



Is the pen mightier than the keyboard? The effect of online testing on measured student achievement[☆]

Ben Backes^{*,1}, James Cowan¹

American Institutes for Research, United States

ABSTRACT

Nearly two dozen states now administer online exams to deliver testing to K-12 students. These tests have real consequences: their results feed into accountability systems, which have been used for more than a decade to hold schools and districts accountable for their students' learning. We examine the rollout of computer-based testing in Massachusetts over 2 years to investigate test mode effects. Crucial to the study design is the state administering the same exam (PARCC) in online and offline formats each year during the transitional period. We find an online test penalty of about 0.10 standard deviations in math and 0.25 standard deviations in English language arts (ELA), which partially but not fully fades out in the second year of online testing.

1. Introduction

Computer-based testing is rapidly spreading across the assessment landscape. There are significant advantages of using computer-based tests (CBTs) to assess student achievement, such as more flexibility in test item design, access to a large repository of items, reduced scope for cheating, and faster turnarounds for score receipt (Parhizgar, 2012). In addition, computer literacy is essential in the workplace, for both finding jobs and succeeding at them (Murray, Sherburn, & Pérez, 2007). Not surprisingly, many national and international assessments are in the process of moving online or have already done so, including the ACT, PISA, and NAEP.² The two consortia of Common Core-based tests that deliver K-12 testing, Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced, are also in the process of transitioning to online testing.³ And as of 2016, nearly two dozen states administer K-12 state assessments online (Farmer, 2016).

Anecdotal reports from the initial implementation of PARCC in

several states have revealed lower scores for students who take PARCC online.⁴ Examples abound: in Ohio, a survey of districts found that 85% of districts administering PARCC on paper received an “A” grade, compared to only 17% of online districts.⁵ In Illinois, 43% of students who took PARCC on paper scored proficient or above in ELA, compared to 36% of students who took the test online. In Maryland, middle school ELA students in Baltimore County scored worse on the online test after controlling for student background and prior achievement.⁶ However, these observed differences could reflect pre-existing differences across students rather than inherent difficulty of online testing.

In this paper, we use the rollout of online PARCC in Massachusetts to investigate whether these findings are reflective of real differences in student achievement or of the causal effect of delivery mode, which we refer to as test mode effects. To our knowledge, this is the first large-scale study of PARCC test mode effects conducted with multiple years of data across an entire state. In 2015, some districts began transitioning to the PARCC assessment. These districts had the choice of using the

[☆] This research was funded by IES Research Grant R305A170119 and supported by the National Center for the Analysis of Longitudinal Data in Education Research (CALDER), which is funded by a consortium of foundations. For more information about CALDER funders, see www.caldercenter.org/about-calder. The data were provided by the state of Massachusetts and the state had the right to review the paper prior to public release in order to ensure that the paper does not disclose any personally identifiable information provided by the state. We thank Carrie Conaway, Kathy Flanagan, Dan Goldhaber, Kris Holden, and Bob Lee for their feedback.

^{*} Corresponding author.

E-mail addresses: bbackes@air.org (B. Backes), jcowan@air.org (J. Cowan).

¹ Backes and Cowan are employed by American Institutes for Research (AIR), which has been involved with delivering Smarter Balanced, which like PARCC is a Common Core assessment. Neither Backes nor Cowan are working or have ever worked in AIR's assessment group, and no one from AIR's assessment group has reviewed the paper. The authors, Backes and Cowan, have final editorial privilege.

² For convenience, we use CBTs and online tests interchangeably throughout.

³ <http://parcconline.org/assessments/test-design/whats-different-this-year>

⁴ PARCC requires that schools have an Internet connection to administer the test: http://www.parcconline.org/files/72/Technology%20Guidelines%20for%20PARCC%20Assessments/389/TechnologyGuidelinesPARCCAssessments-v5_0-Jan2016.pdf

⁵ O'Donnell, P. (2016, March 7). School districts got A grades on paper tests, but F grades online, survey shows. *Cleveland.com*.

⁶ Herold, Benjamin. (2016, February 3). PARCC Scores Lower for Students Who Took Exams on Computers.

paper or online version of the test, and nearly half administered the online format in 2015 or 2016. As described below, three districts allowed schools to make their own decisions about adopting online testing, so our estimation strategy uses school-level variation in online testing rather than district level. Crucially, the PARCC is not computer adaptive and PARCC describes the content of the online test as “significantly overlapped” by the paper version’s content (Pearson, 2016).

Using school-level difference-in-differences models and regressions that control for prior student achievement, we find that students administered an online exam score systematically lower than if they had taken the test on paper. In particular, students taking the online version of PARCC scored about 0.10 standard deviations lower in math and about 0.25 standard deviations lower in English language arts (ELA) than students taking the paper version of the test. After several robustness checks, we reject the most plausible competing explanations for these findings. We find no evidence of differential pre-treatment trends in test scores in online schools. We also find no evidence of an effect on science exams that were administered on paper for all schools, even those that switched to CBTs for math and ELA. Finally, we find that the effects cannot be explained by the switch to the PARCC assessment: schools that switched to CBTs in the year after their switch to PARCC tests experienced declines in achievement that are nearly identical to those that switched during the first year of PARCC testing.

Our estimates of mode effects in math and ELA represent extremely large changes in measured student learning: up to 5.3 months of learning in math and 11.0 months of learning in ELA in a 9 month school year.⁷ Our preferred estimates of the effects of online testing in the second year of administration suggest that mode effects for second-time test takers were about one third as large as the first year in math and about half as large in ELA. Student familiarity with online testing explains part of the reduction in effect size in both subjects, and we find some evidence of general improvements in student performance on the online ELA test. In addition, our exploration of the dynamic effects of online testing suggests that we are truly measuring a mode effect rather than an assessment transition effect. For example, we find substantial online test penalties for third grade students – taking PARCC for the first time – in both 2015 and 2016.

Second, we explore whether states that have fully moved to online testing should expect heterogeneity in test mode effects based on student background, which could cause test mode effects to be conflated with true differences in student achievement across subgroups. For instance, there may be systematic differences in students’ comfort level with CBTs depending on their access to computers in the home and at school. Low income and minority households are significantly less likely to have high speed internet connections (Richtel, 2012). Urban schools are also less likely to have computers with Internet access, making it more difficult to implement or practice with CBTs (Garland, 2012). While we find little systematic evidence of variation in treatment effects by student demographic group in math, we find that ELA mode effects are stronger for students at the bottom of the achievement distribution, for English language learners, and for special education students.

The differential between students’ online and paper based performance could have important consequences, not only for students themselves but also for their teachers, schools, and communities. For instance, standardized tests have a wide range of potential consequences for students, including identification for gifted and talented programs, consideration for special education programs, and being flagged for grade retention. Our findings on CBT transitions are also consequential beyond the impact on individual students in three important ways. First, in many states, the results of standardized assessments support teacher evaluation, school accountability

determinations, student graduation, or the distribution of school resources. Second, families make residential location decisions in part based on measured performance of local schools (Black & Machin, 2011). Third, student test scores are frequently used as an outcome by researchers. Many of the large states frequently used to conduct education research, such as Michigan, Massachusetts, New York, and North Carolina, have recently administered both paper and CBTs simultaneously or plan to in the near future.⁸ These states are frequently sites of education research. Based on a search of papers using the I2 series JEL codes published in American Economic Review, Quarterly Journal of Economics, or Journal of Public Economics between 2013 and 2017, of those listing a state in the abstract, 26% (5 of 19) reported one of these four states. The public K-12 systems in these four states alone contained nearly 7 million students in 2014–15, or approximately 13 percent of public school children in the United States.⁹ Failure to take into account mode effects may lead to biases in empirical analyses of education policies, especially for analyses of school-wide programs. For example, an analysis of a policy change in Massachusetts targeting schools that happened to adopt online testing during the same time frame could find negative effects even if the true policy effect were null or positive.

2. Background

In 2015 and 2016, PARCC administered paper and online versions of the assessment. The paper versions of the PARCC assessment were adapted from the online forms and used a similar set of items. The online versions of the test included some interactive questions, and the paper and online versions were not exactly equivalent in any grade or year. However, both modes did include a subset of linked items to facilitate the reporting of student scores on a common scale (Pearson, 2016; Pearson, 2017). Following the administration of the test, PARCC scored the tests for each mode separately and then transformed results from the paper tests onto the online scale using results from the common set of linked items. The scores were therefore intended to be comparable across modes.

Nonetheless, news reports have documented relatively high scores of students taking paper, as compared to online, tests in several PARCC states (Brown, 2016; Farmer, 2016; O'Donnell, 2016; Tuoti, 2014). Despite the significant difference in measured achievement across test modes, only a few studies have attempted to estimate mode effects on new tests aligned with the Common Core. The PARCC consortium conducted a study based on results from a limited number of schools in a field test and concluded that “there is substantial evidence indicating that the differences in comparability across mode are relatively minor” (Brown et al., 2015, p. 71), although mode effects were larger in writing tasks (Pearson, 2017). However, the PARCC study did not include a measure of prior performance to ensure baseline equivalence between the paper and online samples. On the other hand, Duque (2017), when

⁸ Across all PARCC states, about 80 and 90 percent of students took the assessment online in 2015 and 2016, respectively (Pearson, 2016; Pearson, 2017). In Michigan, about 80 percent of schools administered assessments online in 2015 and 95 percent in 2016 (Michigan Department of Education, 2016). New York State began transitioning to online testing in 2017. Participation in online testing is currently voluntary and schools can elect to administer tests for individual grades and subjects online. The state plans to have all schools using CBTs by 2020 (New York State Education Department, 2016, 2017). North Carolina has been transitioning its tests online since 2015. Starting in 2018, the grade 3–8 assessments will all be available in both online and paper formats (North Carolina Department of Public Instruction, 2017). Other states, such as Florida, have administered a single mode for each assessment, but have rolled out online tests in different grades over multiple years (Florida Department of Education, 2017).

⁹ Digest of Education Statistics, 2016. Table 203.40. https://nces.ed.gov/programs/digest/d16/tables/dt16_203.40.asp

⁷ We describe how we convert our estimates to months of learning using the Lipsey et al. (2012) translations more fully in Section 5.4 below.

controlling for prior achievement, found mode effects of between 0.05 and 0.44 standard deviations in favor of students taking the paper version of the PARCC in Baltimore County Public Schools (BCPS). However, because BCPS determined a school's test mode by the school's student to computer ratio, the results from BCPS may not generalize to other settings. Going beyond PARCC, assessing the research literature on other computerized tests at the K–12 level, Wang et al. (2007) conclude that the average study finds that students taking a paper test score about 10% of a standard deviation higher than those taking a CBT.

One possibility for the mode effects reported is that they reflect true differences in achievement across districts. For instance, if lower-achieving districts were more likely to switch to CBTs, we would expect their subsequent performance on CBTs to be lower even in the absence of a mode effect. Another potential reason for the mode effects observed in the studies mentioned above is temporary adjustment to a new test format. PARCC offers this explanation, with chief of assessment Jeff Nellhaus attributing some of the difference to “student familiarity with the [CBT] platform” (Brown, 2016). If this were the case, it would be possible for mode effects to lower all students' achievement and dissipate over time, and states and districts should want to exercise caution in using transition-year scores in accountability systems but expect the issue to eventually correct itself. Although average achievement may not be strictly comparable across years and should be expected to rise as students become accustomed to the tests, achievement gaps and accountability measures could have a consistent interpretation, provided that the temporary effect on achievement does not differ across subgroups. On the other hand, if mode effects differ by student characteristics or do not fade over time, then switching to online tests could affect school accountability ratings or demographic achievement gaps in unexpected ways.

Several empirical findings suggest that CBT mode effects may depend on student or school characteristics. First, in contrast to paper-and-pencil tests, there may be meaningful differences across schools in the instruments used to complete CBTs. Some schools have limited access to fully updated computers or the high-speed Internet access required for the online testing platform. For instance, urban schools have fewer computers per student and their students are less likely to report using computers frequently for school work (Snyder, de Brey, & Dillow, 2018). Differences in computing facilities may matter because several studies have found that technological aspects of the CBTs may affect student performance. Davis et al. (2013) reported that students found testing devices with larger screens easier to read. And, in an analysis for the PARCC consortium, Liu et al. (2016) found evidence of differences in performance on online tests across different types of devices in at least one state.

Second, CBTs may also measure skills, such as computer literacy, for which student proficiency differs. The online versions of the PARCC use web design features that might be unfamiliar to students who have less prior exposure to computers or the Internet. In Fig. 1, we display reading passages from the sample PARCC assessment's paper and online formats. The paper version of the test (Fig. 1a) displays reading passages across multiple pages in the test booklet. On the other hand, the online version (Fig. 1b) displays the full passage in a box embedded in a single page with multiple-choice questions. Students are expected to scroll down in the text box to view the complete passage. In addition, when taking the test online, students may not be as comfortable performing tasks that are simple on paper (e.g., highlighting a section of a passage).

The online version also uses multiple display formats for multiple-choice questions, as shown in Fig. 2. The paper version (Fig. 2a) asks which answer best describes what the picture adds to the story and then provides several possible responses labeled with a corresponding entry on the answer sheet. Although most questions use a familiar radio

button display for the multiple-choice questions, the assessments also use other formats for some questions. For the question in Fig. 2, the online version (Fig. 2b) includes an additional question (“What is Happening?”) and then asks the student to identify what the passage adds and presents three options in a dropdown menu. The online assessment also uses drag-and-drop and other interactive response methods.

Finally, the two versions of the test use different structures for free response questions. In Fig. 3, we show the formats for an essay question in which students respond to a reading passage. The paper assessment (Fig. 3a) includes two ruled sheets for students to write out their responses. For the online version, students type their responses in a text box located next to the reading prompt (Fig. 3b). The text box includes some basic editing, formatting, and spellcheck features. The box also expands to accommodate students' responses, although the amount of space displayed on the page is less than the corresponding space in the paper version.

If students have differing familiarity with these features, their accuracy in recording responses may differ. Although Massachusetts is a relatively wealthy state, computer penetration lags behind in some areas, especially for minority households. Of the 52 Public Use Microdata Areas (PUMAs) in the state, 19 report household computer access upwards of 95%.¹⁰ However, in several regions, fewer than 80% of households have access, and in many of those, fewer than 70% of Black and Hispanic households have computer access. In the PUMA containing Springfield, an area with a heavy minority population (non-Hispanic Whites make up only 37% of the population) and the second-largest school district in the state, about 35% of Hispanics and 28% of Blacks report no computer access at home. Disparities in computer access may moderate mode effects on standardized tests. Studying an online version of the NAEP writing test, White, Kim, Chen, and Liu (2015) found that home Internet access was associated with response length, use of editing tools, and preference for the computer test. Perhaps as a consequence, high-ability students tend to score disproportionately high on CBTs (Clariana & Wallace, 2002; White et al., 2015).

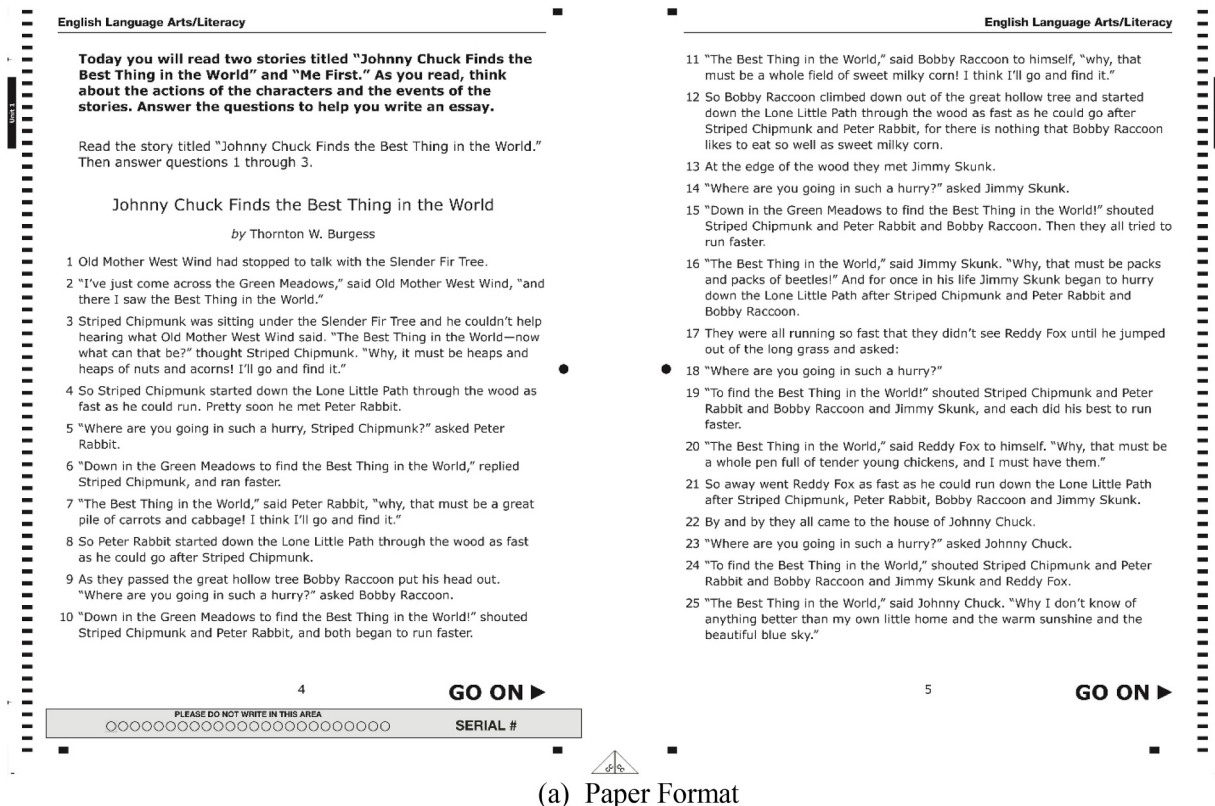
3. Setting and sample

Massachusetts adopted new state curriculum frameworks incorporating the Common Core State Standards in 2011, with implementation beginning in the 2012–13 school year (from here forward, we will refer to school years by the spring of that year). Until 2014, all districts used the Massachusetts Comprehensive Assessment System (MCAS), which was administered on paper. Beginning in 2015, districts chose between MCAS and the new PARCC assessment, with three districts having a mix of online and paper.¹¹ As shown in Table 1, about 72% of elementary or middle schools in our sample administered the PARCC in either 2015 or 2016. PARCC districts had the additional option of offering the test online or on paper. Of those schools administering the PARCC in either 2015 or 2016, 57% administered the test online at least once.

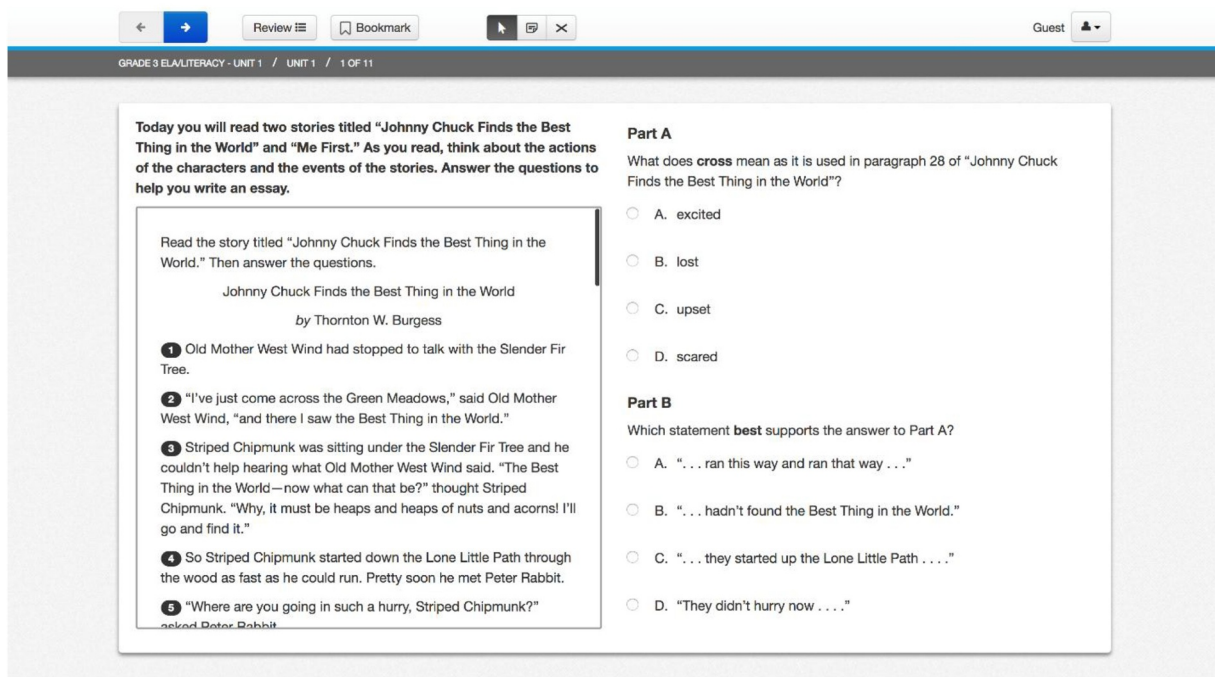
It is important to note that Massachusetts implemented a hold-harmless provision for all schools that switched to PARCC in either 2015 or 2016. During this period, no school's accountability rating could fall as a consequence of their PARCC scores. Because hold harmless applied to all PARCC schools, whether administering online or

¹⁰ Authors' calculations using data from the 2013 and 2014 American Community Survey (Ruggles et al., 2015).

¹¹ Boston, Worcester, and Springfield had the option of assigning individual schools to the online or paper format. Otherwise, districts selected a single test administration for the entire district. In November 2015, the Massachusetts State Board of Education voted to discontinue the PARCC assessment and implement a redeveloped version of the MCAS in all schools beginning in 2017.



(a) Paper Format



(b) Online Format

Fig. 1. Reading passage display formats on online and paper assessments.

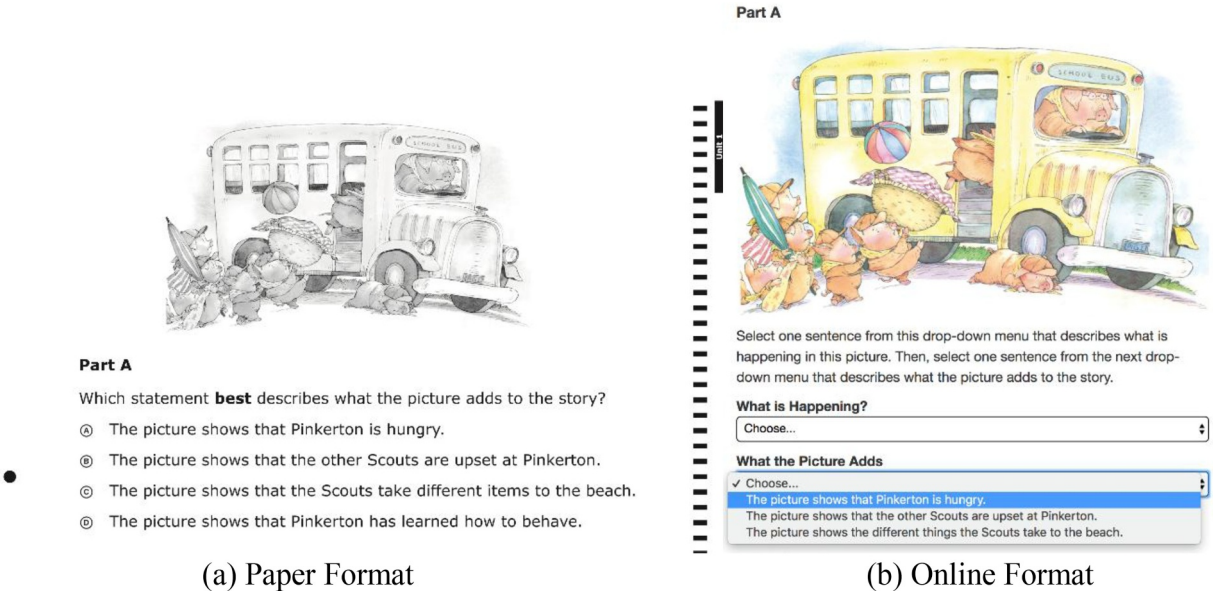


Fig. 2. Multiple-choice question display formats on online and paper assessments.

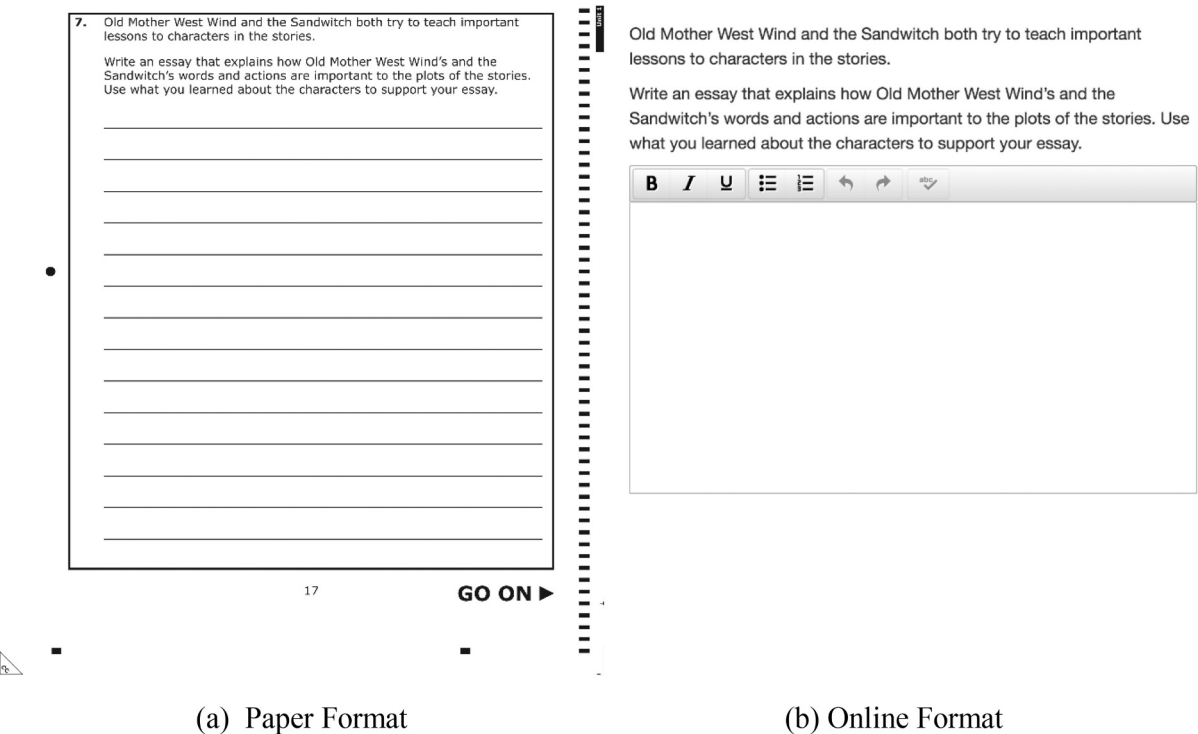


Fig. 3. Free-response question formats on online and paper assessments.

on paper, both groups faced the same accountability environment. Thus, while we present evidence for the internal validity of our results, it is possible that the online test penalty could be smaller in higher-stakes settings where schools may have more incentive to devote more resources to prepare students to take the test online.

In this study, we use student achievement data for students in Massachusetts public schools during the 2011–2016 school years in Grades 3 through 8 (as described below, some of our models will restrict our sample to Grades 5 through 8). We use longitudinal student achievement data that has been linked to student data in the Student

Table 1
Test transition in Massachusetts.

2015 Assessment	2016 Assessment	Number of schools	Number of students
MCAS	MCAS	339	219,075
MCAS	PARCC Paper	164	92,881
MCAS	PARCC Online	39	32,692
PARCC Paper	PARCC Paper	285	166,508
PARCC Paper	PARCC Online	26	18,398
PARCC Online	PARCC Paper	80	49,302
PARCC Online	PARCC Online	267	157,519

Notes: Each row shows the number of schools in the analysis sample with the given assessment choice. We include the bolded schools, which administered PARCC in both 2015 and 2016, in the analysis sample.

Information Management System (SIMS) by the state, which includes information on students' enrollment status, demographics, and program participation.

We make several sample restrictions to properly identify student test mode. First, we limit our sample to schools that administered the PARCC in both 2015 and 2016 to ensure that achievement is measured on a common scale in each year.¹² Because we implement difference-in-differences designs and use lagged school outcomes as key regressors, we also exclude schools that were not operating continuously between 2011 and 2016. Our final sample includes about half of all students enrolled in Grades 3 through 8 between 2011 and 2016 and 88 percent of students in schools administering the PARCC in 2015 and 2016.¹³

We depict achievement trends graphically in Fig. 4, which plots mean standardized test scores by year for schools switching to the PARCC assessment in 2015 based on their mode decisions. This figure shows the first suggestion of mode effects: although test scores are higher in the online districts in 2011–2014, the schools that switch to online testing in 2015 then see their achievement fall behind the schools that remained on paper.¹⁴ We display summary statistics in Table 2, dividing the sample between the two test modes: PARCC online and PARCC paper. Each sample contains hundreds of unique schools and hundreds of thousands of student-year observations. Schools administering the PARCC exam on paper tend to have lower test scores and more low-income students than those administering the PARCC online.

¹² Districts that switched to PARCC in 2015 could not switch back to MCAS in 2016. We also omit 3,229 observations for students in schools where more than 5 percent of students have a test mode that does not match the typical choice in their school. Massachusetts translated PARCC scale scores to equivalent MCAS scale scores (Massachusetts Department of Elementary and Secondary Education, 2016a, 2016b). Given the significant differences between the MCAS and PARCC schools in terms of student observables (Table 2), we do not use the rescaled scores in this analysis. Before 2015, we use the MCAS scores standardized within the set of PARCC schools that comprise this sample. In 2015 and 2016, we similarly standardize the PARCC scores. The standard deviation of test scores in this sample is between 0.96 and 1.01 standard deviations measured in the full sample in each grade, subject, and year, so this standardization does not materially affect the coefficient estimates presented in this paper.

¹³ Besides omitting schools that open during this period, students in PARCC schools can be missing from the sample for reasons such as taking an alternate version of the assessment or opting out of the assessment.

¹⁴ The scores of paper schools increasing when online testing was introduced to the online schools is a mechanical consequence of test scores being standardized to have mean zero and standard deviation one within subject, grade, and year. Because mean scores in the online schools fall substantially, mean scores in the paper schools must rise in order for test scores to remain centered at zero. We show an alternate version of Figure 4 as Appendix Figure 1 where we use scale scores (transformed onto the MCAS scale in case of PARCC assessment) instead of normed scores.

Consistent with Fig. 4, average MCAS achievement in online districts prior to the implementation of PARCC is about 0.09 standard deviations higher in math and 0.11 standard deviations higher in ELA than in paper districts. Although prior test scores are higher in the online districts, current test scores are 0.13 standard deviations lower in ELA and 0.01 standard deviations lower in math. Regarding differences in baseline characteristics, 42% of students taking the paper test and 35% of students taking the online test qualify for subsidized lunches. In the next section, we discuss our strategies for accounting for the clear differences in baseline characteristics.

4. Empirical strategy

Assessments are intended to measure how much a given student knows. We conceptualize test mode effects by writing student i 's test score at time t in school s given test mode m as a function of knowledge, a mode-specific component, and noise:

$$A_{ist}^m = k_{ist} + c_{ist}^m + \varepsilon_{ist}^m.$$

In the preceding, measured achievement depends on a student's true knowledge, k_{ist} , which is constant regardless of whether a test is administered online or on paper; a student-specific test mode effect, c_{ist}^m , and a mean zero error term. We refer to the online test mode effect as representing generic “computer-specific ability,” c_{ist}^m , which may be positive or negative for CBT-takers and is zero for paper-takers and may reflect school factors (e.g., availability of sufficient computers, quality of Internet connection at the school) or the quality of the test design. Thus, relative to paper tests, the properties of CBT-measured achievement can vary because of differences in c_{ist}^m across modes.

For each student in the sample, we only observe achievement on one version of the test per subject and year. That is, we measure

$$A_{ist} = k_{ist} + c_{ist}^C \times CBT_{st} + \eta_{ist},$$

where $\eta_{ist} = \varepsilon_{ist}^C \times CBT_{st} + \varepsilon_{ist}^P \times (1 - CBT_{st})$ is the shock on the given exam. The key empirical challenge is that student ability k_{ist} may be correlated with test mode status. The summary statistics in Table 2 suggest this is the case, and below we describe how we account for these differences in ability.

If students who took the paper version of PARCC tend to do better, even controlling for prior test scores and demographic factors, there are several potential factors for why this may be the case: (1) The paper version may be generically “easier” in some sense, meaning that, on average, a given student would be expected to score higher on the paper test than on the computer test, perhaps due to factors such as ease of reviewing and revising previous responses (Wang et al., 2007). (2) The paper version may be less discriminating among students of differing academic ability so that small preexisting differences in test scores across participating and nonparticipating districts become magnified when switching to online testing (similar to the argument in Cascio & Staiger (2012)), who show that increasing variance of knowledge in later grades partially explains why estimated treatment effects of interventions are smaller in later grades). Or (3) unobservable student or teacher characteristics differ in districts that choose the paper and online versions of the test. The latter may be the case, for instance, if average teacher quality is higher in the districts that chose the paper test.

We take two empirical approaches to estimating mode effects: one that controls for test scores directly and another that uses a difference-in-difference design. First, we use data on student test scores from before the introduction to the online PARCC assessment to adjust for differences in student background. In particular, we estimate

$$A_{ist} = \lambda A_{ist-2} + \alpha X_{ist} + \beta CBT_{st} + \gamma_i + \eta_{ist}. \quad (1)$$

In Eq. (1) above, i indexes students, s indexes schools, and t indexes years. We regress the test score in year t on twice-lagged test scores A_{ist-2} , demographic characteristics X_{ist} , and an indicator for whether

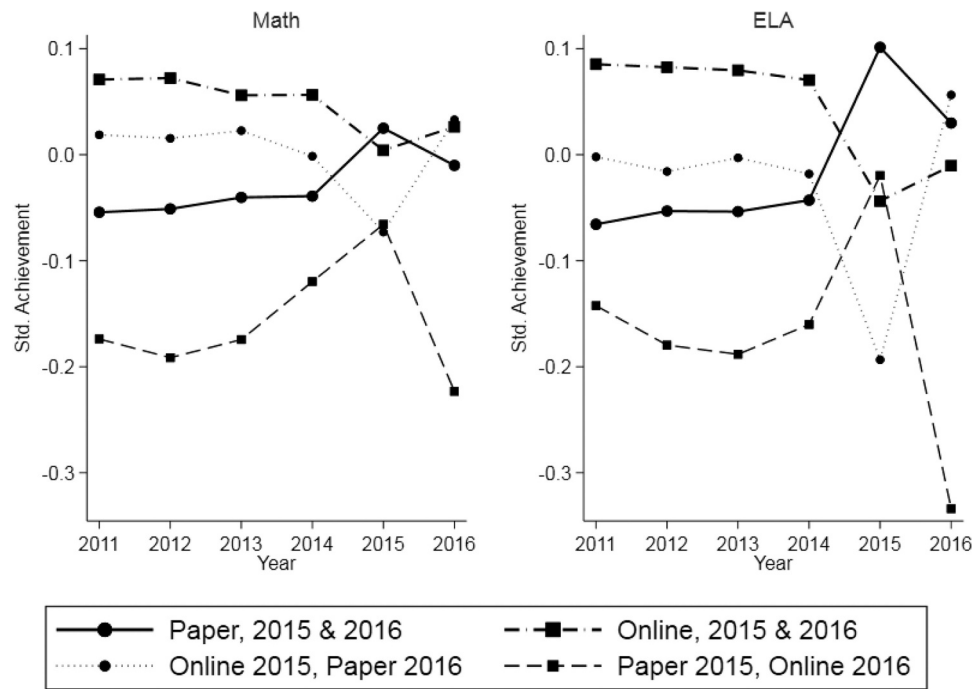


Fig. 4. Standardized achievement by year

Notes: Mean standardized test scores by year for schools administering PARCC in both 2015 and 2016 by online and paper administration. All test scores have been standardized by grade and year within the sample of schools administering the PARCC in 2015 and 2016.

Table 2
Summary statistics.

	Lag score sample (2015–16)		Difference	DID sample (2011–16)		Difference
	Paper	Online		Paper	Online	
Math test	0.00 (1.02)	–0.00 (0.98)	–0.01** (0.00)	–0.03 (1.01)	0.02 (0.99)	0.05*** (0.00)
ELA test	0.06 (0.99)	–0.06 (1.00)	–0.13*** (0.00)	–0.01 (1.01)	0.01 (0.99)	0.03*** (0.00)
Twice-lagged math test	–0.04 (1.01)	0.05 (0.99)	0.09*** (0.00)	–0.05 (1.01)	0.04 (0.99)	0.09*** (0.00)
Twice-lagged ELA test	–0.06 (1.01)	0.06 (0.99)	0.11*** (0.00)	–0.06 (1.01)	0.04 (0.99)	0.11*** (0.00)
Male	0.51 (0.50)	0.51 (0.50)	0.11*** (0.00)	0.51 (0.50)	0.51 (0.50)	–0.00 (0.00)
Hispanic	0.18 (0.39)	0.14 (0.34)	–0.05*** (0.00)	0.17 (0.38)	0.15 (0.35)	–0.02*** (0.00)
African American	0.12 (0.32)	0.08 (0.27)	–0.04*** (0.00)	0.12 (0.32)	0.09 (0.28)	–0.03*** (0.00)
Asian	0.06 (0.24)	0.06 (0.24)	–0.04*** (0.00)	0.06 (0.23)	0.06 (0.23)	0.00 (0.00)
Free lunch eligible	0.39 (0.49)	0.31 (0.46)	–0.08*** (0.00)	0.38 (0.49)	0.32 (0.46)	–0.07*** (0.00)
Reduced price lunch eligible	0.03 (0.17)	0.04 (0.19)	0.01*** (0.00)	0.05 (0.21)	0.05 (0.21)	–0.00 (0.00)
Limited English proficient	0.08 (0.26)	0.04 (0.21)	–0.03*** (0.00)	0.08 (0.27)	0.06 (0.24)	–0.02*** (0.00)
Special education	0.17 (0.38)	0.16 (0.37)	–0.01*** (0.00)	0.17 (0.37)	0.16 (0.37)	–0.01*** (0.00)
Observations	128,643	127,533		506,234	687,694	

Notes: Summary statistics for each analytical sample by school test and test mode choice. Columns (1) to (3) include students in grades 5 through 8 in 2015 and 2016. The paper and online columns indicate contemporaneous test mode. Columns (4) through (6) include students in grades 3 through 8 in 2011–2016. The online column indicates schools that administered an online test in either 2015 or 2016; the paper column indicates schools that never administered an online test. T-test difference in means is provided in the final column with standard errors in parentheses.

* $p < 0.10$.

** $p < 0.05$.

*** $p < 0.01$.

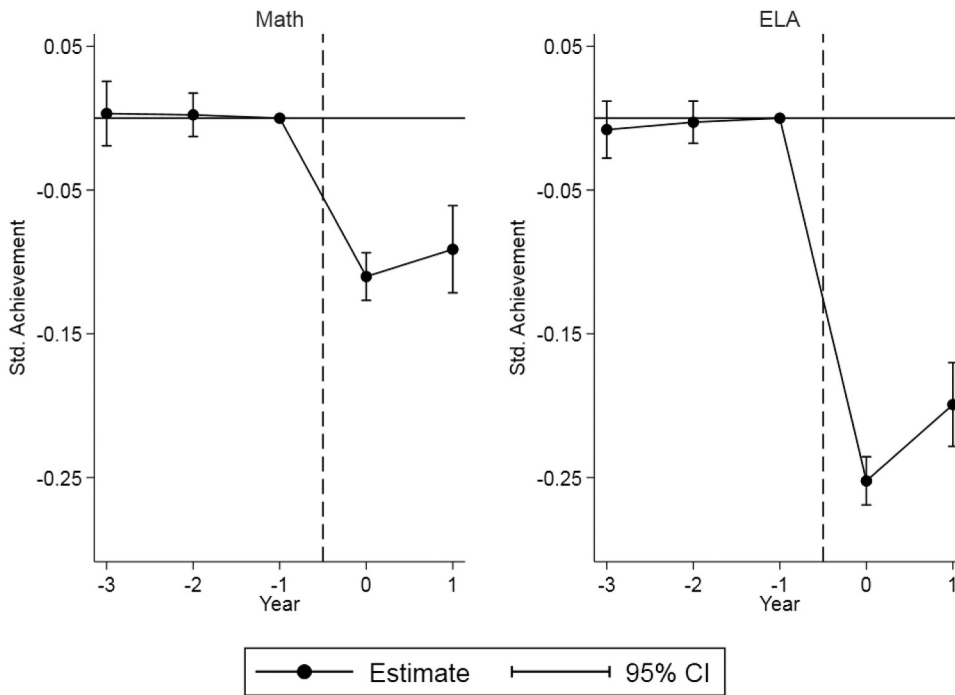


Fig. 5. Effects of computer-based testing by year relative to implementation
Notes: Estimates of impulse response function from regression of achievement on current test mode and three leads and one lag, school-by-grade and grade-by-year fixed effects, and school characteristics. Estimates are normalized relative to the year before implementation. Standard errors clustered at the school level.

the school administers a CBT in year t , C_{st} .¹⁵ Thus, Eq. (1) estimates the differential performance of students who took the test online, relative to those who took the test on paper, which we refer to as the mode effect. X_{ist} also includes a set of school variables, including aggregates of the student-level variables and the full history of school average achievement in math and ELA between 2011 and 2014. Because we have 2 years of PARCC data, all lagged achievement variables in Eq. (1) are measured before the introduction of online testing to avoid having online scores as both an outcome and control variable in 2016. In addition, Eq. (1) is estimated on a sample of students in Grades 5–8 because of the need for twice-lagged scores. The coefficient β then measures the average loss (or gain) in test scores associated with taking a CBT relative to observationally similar students who took a paper test.

Second, we use data from before the introduction of online testing and assess mode effects using a difference-in-differences design. Specifically, we include school fixed effects to remove any unobserved, time-invariant differences across schools:

$$A_{ist} = \alpha X_{ist} + \beta CBT_{st} + \chi_s + \gamma_t + \eta_{ist}. \quad (2)$$

The advantage of this approach relative to Eq. (1) above is that the test mode effect β is estimated from within-school variation over time. This allows for the possibility of time-invariant preexisting differential factors such as teacher or school quality in paper versus online districts. In addition, because we include school fixed effects in place of students' prior achievement, we can include students in third and fourth grade and thus use the full sample of Grades 3–8. As before, the coefficient β then represents the additional gain (or loss) associated with taking the

test online, holding these time-invariant differences across schools constant.¹⁶

Before moving to the empirical results, we present some preliminary checks on the validity of the research designs. We estimate the difference-in-differences models using three leads and one lag of the online testing indicator and plot the cumulative effects of online testing by year in Fig. 5. We find little evidence of preexisting trends in either math or ELA.

Beyond the usual concerns about selection on observables, the coincidence in the timing between the introduction of the PARCC assessment and online testing potentially creates an additional problem for estimation approaches that rely either on lagged scores, such as Eq. (1), or within-school variation in test scores, as in Eq. (2). In both instances, we use test scores from the MCAS to control for baseline achievement, whereas online testing results are generally available only on the PARCC. However, because the 2011 Massachusetts Curriculum Framework (aligned with MCAS) is very similar to the Common Core State Standards (aligned with PARCC), this may not be a concern.¹⁷ In addition, in earlier work, we found that the year to year correlation in test scores was similar during Massachusetts' transition to PARCC as in other years (Backes et al., 2018).

5. Results

5.1. Main results

We begin by displaying our main results in Table 3. In Columns (1) and (5), we regress test scores in year t on twice-lagged test scores, demographic information, and means of each at the school and school-grade-year levels. The achievement outcomes are measured in 2015 and 2016, whereas the lagged achievement measures are derived from MCAS tests administered in 2013 and 2014. Recall that we use twice-

¹⁵ We use official scale scores standardized by year and grade for both the current and lagged achievement variables. Massachusetts transforms the individual student ability estimates into MCAS scale scores using linear transformations that differ across performance categories (Massachusetts Department of Elementary and Secondary Education, 2015). To ensure that the results are not sensitive to using control variables with a potentially non-interval interpretation, we also use a normal curve equivalent transformation of the lagged MCAS achievement variables. None of the results shown is sensitive to using these variables in place of the standardized scores. In addition, all the empirical analyses estimate standard errors that are robust to arbitrary heteroskedasticity in the error term.

¹⁶ As shown below, we also experiment with using student fixed effects instead of school fixed effects. Results are very similar.

¹⁷ "The Massachusetts Working Draft Standards overlap with about 90% of the Common Core." <http://www.doe.mass.edu/bese/docs/fy2011/2010-07/item1.html>

Table 3
Estimated test mode effects.

	Math (1)	(2)	(3)	(4)	ELA (5)	(6)	(7)	(8)
Online	−0.10*** (0.01)	−0.10*** (0.01)	−0.09*** (0.01)	−0.11*** (0.01)	−0.24*** (0.01)	−0.25*** (0.01)	−0.24*** (0.01)	−0.24*** (0.01)
Observations	232,107	778,332	693,994	1,186,422	232,891	777,874	694,664	1,184,736
Controls:								
Lags	X		X		X		X	
School FE		X	X	X		X	X	X
Sample:								
Grades	5–8	5–8	5–8	3–8	5–8	5–8	5–8	3–8
Years	2015–16	2011–16	2011–16	2011–16	2015–16	2011–16	2011–16	2011–16

Notes: Regressions of standardized math (Columns (1)–(4)) and ELA (Columns (5)–(8)) scores on test mode. Regressions in Columns (1) and (5) use achievement data from 2014–15 and 2015–16 and include a cubic polynomial in twice-lagged achievement, student race/ethnicity, free-and-reduced-price-lunch status, special education status, limited English proficiency status, grade-by-year indicators, and each of the means of each of these variables at the school-year and school-grade-year levels. Regressions also include mean school achievement for each year between 2011 and 2014. Regressions in Columns (2)–(4) and (6)–(8) use achievement data from 2011–12 through 2015–16 and replace all school- and individual-level prior achievement variables with school-by-grade fixed effects. Regressions in Columns (2), (3), (6), and (7) use Grades 5–8 only; those in Columns (4) and (8) also include Grades 3 and 4. Regressions in Columns (3) and (7) additionally control for students' twice lagged achievement. Standard errors clustered by school in parentheses.

* $p < 0.10$.

** $p < 0.05$.

*** $p < 0.01$.

lagged test scores to avoid having online scores as both an outcome and control variable: the test scores used as controls are always from MCAS paper assessments. In the remaining columns, we replace the individual- and school-level student achievement controls and estimate difference-in-differences models. In Columns (2),(3) and (6),(7), we use the same grades as the ordinary least squares (OLS) models for comparison. We then include data from the full set of tested grades (3–8) between 2011 and 2016.

The results are generally similar across models. We estimate online test effects of −0.10 in math and −0.24 in ELA controlling for prior test scores.¹⁸ When we estimate mode effects on the same grades with difference-in-differences models, we estimate effects of −0.10 and −0.25, respectively. Adding twice-lagged test scores to the difference-in-difference models has little impact on the estimated effect of online testing. Our preferred estimates in Columns (4) and (8), which include data from all grades, are quite similar: −0.11 standard deviations in math and −0.24 standard deviations in ELA. The results of Table 3 indicate that students taking PARCC online score lower than observationally similar students who take the test on paper.

5.2. Robustness checks

In this section, we investigate whether our estimates of test mode effects could be driven by unrelated changes in school achievement. In Table 4, we therefore conduct several further tests of the main research design. To test whether preexisting trends in school outcomes explain our estimated mode effects, we conduct a placebo test on science achievement in Grades 5 and 8, which was administered on paper forms throughout this period. We should therefore expect to find null results on these placebo tests. In Columns (1) and (2), we replicate the OLS and

difference-in-differences regressions in Table 3 using the science achievement scores as a dependent variable. In each case, the set of explanatory variables is identical to those estimated in previous regressions. We find no evidence of mode effects in either specification: The estimates are near zero (−0.005 and −0.001, respectively) and statistically insignificant.¹⁹

In the remaining columns, we use different samples of schools to further test the robustness of our results. The primary results in Table 3 use the PARCC paper schools to identify the counterfactual time trend for online schools. We therefore implicitly assume that the transition from MCAS to PARCC testing affected both types of schools similarly. Testing for pre-existing trends in online PARCC schools relative to paper PARCC schools does little to support this specific assumption given the simultaneous introduction of PARCC and online testing. Although the Massachusetts and Common Core standards were similar, we may still conflate test mode and test type effects.

In columns (3) and (6), we restrict our sample to 2015 and 2016 and re-estimate the DID models. These models only use variation in test format generated by schools that switched modes between 2015 and 2016. Because all schools in the sample administer the PARCC during this window, our estimates should not be sensitive to changes in the test content. Thus, if our estimates are measuring a test mode effect rather than volatility associated with the introduction of a new assessment, we would expect to find mode effects in 2016 for schools that administered PARCC in 2015 and 2016 that switched to online testing in 2016. Indeed, the estimates are quite similar to those in Table 3: we find that switching to an online test mode reduces achievement by 0.12 standard deviations in math and 0.27 standard deviations in ELA.

Next, we broaden the sample to include schools switching to the PARCC in 2016 (schools in rows 2 and 3 of Table 1). Most of these schools switched to the paper test, and their prior average achievement is significantly lower than schools administering the PARCC online. Nonetheless, their inclusion does little to change the estimated mode

¹⁸ While the difference across subjects is notable, our analysis of average differences across scale scores is not well suited for an investigation of factors driving these mode effects. One piece of suggestive evidence comes from sub-scale scores for ELA, which we do not include in the paper because we only have them for one year. Nevertheless, using our OLS specifications, we find larger first-year ELA effects on the Writing portion of the test (0.35 standard deviations) than for Reading (0.20 standard deviations).

¹⁹ We also estimate difference-in-differences models that include school-specific linear time trends. The results are quite similar to the main difference-in-differences estimates, although we note that Figure 5 does not suggest clear evidence of preexisting trends among the online adopters.

Table 4
Robustness checks.

	Science (placebo)		Math: Columns 3–5			ELA: Columns 6–8		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Online test	– 0.00 (0.02)	– 0.00 (0.01)	– 0.12*** (0.01)	– 0.10*** (0.01)	– 0.11*** (0.01)	– 0.27*** (0.01)	– 0.21*** (0.01)	– 0.25*** (0.01)
Model	OLS	DID	DID	DID	Student FE	DID	DID	Student FE
Grades	5,8	5,8	3–8	3–8	3–8	3–8	3–8	3–8
Years	2015–2016	2011–2016	2015–2016	2011–2016	2011–2016	2015–2016	2011–2016	2011–2016
N	116,881	386,246	389,273	1,567,707	1,186,422	389,657	1,565,192	1,184,736

Notes: Regressions of standardized science (Columns (1) and (2)), math (Columns (3)–(6)), and ELA (Columns (6) - (8)) scores on school's PARCC test mode. Regressions in Columns (1) and (2) use the sample of schools in the primary analysis in Table 3 with 5th and 8th grade MCAS science scores. Regression in Column (1) uses achievement data from 2014–15 and 2015–16 and includes a cubic polynomial in twice-lagged achievement in math and ELA, student race/ethnicity, free-and-reduced-price-lunch status, special education status, limited English proficiency status, grade-by-year indicators, and each of the means of each of these variables at the school-year and school-grade-year levels. The regression also include mean math and ELA school achievement for each year between 2011 and 2014. The regression in Column (2) uses achievement data from 2011–12 through 2015–16 and replaces all school- and individual-level prior achievement variables with school-by-grade fixed effects. Regressions in Columns (3) and (6) use only data from PARCC schools in 2015 and 2016. Regressions in Columns (4) and (7) also add the schools adopting the PARCC assessment in 2016 to the sample (schools in Rows 2 and 3 of Table 1). Regressions in Columns (5) and (8) use student fixed effects instead of the school-by-grade fixed effects and do not control for any individual-level variables. N denotes number of student-year observations. Standard errors clustered by school in parentheses.

* $p < 0.10$.** $p < 0.05$.*** $p < 0.01$.**Table 5**
Mode effects by year of administration.

	Math (1)	(2)	(3)	ELA (4)	(5)	(6)
Online	– 0.17*** (0.02)	– 0.14*** (0.02)	– 0.13*** (0.02)	– 0.31*** (0.02)	– 0.30*** (0.02)	– 0.28*** (0.02)
x 2nd year school	0.06*** (0.01)	– 0.01 (0.02)	– 0.02 (0.02)	0.11*** (0.01)	0.08*** (0.02)	0.04*** (0.01)
x 2nd year grade		0.08*** (0.02)			0.04 (0.02)	
x 2nd year student			0.10*** (0.02)			0.09*** (0.01)
x 4th grade	0.04** (0.02)	0.00 (0.02)	– 0.00 (0.02)	0.10*** (0.02)	0.08*** (0.02)	0.06*** (0.02)
x 5th grade	0.09*** (0.02)	0.05** (0.02)	0.04** (0.02)	0.02 (0.02)	0.00 (0.02)	– 0.02 (0.02)
x 6th grade	0.05* (0.03)	0.01 (0.03)	0.01 (0.03)	– 0.03 (0.02)	– 0.05** (0.02)	– 0.07*** (0.02)
x 7th grade	0.06** (0.03)	0.02 (0.03)	0.02 (0.03)	0.03 (0.03)	0.02 (0.03)	– 0.00 (0.03)
x 8th grade	– 0.02 (0.03)	– 0.05* (0.03)	– 0.06** (0.03)	0.07*** (0.03)	0.06** (0.03)	0.04 (0.03)
N	1,186,422	1,186,422	1,186,422	1,184,736	1,184,736	1,184,736

Notes: Difference-in-differences estimates of mode effects by grade and year of administration. All models include same variables as in Table 3. Standard errors clustered by school in parentheses. N denotes number of student-year observations.

* $p < 0.10$.** $p < 0.05$.*** $p < 0.01$.

effects in either math or ELA. Finally, the introduction of online testing may affect which students take the standardized tests. Students' opting out of standardized testing was widely reported in some states switching to Common Core aligned tests (Ujifusa, 2015), and the movement to online tests may have been a more salient signal of changes in testing policy. To assuage concerns that changes in the composition of test takers explains our results, we replace school-by-grade fixed effects with student fixed effects in Columns (5) and (8),

with very similar results to our main results in Table 3.²⁰ We conclude that sorting into test mode based on preexisting trends is unlikely to drive our findings.

²⁰ In addition, we directly test for effects of online testing on the likelihood that enrolled students had valid testing scores in either math or ELA, replicating the DID specification from Table 3. The coefficient on online testing is 0.001 and statistically insignificant.

5.3. Dynamic effects of computer-based testing

Over time, schools or districts may improve their ability to administer online tests, or students may become more familiar with the testing software. We therefore test whether mode effects differ on subsequent administrations of the test. We replicate the results in Table 3 using indicators for the number of cumulative administrations of online PARCC tests. For the remainder of the paper, we use the DID specification used in Columns (3) and (6) in Table 3 so that we can use the entire Grade 3 - 8 sample.²¹ We measure prior exposure to online tests in three ways. First, we measure the number of times a school has previously administered the PARCC test online. Second, we measure the number of times a cohort has taken the online PARCC. This is generally the same as the school-level measure, except for students in third grade who have not previously taken the PARCC. That is, for schools that administered PARCC online in 2015, fourth graders and above in 2016 have taken the PARCC online previously, while those in third grade in 2016 have not. Finally, we measure the number of times an individual student has taken the PARCC online. This differs from the two measures above for students who switch to an online school from a paper school in 2016 (or vice versa). We use these sources of variation to separately identify the effects of student familiarity and school implementation. Coefficients on subsequent online tests measured at the individual level that are significantly smaller than the first time would be taken as evidence of student-specific fade out, while coefficients on the school level measure that are significantly smaller would suggest that district or school test coordination improves over time.

We present the results of this exercise in Table 5. In each column, the coefficient on the online indicator is presented first and the coefficient for a set of interaction terms is presented second. In each case, we also include interactions of the mode effect with grade. We focus on the interaction effects on the number of school and student administrations. In Columns (1) and (4), we estimate the effects of online testing in the first and second year a school administers the exam. The coefficient on the online indicator provides an estimate of the first-time mode effect for third grade students and the interaction with second year provides an estimate of the change in the mode effect during the second year of administration. In math, we estimate that the mode effect diminishes by 0.06 standard deviations between the first and second year. In ELA, the mode effect diminishes by 0.11 standard deviations.²²

In the next columns, we attempt to disentangle the effects of student familiarity from schoolwide improvements in administration by including controls for prior student experience with PARCC. As we discussed above, these coefficients describe slightly different sources of variation in the number of test administrations. In Columns (2) and (5), we add a control for whether students in a cohort have had a prior administration of the PARCC online. This differs from the number of prior school administrations only for students in third grade, who have not previously taken the PARCC in either 2015 or 2016 and therefore have no prior online PARCC experience. Thus, we compare changes in the mode effect in third grade to changes in the mode effect in fourth through eighth grades for students who have previously taken an online PARCC test. In Columns (3) and (6), we include an indicator for the

Table 6
Heterogeneous effects.

	Math	ELA		Math	ELA
Full Sample			Asian		
Online	−0.11 (0.01) [1.00]	−0.24 (0.01) [1.00]	Online	−0.09 (0.02) [0.34]	−0.21 (0.02) [0.09]
Male			Free/Reduced-Price Lunch		
Online	−0.11 (0.01) [0.71]	−0.24 (0.01) [0.56]	Online	−0.12 (0.01) [0.33]	−0.26 (0.01) [0.01]
Hispanic			Limited English Proficient		
Online	−0.12 (0.02) [0.41]	−0.25 (0.02) [0.30]	Online	−0.11 (0.02) [0.99]	−0.28 (0.03) [0.04]
African American			Special Education		
Online	−0.13 (0.02) [0.27]	−0.27 (0.02) [0.09]	Online	−0.14 (0.01) [0.00]	−0.29 (0.01) [0.00]
<i>N</i>	234,397	235,194	<i>N</i>	234,397	235,194

Notes: Regressions of standardized math and ELA scores on test mode. Regressions use same base sample (grades 3–8) and specifications as DID estimates in Table 3. Each cell represents a separate regression using the specified subsample. Standard errors clustered by school in parentheses. *P*-value of test against equality to full sample coefficient in brackets. *N* denotes number of student-year observations.

**p* < 0.10.

***p* < 0.05.

****p* < 0.01.

second time a student takes a PARCC assessment. This variable differs from the school administration variable for students in third grade, but it also identifies the effect of prior experience with online PARCC from students who switch into online schools from paper schools in different years.

We find some evidence that a reduction in mode effects results from both student familiarity with the online format and school experience administering the exam. In math, student familiarity explains the full reduction in the size of the mode effect between the first and second years. Once we adjust for the number of prior times a student has taken the exam (the coefficients on prior student administrations and prior grade administrations), we find little evidence of schoolwide reductions in the mode effect between the first and second years. In Column (2), we see no improvement in the mode effect for third graders between 2015 and 2016, which suggests that improvement in online test outcomes result from students becoming more familiar with the exam. Similarly, when we control directly for students' prior number of online test administrations in Column (3), we find little evidence of schoolwide improvements from 2015 to 2016. This suggests that student familiarity explains much of the change.

On the other hand, we do find evidence of general administrative effects on the ELA test. In Column (5), we see relatively large reductions in the mode effect among third graders (the coefficient on second year school), which is consistent with general schoolwide improvements in online test results. The reduction in the mode effect for older students is slightly larger than for third graders (coefficient on second year grade), but this difference is not statistically significant. However, when we control directly for the number of times a student has previously taken the PARCC, we do find more evidence of a familiarity effect. The coefficient on a student's second administration is 0.09 standard deviations, and including this measure reduces the schoolwide administration effect from 0.08 to 0.04 standard deviations. Overall, student familiarity with the PARCC

²¹ Results are similar when using OLS models.

²² The mode effects could also diminish over time if schools with less successful implementation of online testing were disproportionately likely to switch back to paper testing in 2016. We do find some evidence of this: first-year test mode effects were larger by 0.02 and 0.05 standard deviations in math and ELA, respectively, for schools that did not administer the test online in both years, although the interaction is only statistically significant in ELA. Accounting for this heterogeneity in mode effects reduces the improvement in second-year online scores in columns (1) and (4) by about 0.01 standard deviations in math and 0.02 standard deviations in ELA.

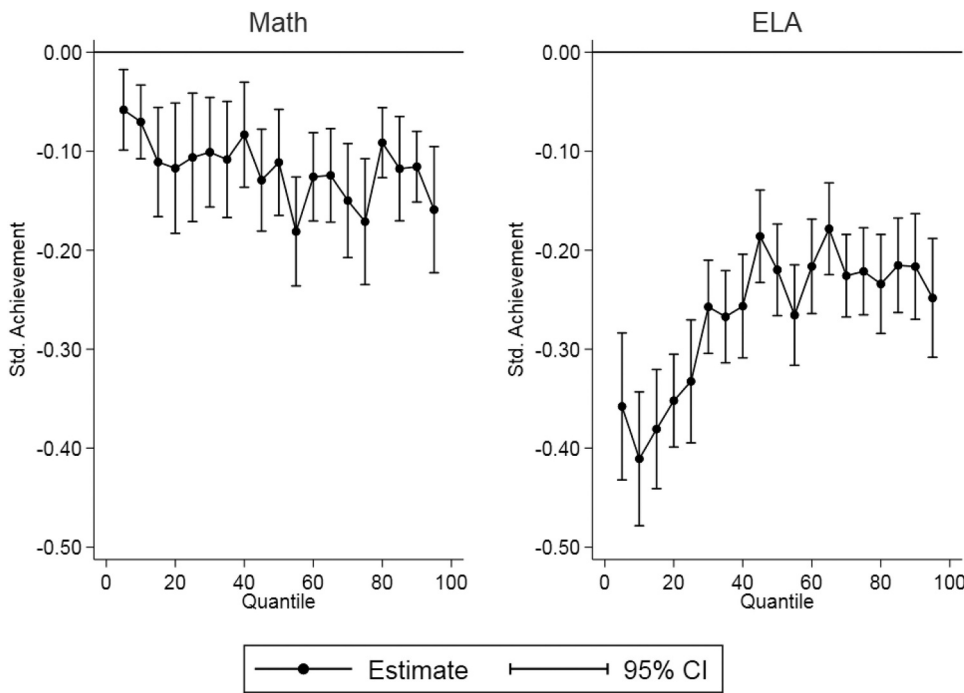


Fig. 6. Quantile treatment effects of computer-based testing

Notes: Estimated quantile treatment effects at each decile of the test score distribution for math and ELA achievement. The sample includes all schools in 2015 and 2016 administering the PARCC assessment in Grades 5 through 8. We estimate quantile treatment effects using the two-step estimator suggested by Firpo (2007), trimming the sample to include observations with propensity scores between 0.1 and 0.9. The first-stage propensity score includes twice-lagged achievement, student race/ethnicity, free-and-reduced-price-lunch status, special education status, limited English proficiency status, grade and year indicators, and each of the means of each of these variables at the school-year and school-grade-year levels. Regressions also include mean school achievement for each year between 2011 and 2014. Standard errors estimated by clustered bootstrap at the school level with 199 iterations.

assessment explains almost all of the reduction in the mode effect for the math test. Although student familiarity may also be important for explaining reductions in the mode effect for ELA, it appears there is also a role for general school improvements in test administration. Nonetheless, students taking online tests scored lower than those taking paper tests in both subjects during the second year of administration.

5.4. Benchmarking effect sizes

Using Table 5 in conjunction with the findings of Lipsey et al. (2012), we compute the size of large test mode effects in terms of measured months of learning. In particular, we use Lipsey et al.'s estimates of annual learning gains in a given grade and subject, to translate the effect sizes in Columns (3) and (6) of Table 5 into months of learning given a 9 month school year.²³ The effect size ranges from 1.3 months (grade 3) to 5.3 months (grade 8) in math and 4.2 months (grade 3) to 11.0 months (grade 7) in ELA.

6. Heterogeneous effects

6.1. Heterogeneity in mode effects by student characteristics

To assess the extent to which certain subgroups are disproportionately impacted by the shift to online testing, we re-estimate Eq. (1) separately for the demographic groups identified in the administrative data. We show the results in Table 6, which indicates re-

latively little variation in mode effects by student demographic group. The main exception to these patterns is for English language learners, whose estimated performance on online tests is statistically significantly lower than the mode effect for the full sample, and special education students, who score lower on online tests than other students in both subjects.²⁴

6.2. Quantile treatment effects

We assess the distributional effects of CBTs by estimating quantile treatment effects (QTE) for the online test. The QTEs describe the effect of online testing on given quantiles of the test score distribution and provide an indication of whether the format of the test affects students differentially by their academic aptitude. We estimate the QTE using the two-step method suggested by Firpo (2007). In the first step, we estimate a propensity score for test mode using the same set of covariates in the lag score model in Eq. (1). Identification therefore rests on the same selection on unobservable assumptions as our primary specification. Recall that this set of covariates includes lags of both school and student outcomes.²⁵ In the second stage, we construct differences in the counterfactual distributions of test scores at each quantile by quantile regression using the propensity scores as weights. The QTE estimator is similar to the more familiar propensity score weighting estimators for mean treatment effects; the primary difference is that the model replaces the weighted difference in outcomes between treated and control units in the second stage with a

²³ An example of this calculation is $9 * [(0.13 + 0.06) / 0.32]$, where 0.13 + 0.06 is the grade 8 math mode effect (Column 3 of Table 5) and 0.32 is Lipsey et al. (2012)) estimate of the annual learning gain between spring of grade 7 and spring of grade 8.

²⁴ For these groups with stronger mode effects, differences between the full sample and subgroup online performance tend to be attenuated in the second year.

²⁵ We use the Stata command `ivqte` to implement the estimation of QTE (Frolich & Melly, 2010), which uses a logit estimator for the propensity score.

difference in sample quantiles constructed using the propensity scores as weights.²⁶

We display the estimated QTE for each vigintile of the test score distribution in Fig. 6. In math, we find little evidence that the effects differ across the achievement distribution. The estimated QTE range from -0.06 at the 5th percentile of the test score distribution to -0.17 at the 75th percentile. However, 12 of the 19 estimates are between -0.08 and -0.13 . We find more significant evidence of variation in the QTE on the ELA test. We estimate larger mode effects on the bottom of the achievement distribution, with QTE below the median ranging from -0.41 at the 10th percentile to -0.18 at the 65th percentile. Above the median, the estimated QTE are all between -0.18 and -0.25 . Choices of test mode therefore appear to have relatively minor distributional consequences for the math test, but more significant effects on the distribution of ELA achievement.

7. Discussion

Using the first two years of the transition to online PARCC in Massachusetts, we find strong evidence that media reports of students scoring lower on CBTs represent true test mode effects that cannot be explained by preexisting trends in the performance of schools that initially moved to online testing or by the prior achievement of students who take the test online. The effect on ELA (-0.24 standard deviations) is larger than the effect on math (-0.10 standard deviations), but both are large changes in measured student performance that should concern education agencies using online PARCC scores for accountability purposes. In addition, the effects in ELA are most pronounced for students at the bottom of the test score distribution.

These estimated effects are large enough to drive substantial changes to a teacher's position in the value added distribution. In a review of 10 studies, Hanushek and Rivkin (2010) find that the average standard deviation of estimated teacher effectiveness is 0.17 standard deviations in math and 0.13 in reading. Combining these estimates with our estimates, the difference in student achievement across modes is greater than the difference between the 40th and 60th percentile of the value added distribution in math and between the 20th and 80th percentile in ELA. When half the schools in a state administer online tests – as was the case in Massachusetts – the mode effects are large enough to move a teacher at the 62nd and 82nd percentiles of the math and ELA value-added distributions, respectively, to the statewide median.²⁷ The distributional consequences of mode effects are even larger in states where the share of online and paper test takers is unequal. Given that states often pool teacher and school value-added measures over a number of years, test mode may have persistent effects on accountability measures. As of 2017, 39 states require that their districts include student growth measures in their evaluation systems, including 13 states where growth must constitute at least 33 percent of a teacher's overall score (NCTQ, 2017).

²⁶ To avoid assigning large weights to individual observations, we trim the sample to include observations with propensity scores between 0.1 and 0.9 using the rule of thumb procedure suggested by Crump, Hotz, Imbens, and Mitnik (2009) for the average treatment effect; however, estimated effects are not sensitive to this choice.

²⁷ We assume that teacher effectiveness is normally distributed, which Goldhaber and Startz (2017) show to be a reasonable approximation. The distribution of teacher value-added in states offering tests both online and on paper forms is then a mixture of two normal distributions with a difference in means equal to the mode effect and mixture weights equal to the share of teachers assigned students in each mode.

Although some policymakers have questioned the consequential use of online assessments during the first year of implementation,²⁸ we also find test mode effects in the second year of online testing, especially in ELA. States or districts that administer PARCC online to some students and on paper to other students should be aware that the paper students will likely score systematically higher, even in the second year.

Our findings indicate that policies that reward or sanction students, teachers, or schools based on student test scores should take test mode effects into account. In Massachusetts, for example, schools that chose to administer PARCC in 2015 and 2016 (whether online or on paper) were subject to a “hold harmless” provision, in which schools would be “held harmless for any negative changes in their [...] accountability and assistance levels”.²⁹ From 2017 forward, as the state transitions to a new assessment, it will use statistical adjustments to correct for mode effects.³⁰ Other states may want to consider similar strategies during the transition phase to online testing.

Researchers may also need to be mindful of mode choice in their use of assessment data, especially when analyzing longitudinal data that span testing transitions. In states that switch all schools from paper to online tests, common among states administering the Smarter Balanced Assessment Consortium tests, for instance, heterogeneity in test mode effects may confound counterfactual time trends in difference-in-differences models where test mode and other student characteristics are unbalanced across treatment and control conditions.³¹ Even in data scenarios like ours – where the state simultaneously tested online and on paper – our results suggesting heterogeneity in the mode effect indicate that controlling for test format alone may be insufficient to eliminate bias from mode effects in some research designs. In both cases, researchers may wish to consider either matching estimators or difference-in-differences methods that permit some relationship between time trends and school characteristics.³²

Finally, there are several reasons that this paper is not intended to recommend for or against transitioning to online testing. First, it is possible that the effects that we measure in the first two years of online PARCC will fade out over time. Second, a full cost-benefit analysis is beyond the scope of this paper and there are likely benefits associated with online testing (e.g., cost of administration, grading, etc.); we are simply measuring one of the costs of transitioning to online testing. And third, given that the spread of online testing is likely inevitable, it is more productive to ask others to think carefully about how to handle test mode effects in both evaluation and research settings.

²⁸ For example, in response to the observed test mode effects in Baltimore County, Russell Brown, the district's chief accountability and performance-management officer, said that “I think it draws into question the validity of the first year's results for PARCC.” Brown, Emma. Report: Kids who took Common Core test online scored lower than those who used paper. *The Washington Post*, 4 February 2016. <https://www.washingtonpost.com/news/education/wp/2016/02/04/report-kids-who-took-common-core-test-online-scored-lower-than-those-who-used-paper>

²⁹ Massachusetts Assessment Decision FAQs. <http://www.doe.mass.edu/mcas/decisionfaq.html>

³⁰ Information obtained through personal correspondence with the state.

³¹ Whether this is the case depends on the extent to which the mode effects we find are a result of the test delivery format or the test scoring procedures that equate the online and paper forms. If the heterogeneity is caused by the latter, then states administering a single examination type may not face the same issues.

³² For instance, if the test mode is observed, then researchers may wish to include models with mode effects and interactions with observed student characteristics. If the state switches entirely to online testing, then heterogeneous mode effects might violate the parallel trends assumptions in difference-in-differences designs.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.econedurev.2018.12.007](https://doi.org/10.1016/j.econedurev.2018.12.007).

Appendix

Fig. A1, Table A1.

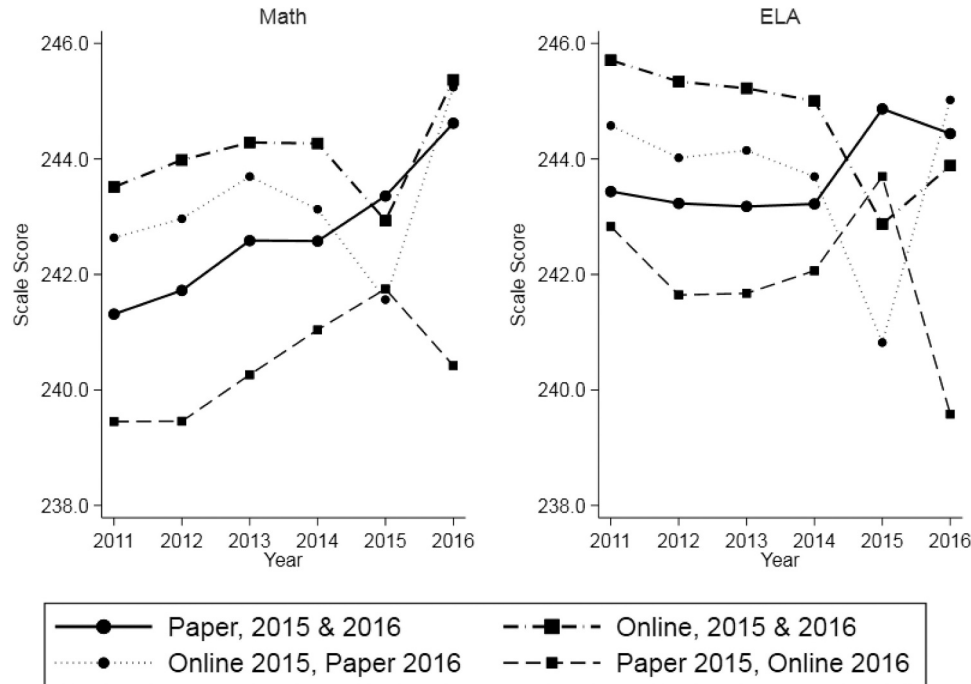


Fig. A1. . Scale scores by year

Notes: Mean test scores by year for schools administering PARCC in both 2015 and 2016, measured on the MCAS scale. PARCC scale scores in 2015 and 2016 are translated to the 2011–2014 MCAS scale by the state.

Table A1

. Analysis sample descriptive statistics.

	Lag score sample (2015–16)			DID sample (2011–16)		
	Excluded	Analysis	Difference	Excluded	Analysis	Difference
Math test	0.08 (0.99)	0.01 (0.99)	−0.08*** (0.00)	0.06 (0.99)	0.01 (0.99)	−0.05*** (0.00)
ELA test	0.10 (0.98)	−0.00 (0.98)	−0.10*** (0.00)	0.07 (0.99)	0.00 (0.99)	−0.06*** (0.00)
Twice-lagged math test	0.06 (0.98)	0.02 (0.99)	−0.04*** (0.00)	0.06 (0.98)	0.02 (0.99)	−0.03*** (0.00)
Twice-lagged ELA test	0.07 (0.99)	0.02 (0.99)	−0.05*** (0.00)	0.07 (0.98)	0.02 (0.99)	−0.05*** (0.00)
Male	0.51 (0.50)	0.51 (0.50)	0.00** (0.00)	0.51 (0.50)	0.51 (0.50)	0.00*** (0.00)
Hispanic	0.15 (0.36)	0.16 (0.37)	0.01*** (0.00)	0.15 (0.36)	0.16 (0.36)	0.01*** (0.00)
African American	0.05 (0.21)	0.10 (0.30)	0.05*** (0.00)	0.05 (0.22)	0.10 (0.30)	0.05*** (0.00)
Asian	0.07 (0.26)	0.06 (0.24)	−0.01*** (0.00)	0.07 (0.25)	0.06 (0.23)	−0.01*** (0.00)
Free lunch eligible	0.30 (0.46)	0.35 (0.48)	0.05*** (0.00)	0.29 (0.45)	0.34 (0.48)	0.05*** (0.00)
Reduced price lunch eligible	0.03 (0.17)	0.04 (0.18)	0.00*** (0.00)	0.04 (0.20)	0.05 (0.21)	0.01*** (0.00)
Limited English proficient	0.05 (0.21)	0.06 (0.24)	0.01*** (0.00)	0.06 (0.24)	0.07 (0.26)	0.01*** (0.00)
Special education	0.16 (0.37)	0.17 (0.37)	0.01*** (0.00)	0.16 (0.37)	0.17 (0.37)	0.01*** (0.00)
Observations	227,981	256,176		1047,732	1193,928	

Notes: Summary statistics for the analytical and full state samples. Columns (1) to (3) include students in grades 5 through 8 in 2015 and 2016. Columns (4) through (6) include students in grades 3 through 8 in 2011–2016. T-test difference in means is provided in the final column with standard errors in parentheses.

* $p < 0.10$.

** $p < 0.05$.

*** $p < 0.01$.

References

- Backes, B., Cowan, J., Goldhaber, D., Koedel, C., Miller, L. C., & Xu, Z. (2018). The common core conundrum: To what extent should we worry that changes to assessments will affect test-based measures of teacher performance? *Economics of Education Review*, 62, 48–65.
- Black, S. E., & Machin, S. (2011). *Housing valuations of school performance. Handbook of the economics of education*, 3, Elsevier 485–519.
- Brown, E. (2016). Report: kids who took common core test online scored lower than those who used paper. *The Washington Post*. February 4 <https://www.washingtonpost.com/news/education/wp/2016/02/04/report-kids-who-took-common-core-test-online-scored-lower-than-those-who-used-paper/>.
- Brown, T., Chen, J., Ali, U., Costanzo, K., Chun, S., & Ling, G. (2015). *Mode comparability study based on spring 2014 field test data*. Washington, DC: Partnership for Assessment of Readiness for College and Careers.
- Cascio, E. U., & Staiger, D. O. (2012). Knowledge, tests, and fadeout in educational interventions. *National Bureau of Economic Research Working Paper No. 18038*.
- Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*, 33(5), 593–602.
- Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1), 187–199.
- Davis, L. L., Strain-Seymour, E., & Gay, H. (2013). Testing on tablets: Part II of a series of usability studies on the use of tablets for K-12 assessment programs. *Pearson White Paper*.
- Duque, M. (2017). *Is there a PARCC mode effect?* Cambridge, MA: Center for Education Policy Research. Retrieved from <https://sdp.cepr.harvard.edu/files/cepr-sdp/files/sdp-fellowship-capstone-parcc-mode.pdf>.
- Farmer, B. (2016, February 20). *The state that pulled the plug on computer testing*. NPR.
- Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1), 276–359.
- Florida Department of Education. (2017). *Florida standards assessments technical report* Florida standards assessments technical report I Florida Department of Education Retrieved from.
- Frolich, M., & Melly, B. (2010). Estimation of quantile treatment effects in Stata. *The Stata Journal*, 10(3), 423–457.
- Goldhaber, D., & Startz, R. (2017). On the distribution of worker productivity: The case of teacher effectiveness and student achievement. *Statistics and Public Policy*, 4(1), 1–12.
- Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, 100(2), 267–271.
- Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., et al. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. National Center for Special Education Research.
- Liu, J., Brown, T., Chen, J., Ali, U., Hou, L., & Costanzo, K. (2016). Mode Comparability Study Based on Spring 2015 Operational Test Data. *Educational Testing Service*.
- Massachusetts Department of Elementary and Secondary Education. (2015). *2014 MCAS and MCAS-alt technical reports*. Malden, MA: Massachusetts Department of Elementary and Secondary Education.
- Massachusetts Department of Elementary and Secondary Education. (2016a). *2015 MCAS and MCAS-alt technical report*. Malden, MA: Massachusetts Department of Elementary and Secondary Education.
- Massachusetts Department of Elementary and Secondary Education. (2016b). *Representative samples and PARCC to MCAS concordance studies*. Malden, MA: Massachusetts Department of Elementary and Secondary Education.
- Michigan Department of Education. (2016). *2016–2017 guide to state assessments*. Michigan Department of Education.
- Murray, M., Sherburn, R., & Pérez, J. (2007). Information technology literacy in the workplace: A preliminary investigation. *Information Technology*, 3(1), 132–136.
- NCTQ. (2017). National Council on Teacher Quality https://www.nctq.org/dmsView/NCTQ_2017_State_Teacher_Policy_Yearbook.
- New York State Education Department. (2016). New York State Education Department.
- New York State Education Department. (2017). New York State Department of Education. Retrieved from <http://www.p12.nysed.gov/assessment/ei/2017/2017-faq-for-parents-3-8-tests.pdf>.
- North Carolina Department of Public Instruction. (2017). *The North Carolina testing program 2017–18* North Carolina Department of Public Instruction. Retrieved from <http://www.ncpublicschools.org/docs/accountability/generalinfo/1718testovervw.pdf>.
- O'Donnell, P. (2016). School districts got A grades on paper tests, but F grades online. *survey shows*. cleveland.com, March 8 https://www.cleveland.com/metro/index.ssf/2016/03/school_districts_got_a_grades_on_paper_tests_but_f_grades_online_survey_shows.html.
- Parhizgar, S. (2012). Testing and Technology: Past, Present and Future. *Theory and Practice in Language Studies*, 2(1), 174–178.
- Pearson. (2016). *Final technical report for 2015 administration*. Pearson.
- Pearson. (2017). *Final technical report for 2016 administration*. Pearson.
- Richtel, M. (2012, May 29). *Wasting time is new divide in digital era*, A1.
- Snyder, T. D., de Brey, C., & Dillow, S. A. (2018). *Digest of education statistics 2016 (No. NCES 2017-094)*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from <https://nces.ed.gov/pubs2017/2017094.pdf>.
- Tuotti, G. (2014, June 19). Despite issues, PARCC gets good grades. *Taunton Daily Gazette*. <http://www.tauntongazette.com/article/20140619/News/140615326>.
- Ujifusa, A. (2015, August 12). N.Y. Opt-Out Rate Hits 20 Percent on Common-Core Tests. *Education Week*. http://blogs.edweek.org/edweek/state_edwatch/2015/08/ny_opt-out_rate_hits_20_percent_on_common-core_tests.html.
- Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement*, 67(2), 219–238.
- White, S., Kim, Y. Y., Chen, J., & Liu, F. (2015). *Performance of fourth-grade students in the 2012 NAEP computer-based writing pilot assessment: Scores, text length, and use of editing tools (No. NCES 2015-119)*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from <https://nces.ed.gov/nationsreportcard/subject/writing/pdf/2015119.pdf>.