CALDER

NATIONAL CENTER *for* ANALYSIS *of* LONGITUDINAL DATA *in* EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

AIR

Is Online a Better Baseline? Comparing the Predictive Validity of Computer- and Paper-Based Tests

Ben Backes
James Cowan

WORKING PAPER No.241-0820 • August 2020

Is Online a Better Baseline? Comparing the Predictive Validity of Computer- and Paper-Based Tests

Ben Backes
American Institutes for Research/CALDER

James Cowan
American Institutes for Research/CALDER

Contents

Acknowledgments

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 Elana McDermott, Bob Lee, and Dan Goldhaber for their feedback. All opinions expressed in this paper are those of the authors and do not necessarily reflect the views of our funders or the institutions to which the authors are affiliated.

CALDER working papers have not undergone final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication. Any opinions, findings, and conclusions expressed in these papers are those of the authors and do not necessarily reflect the views of our funders.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street NW, Washington, DC 20007
202-403-5796 • www.caldercenter.org

Is Online a Better Baseline? Comparing the Predictive Validity of Computer- and Paper-Based Tests

Ben Backes, James Cowan

CALDER Working Paper No. 241-0820
August 2020

Abstract

Prior work has documented a substantial penalty associated with taking the Partnership for Assessment of Readiness for College and Careers (PARCC) online relative to on paper (Backes & Cowan, 2019). However, this penalty does not necessarily make online tests less useful. For example, it could be the case that computer literacy skills are correlated with students' future ability to navigate high school coursework, and thus more predictive of later outcomes. Using a statewide implementation of PARCC in Massachusetts, we test the relative predictive validity of online and paper tests. We are unable to detect a difference between the two and in most cases can rule out even modest differences. Finally, we estimate mode effects for the new Massachusetts statewide assessment. In contrast to the first years of PARCC implementation, we find very small mode effects, showing that it is possible to implement online assessments at scale without large online penalties.

1. Introduction

Standardized testing occupies a central role in the measurement of student learning. In many states, the results of standardized assessments support teacher evaluation, school accountability determinations, student graduation, or the distribution of school resources. Recent years have seen a rapid transition to online testing, and the vast majority of states' test administrations are partially or fully online (Olson, 2019). Proponents of online testing list stated benefits of online testing including faster turnaround time, cheating detection, and more flexibility with item design (Garland, 2012). But whether these features of online tests translate into improved measurement of student learning outcomes remains an open question.

Online tests support new question types that may better measure student learning. Some research has found that state accountability tests implemented under the No Child Left Behind Act, which often relied heavily on multiple-choice questions, did not adequately test higher order thinking skills (Yuan & Le, 2012). Online platforms, by contrast, offer interactive question types that may better test students' problem-solving skills. Measuring deeper learning concepts through these technologically enhanced questions was one motivation for adopting online tests as part of the transition to the Common Core State Standards (Darling-Hammond, 2017; Slover & Muldoon, 2018). But it is not clear empirically that tests of critical thinking skills are more highly correlated with long-run learning outcomes than performance on other standardized assessments (National Research Council, 2011).[1]

Online tests also could measure, in part, students' familiarity with computers. The tests require students to navigate computer programs and use editing tools to respond to open-ended questions. White et al. (2015) and Sandene et al. (2005) have found that prior computer experience predicted student performance in pilots of the National Assessment of Educational Progress in mathematics and English language arts (ELA). Goldberg and Pedulla (2002) documented similar patterns for some online versions of the Graduate Record Examinations. And Backes and Cowan (2019) found that taking the online version of the Partnership for Assessment of Readiness for College and Careers (PARCC) assessment reduced measured student achievement. If computing skills are important inputs into academic success, then tests that measure these skills may better predict students' academic outcomes. However, the evidence on computers as academic inputs is relatively mixed. Bulman and Farlie (2016) review the literature on home and school computer use and conclude there is no strong evidence of positive effects on measurable student outcomes.

With the prevalence of online testing rising over time, it is important to understand how the administration of tests on computers affects measured student learning. In this paper, we measure the relative predictive validities of online versus paper PARCC scores. In particular, we use 2 years in which the state of Massachusetts administered PARCC statewide, with approximately half of students taking PARCC online and half on paper forms. In previous work (Backes & Cowan, 2019), we estimated online penalties of about 0.25 standard deviations in math and 0.10 standard deviations in ELA, with larger mode effects for students at the bottom of the test score

[1] Another potential advantage of online tests is the possibility of adaptive testing. However, because PARCC is not an adaptive test, we do not include adaptive testing in our discussion of the differences between paper- and computer-based tests.

distribution in ELA. Despite the lower scores of students taking the test online, we are unable to find any differences in the ability of test scores to predict student outcomes 2 years into the future. We also provide additional evidence about the effects of test mode on measured student achievement. Using data from another test transition in Massachusetts, we find significantly lower mode effects for online tests than on the PARCC assessment, which we attribute to post hoc adjustments to the testing scale.

2. State Context

This study covers 2 years (2015 and 2016) in which Massachusetts administered paper and online versions of the PARCC assessment simultaneously to different sets of schools. The paper versions of the PARCC assessment were adapted from the online forms, and they used a similar set of items. That said, the online versions of the test included some interactive questions not present in the paper version, meaning that the paper and online versions were not exactly equivalent. Both modes included a subset of linked items to facilitate the reporting of student scores on a common scale (Educational Testing Service, Pearson, & Measured Progress, 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.

A prior study has found strong evidence that the scores were not comparable across modes, showing a negative causal effect in both math and ELA scaled scores of taking the test online rather than on paper (Backes & Cowan, 2019). Of the 2 years in the study, the largest test mode penalties were found in the first year but continued into the second year. This penalty did not appear to be explained by differences in school context—the schools that switched to online testing were disproportionately high achieving—and the analysis passed a series of placebo tests. For example, schools that switched to the online PARCC assessment and saw a drop in math and ELA scores did not see a corresponding drop in grades 4 and 8 science, which was administered on paper to all schools throughout the time period.

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 the year under discussion). Until 2014, all districts used the Massachusetts Comprehensive Assessment System (MCAS), which was administered on paper. In 2015 and 2016, districts chose between MCAS and the new PARCC assessment, with three districts having a mix of online and paper tests.[2] About 72% of elementary or middle schools in our sample administered PARCC in either 2015 or 2016. PARCC districts had the additional option of offering the test online or on paper. Of those schools administering PARCC in either 2015 or 2016, 57% administered the test online at least once. In 2017, Massachusetts switched away from PARCC to a new assessment, MCAS 2.0,

[2] 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.

which also consisted of a mix of online and paper offerings. In a later section, we show that the online penalty associated with MCAS 2.0 was substantially smaller than that for PARCC.

3. Data and Summary Statistics

The primary explanatory variables of interest are student test scores on the PARCC assessment—some online and others on paper—in grades 3 through 8 in the 2014–15 and 2015–16 school years (as described later, some of our models will restrict our sample to grades 7 and 8 in order to estimate later high school outcomes). We then link these data to student outcomes 2 years forward, in 2017 and 2018. We use longitudinal student achievement data that have been linked to student data in the Student Information Management System by the state, which includes information on students' enrollment status, demographics, transcripts, and program participation. We limit our sample to schools that administered PARCC in both 2015 and 2016 to ensure that achievement is measured on a common scale in each year.[3]

Summary statistics for the sample are shown in Table 1. Students in schools that would switch to online testing had higher math and ELA scores in 2014, the final year before PARCC was adopted. Average MCAS achievement in online schools prior to the implementation of PARCC was about 0.08 standard deviations higher in math and 0.09 standard deviations higher in ELA than in paper schools. In addition, students in schools that switched to online testing are less likely to be Hispanic or Black, and much less likely to be eligible for free or reduced-price lunch (FRL).

Due to the differences in school characteristics across the paper and online schools, we construct an additional sample that reweights students in the paper sample based on an estimated propensity score for having an online test (Busso, DiNardo, & McCrary, 2014). This step is potentially important because if we were to find differences in the predictive power of tests for online versus paper, it would be difficult to know whether it were due to differential predictive power by test mode or the relationship between future outcomes and test scores being stronger for certain types of students. For example, perhaps math test scores are more predictive of future math grade point average (GPA) for more advantaged students. If this were the case, we might expect to see more predictive power for online PARCC math scores even in the absence of a true difference in predictiveness due to the sample being more advantaged.

To construct alternative weights for the paper sample, we estimate a logistic regression predicting online status with the following variables: math and ELA scores in 2014 at the grade and school levels, total students at the grade and school levels, percentage male at the grade and school levels, percentage FRL eligible at the grade and school levels, percentage of each race at

[3] 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 in which more than 5% 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, 2016). 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.

the grade and school levels, and individual indicators for limited English proficient, special education, FRL, male, and race. Results are suppressed for brevity (but available upon request). Along with the obvious predictors gleaned from differences in Table 1, total students at the grade and school levels are predictors of online testing, as some of the large districts were the ones that moved to online testing. As the final column of Table 1 shows, observable characteristics are balanced between the paper and online samples. Although the outcome variables are not included in the propensity score model, they are also generally balanced between test modes. Students in the reweighted paper sample have 0.45 fewer unexcused absences, but differences in all other outcomes are small and not statistically different than zero.

To offer a preview of our results using GPA as an example, Figure 1 shows a scatterplot of math test scores from time *t* versus core-subject high school GPA in time *t + 2* for a sample of students whose grade in time *t* would place them on track to reach high school by *t + 2* (i.e., seventh and eighth graders in 2015 and eighth graders in 2016). As shown in the figure, there is a similar, positively sloped relationship between GPA and test scores for both the online and paper samples. When one separately regresses GPA in 2 years on a cubic function of math scores on the paper versus online samples, the coefficients on the linear, quadratic, and cubic terms are nearly identical in both the online and the paper regressions. However, the online test penalty means that for a given math test score, the prediction of high school GPA is higher for the online sample than for the paper sample. In this example, the constant is 0.17 GPA points higher in the online regression than in the paper regression (i.e., the slopes of the curves are similar, but the online curve is shifted upward). Results are similar for subject-specific GPA (i.e., math scores predicting math GPA), for ELA, and for using both math and ELA to predict outcomes, as we discuss later.

Figure 2 suggests that while the curve relating test scores to GPA is shifted for the online sample, the overall relationship is similar. That the difference between online and paper test scores appears to be a linear transformation suggests the relative abilities of online and paper tests to predict future GPA are similar. Next, we present more evidence that this is indeed the case.

4. Results

Our main results are shown in Table 2. The table is organized into sections, with columns (1) through (3) showing the correlation between a given outcome and math test scores, columns (4) through (6) indicating the corresponding correlations for ELA and outcomes, and columns (7) through (9) showing the square root of the adjusted *r*-squared for a regression of the outcome on cubic polynomials in math and ELA scores jointly.

Within each section of the table, we show the correlation for students who took PARCC online (e.g., column 1), for those who took PARCC on paper (column 2), and the reweighted PARCC paper sample (column 3). We also test for differences between the paper and online correlations using a percentile blocked bootstrap at the school level.

Before getting into the specific results, we argue the overarching message from the findings is that there are no meaningful differences between paper and online tests in their ability to predict later outcomes. The differences in the correlations are generally small and mostly not statistically significant. The few that are statistically significant are also not directionally consistent.

General outcomes. The first group of outcomes includes test scores in 2 years, as well as non-test outcomes such as absences and suspensions. For this group of outcomes, the correlations and adjusted *r*-squares are nearly identical across test modes. Unsurprisingly, test scores are positively correlated with future test scores and negatively correlated with missing school in the future (through either absences or suspensions). We do find some evidence that online tests are more predictive of future ELA tests in the reweighted sample. In each model, the correlation is about 0.01 higher for online tests and statistically significant at the 5% level. We do not find consistent differences for any of the other outcomes. For total absences, the 95% confidence interval for the difference in *R* spans from about −0.020 to 0.015 for the reweighted sample (available from authors upon request), meaning we can rule out even moderate changes in predictive power.

Predicting specific later test scores. A few general patterns are worth noting. First, unsurprisingly, same-subject tests better predict future scores than tests in other subjects. In addition, the relationship between current and future math scores is stronger than the relationship between current and future ELA scores. Finally, paper tests do a slightly better job of predicting later paper test scores. For example, the root *r*-squared of a regression on grade 10 paper math test scores on grade 8 paper math and ELA scores is 0.84, compared to 0.82 for students taking the grade 8 test online.[4] The results for future online tests are less clear; however, we do find that online ELA tests better predict future performance on online ELA tests than paper ELA tests. We do not find a similar result for math, which may partially reflect the more extensive use of editing tools on online ELA tests.

Predicting high school outcomes. The correlations between test scores and high school GPA range from about 0.50 to 0.60. Looking across test modes, we do not find much evidence that online tests better predict academic performance than paper tests. The coefficients tend to be higher for the online tests, but the correlations (columns 1 versus 2 and columns 4 versus 5) and adjusted *r*-squared values (columns 7 versus 8) are quite similar. We find a single result significant at the 10% level, but none of the other differences are statistically significant, although in some cases we are unable to rule out moderate differences of up to about 0.05. Continuing on to the Advanced Placement® (AP®) tests, the correlations and *r*-squared values are generally low for both the online and paper samples. In contrast to the GPA results, the correlations for paper tests tend to be higher, but the differences are small and none are statistically significant.

5. Estimation of Mode Effects in the Newer MCAS 2.0

In this section, we use a more recent test transition in Massachusetts as a piece of evidence about whether scale score penalties from online tests are always large, as they were for PARCC. In 2017 and 2018, Massachusetts switched to a new test, which we refer to as Next-Generation

[4] For another example of high correlations between old and new test regimes across multiple states, see Backes and colleagues (2018).

MCAS. The math and ELA MCAS were administered online in grades 3 through 8 beginning in 2017, with optional paper administrations in some grades.[5] In 2017, the state required the grades 4 and 8 MCAS to be offered online, and 95% of fourth and eighth graders took the online MCAS. In addition, 43% of primary school students in other grades took the test online. In 2018, Massachusetts required online administration of the grades 5 and 7 MCAS. Nearly 90% of students took the test online that year, with grade 3 (62%) and grade 6 (78%) lagging behind the other grades.

Our research design relies on the rollout of the online MCAS. We use the entire panel of test scores from 2011 through 2018 and estimate models comparing changes in test scores in grades and schools switching to online tests to those remaining on paper tests. Formally, we estimate

$$Y_{igst} = X_{igst}\beta + Online_{gst}\delta + \alpha_{gs} + \lambda_t + \epsilon_{igst} \qquad (1)$$

where α_{gs} is a grade-by-school fixed effect and λ_t is a test-by-year effect.[6] The research design incorporates two sources of variation in test mode: differences across grades in the application of the MCAS online testing mandate and differences within grades between schools voluntarily switching to the online format. We plot the effects of online MCAS testing relative to the year prior to implementation in Figure 1. We do not find any statistically significant evidence of differential pretreatment trends among the online testing cells.

We additionally estimate models that isolate each of these distinct sources of variation in test mode. If the online testing mandates are associated with other changes to the test across grades, then we may conflate test difficulty with mode effects. We therefore replace the year effects with year-by-grade effects. Because these models remove variation within grades and years in online testing status, they compare changes in test scores among schools and grades voluntarily switching to the online test format to those remaining on the paper test. We then instrument for online testing status with an indicator for whether the state required the test be given online. Because this approach only uses variation in mode generated by the online testing mandate, it does not rely on schools' voluntary adoption of online tests for identification.

Results are shown in Table 3. In the first column, we estimate the PARCC mode effect using data from the 2011–2016 school years.[7] We estimate mode effects of about −0.08 in math and −0.21 in ELA on the PARCC assessment. These estimates are slightly smaller than the estimates from Backes and Cowan (2019), using data on only PARCC adopters. Using the specification in

[5] Schools could obtain a waiver to administer the MCAS tests on paper, but waivers were granted only for cases in which online testing was not possible at a school (e.g., due to lack of technology required) or if every student in the school would take a paper-based test due to testing accommodations. For more information, see http://www.doe.mass.edu/news/news.aspx?id=25169.

[6] In earlier work (Backes & Cowan, 2019), we found that estimates from this fixed-effects model are similar to one using prior achievement at the student level. However, the fixed-effects model is advantageous in that it does not require the inclusion of students' lagged test scores; thus, it can be estimated on a larger pool of students.

[7] Backes and Cowan (2019) use a sample of schools administering the PARCC in both 2015 and 2016. Given the interest in mode effects on the MCAS, we include MCAS schools in this study.

Equation (1), which incorporates variation from voluntary adoption and online testing mandates, we estimate mode effects of −0.03 in math and −0.02 in ELA. The confidence intervals are precise enough to rule out mode effects of about −0.05 in math and −0.04 in ELA, suggesting that the mode effects on MCAS 2.0 were considerably smaller than on the PARCC assessment.

In column 3, we include grade-by-year effects to focus on effects of online tests for schools voluntarily switching to the online format. The estimates are nearly identical to those in column 2. We repeat these analyses in columns 4 and 5 using schools that administered the MCAS in both 2015 and 2016. The point estimates are not statistically significant for either test, although they are quite similar in magnitude on the ELA test. Notably, because these schools did not offer online testing prior to 2017, these findings indicate that the reduction in the online penalty from PARCC to Next-Generation MCAS could not be driven by prior familiarity with online testing. Finally, in column 6, we instrument for online testing with an indicator for whether the grade was subject to a testing mandate in the given year. Although not statistically significant, the point estimates are similar to the baseline specification for ELA and smaller in math.

Overall, online mode effects appear to be significantly smaller on the Next-Generation MCAS than for the PARCC assessments. We estimate mode effects of about −0.02 standard deviations in ELA, and between 0.00 and −0.03 in math. There appear to be two possible explanations for why there is a major online penalty for PARCC and not for MCAS. The first is that the online version of PARCC was genuinely harder due to item design. Although we cannot assess this possibility directly, we note that the Next-Generation MCAS did license some items from the PARCC item bank (Massachusetts Department of Elementary and Secondary Education, 2017, 2018). The second is that the equating between the online and paper versions of the tests was conducted differently in the two assessment systems. PARCC equated online and paper forms using a set of linked items that were common across the tests (Educational Testing Service et al., 2016; Pearson, 2017). The Massachusetts Department of Elementary and Secondary Education used a set of linked items, but also subsequently adjusted the scale scores using a sample of test takers matched on observable characteristics, including prior test scores. Notably, the mode effects after equating based on linked items but prior to adjustment were similar in magnitude to estimates from the PARCC assessments (Backes & Cowan, 2019; Massachusetts Department of Elementary and Secondary Education, 2017, 2018).

6. Lessons Learned

Online tests are perceived as an improvement over paper tests, offering “reduced testing time and cost, quicker results, greater access for English language learners and students with disabilities, individualized questions, automated scoring, and technology-enhanced performance tasks that can assess more complex skills” (Olson, 2019). In this study, we indeed find suggestive evidence that online tests measure distinct skills relative to paper tests: in our sample, online tests better predict future online tests (even on a different assessment) than paper tests (and vice versa). However, we do not find consistent evidence that online tests better predict other future outcomes. In particular, we do not find that either test mode produces better predictions of high school GPA or absenteeism. We also find that the online testing penalty found in prior studies can be mitigated in large-scale testing systems. Although both the PARCC and Next-Generation MCAS exhibited mode effects after equating scores using a sample of matched items, the mode effects on the Next-Generation MCAS were significantly reduced through matching procedures.

References

Backes, B., & Cowan, J. (2019). Is the pen mightier than the keyboard? The effect of online testing on measured student achievement. *Economics of Education Review*, *68*, 89–103.

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.

Barnum, M. (2018, April 10). Did computer testing muddle this year's NAEP results? Testing group says no; others are unconvinced. *Chalkbeat*. Retrieved from https://www.chalkbeat.org/posts/us/2018/04/10/did-computer-testing-muddle-this-years-naep-results-testing-group-says-no-others-are-unconvinced/

Bulman, G., & Fairlie, R. W. (2016). Technology and education: Computers, software, and the internet. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 5, pp. 239–280). Elsevier B.V.

Busso, M., DiNardo, J., & McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *The Review of Economics and Statistics, 96*(5), 885–897.

Darling-Hammond, L. (2017*). Developing and measuring higher order skills: Models for state performance assessment systems*. Learning Policy Institute. Retrieved from https://learningpolicyinstitute.org/sites/default/files/product-files/Models_State_Performance_Assessment_Systems_REPORT.pdf

Educational Testing Service, Pearson, & Measured Progress. (2016). *Final technical report for 2015 administration*.

Garland, S. (2012, June 26). Online testing revolution comes to schools. *The Washington Post*. Retrieved from https://www.washingtonpost.com/local/education/online-testing-revolution-comes-to-schools/2012/06/26/gJQACaIJ4V_story.html

Goldberg, A. L., & Pedulla, J. J. (2002). Performance differences according to test mode and computer familiarity on a practice Graduate Record Exam. *Educational and Psychological Measurement, 62*(6), 1053–1067.

Massachusetts Department of Elementary and Secondary Education. (2017). *2017 Next-Generation MCAS and MCAS-Alt technical report*. Malden, MA: Massachusetts Department of Elementary and Secondary Education.

Massachusetts Department of Elementary and Secondary Education. (2018). *2018 Next-generation MCAS and MCAS-Alt technical report*. Malden, MA: Massachusetts Department of Elementary and Secondary Education.

National Research Council. (2012). *Assessing 21st century skills: Summary of a workshop*. Washington, DC: National Academies Press.

Olson, L. (2019). The New Testing Landscape. *FutureEd.* https://www.future-ed.org/wp-content/uploads/2019/09/FutureEdTestingLandscapeReport.pdf

Pearson. (2017). *Final technical report for 2016 administration*.

Sandene, B., Horkay, N., Bennett, R. E., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online Assessment in Mathematics and Writing: Reports from the NAEP Technology-*

Based Assessment Project, Research and Development Series. NCES 2005-457. National Center for Education Statistics.

Slover, L., & Muldoon, L. (2018, September 6). How the Common Core changed standardized testing. Retrieved from https://www.educationnext.org/common-core-changed-standardized-testing/

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. Working Paper Series. NCES 2015-119. *National Center for Education Statistics*.

Yuan, K., & Le, V.-N. (2012). *Estimating the percentage of students who were tested on cognitively demanding items through the state achievement tests* (No. WR-967-WFHF). RAND Corporation.

Tables and Figures

Figure 1

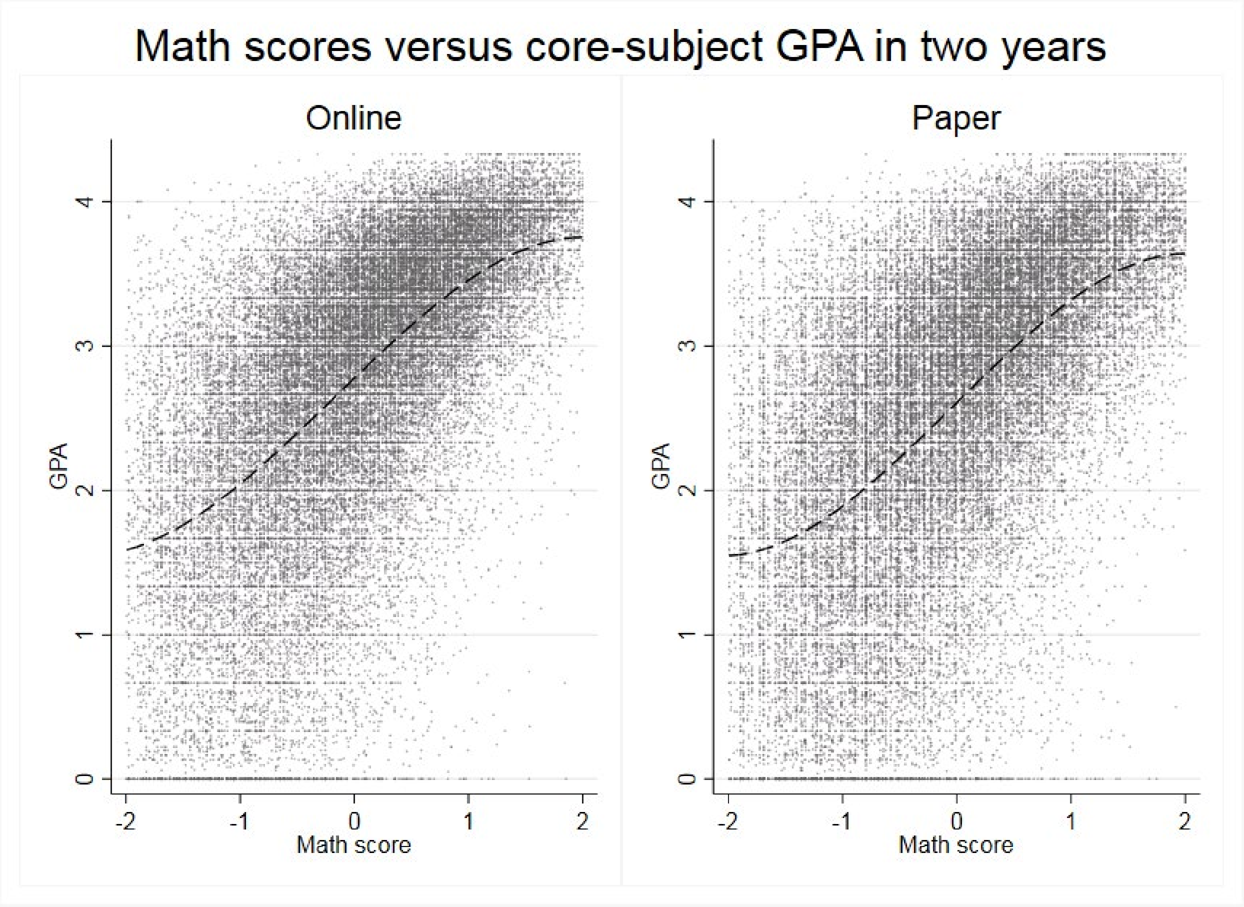

Figure 2. Test Mode Effects by Year Relative to Implementation

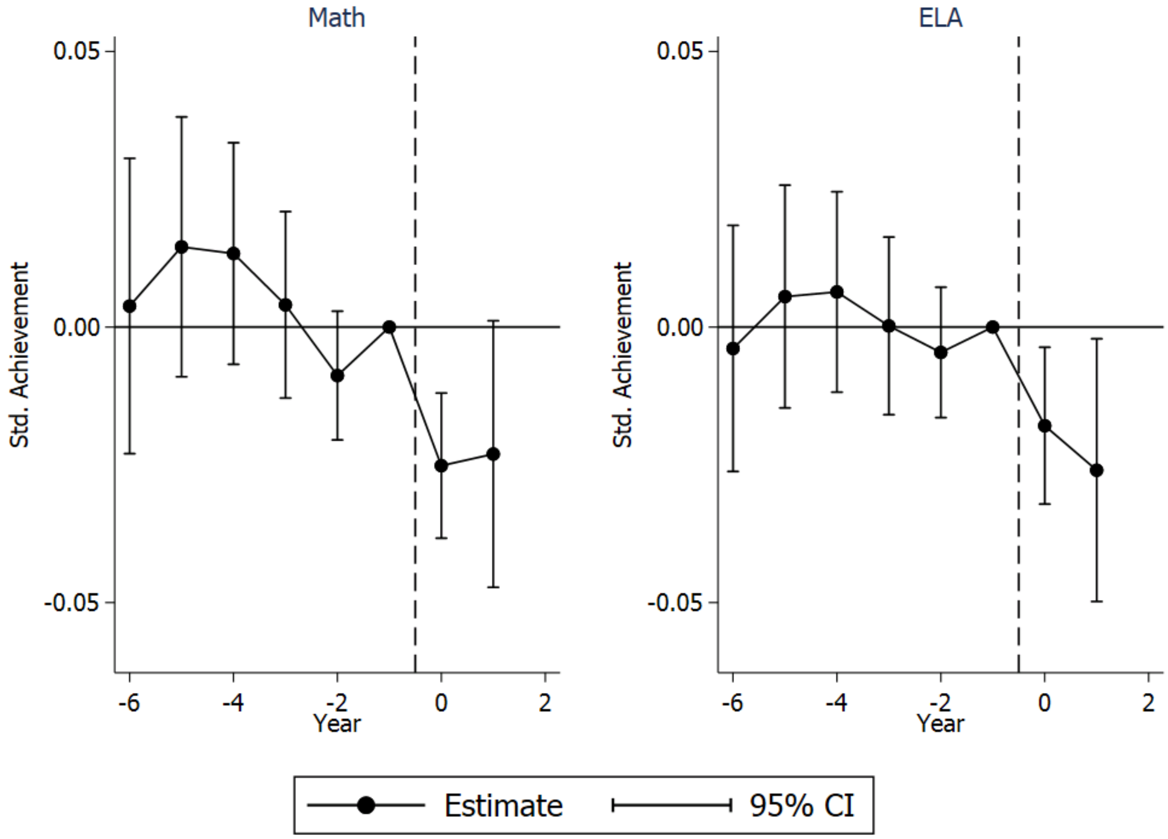

Table 1: Summary statistics

	Online	Paper	Difference	Reweighted Difference
Student Characteristics				
Male	0.51	0.51	0.00	-0.00
Hispanic	0.14	0.19	-0.05***	-0.00
African American	0.08	0.12	-0.04***	-0.00
Asian	0.06	0.06	-0.00	-0.00
Free lunch eligible	0.32	0.40	-0.08***	0.00
Reduced price lunch eligible	0.04	0.03	0.01**	0.00
Limited English proficient	0.06	0.09	-0.03***	0.00
Special education	0.16	0.17	-0.01**	-0.00
School mean math score (2014)	0.04	-0.03	0.08**	-0.02
School mean ELA score (2014)	0.05	-0.04	0.09**	-0.01
Observations	181,029	191,855		
Student Outcomes (t+2)				
ELA test	0.07	-0.02	0.09***	0.01
Math test	0.07	-0.03	0.10***	0.01
Observations	144,794	153,911		
Total absences	8.25	8.80	-0.55***	-0.04
Unexcused absences	4.48	4.58	-0.11	0.45*
Days suspended	0.08	0.10	-0.01	-0.01
Observations	181,029	191,855		
High School Outcomes (Grades 7-8)				
Core subject GPA (t+2)	2.78	2.62	0.16**	0.01
Math GPA (t+2)	2.63	2.47	0.16**	0.01
ELA GPA (t+2)	2.79	2.62	0.17**	0.02
Observations	38960	37758		
AP tests taken by grade 10	0.04	0.04	0.00	-0.00
AP tests passed by grade 10	0.02	0.02	0.00	-0.00
AP math tests taken by grade 10	0.01	0.00	0.00	0.00
AP math tests taken by grade 10	0.00	0.00	0.00	0.00
Observations	44,837	44,593		

Note: Students categorized as online sample if the school they are located in during a given year was online in either 2015 or 2016. The reweighted difference column contains the difference between the online and reweighted paper sample based on the estimated propensity score. Standard errors on differenced (not shown) clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Predictive Correlations between test scores and later outcomes

	Math Test			ELA Test			Combined		
	Online	Paper (no weight)	Paper (re-weighted)	Online	Paper (no weight)	Paper (re-weighted)	Online	Paper (no weight)	Paper (re-weighted)
Future Student Outcomes (t+2)									
ELA test	0.68	0.68	0.67**	0.77	0.77	0.76**	0.79	0.79	0.78**
Math test	0.82	0.82	0.82	0.70	0.70	0.69	0.83	0.83	0.83
Observations	144,794	153,911	153,911	144,794	153,911	153,911	144,794	153,911	153,911
Total absences	-0.22	-0.24	-0.22	-0.18	-0.21**	-0.19	0.23	0.24	0.23
Unexcused absences	-0.23	-0.24	-0.22	-0.20	-0.22	-0.20	0.23	0.25	0.23
Days suspended	-0.07	-0.07	-0.07	-0.08	-0.07	-0.07	0.09	0.08	0.08
Observations	181,029	191,855	191,855	181,029	191,855	191,855	181,029	191,855	191,855
High School Outcomes (t+2, restricted to grades 7-8 in t)									
Core subject GPA	0.61	0.60	0.59	0.61	0.59	0.59	0.65	0.64	0.63
Math GPA	0.56	0.55	0.53	0.51	0.50	0.48*	0.58	0.57	0.55
ELA GPA	0.52	0.50	0.50	0.54	0.53	0.53	0.57	0.56	0.55
Observations	38,960	37,758	37,758	38,960	37,758	37,758	38,960	37,758	37,758
AP tests taken by grade 10	0.16	0.18	0.19	0.16	0.17	0.17	0.20	0.22	0.22
AP tests passed by grade 10	0.17	0.18	0.19	0.16	0.17	0.17	0.23	0.25	0.24
AP math tests taken by grade 10	0.09	0.09	0.10	0.07	0.07	0.07	0.15	0.14	0.14
AP math tests taken by grade 10	0.10	0.10	0.10	0.08	0.07	0.07	0.18	0.16	0.16
Observations	44,837	44,593	44,593	44,837	44,593	44,593	44,837	44,593	44,593
Online Test (t+2, restricted to grade 6 in t)									

ELA test	0.70	0.70	0.68**	0.79	0.79	0.78*	0.81	0.81	0.79*
Math test	0.84	0.84	0.83	0.71	0.72	0.70	0.85	0.85	0.85
Observations	30,286	29,019	29,019	30,286	29,019	29,019	30,286	29,019	29,019
Paper Test (t+2, restricted to grade 8 in t)									
ELA test	0.67	0.68*	0.68	0.76	0.78**	0.78*	0.78	0.80**	0.80**
Math test	0.80	0.82***	0.81*	0.68	0.69	0.68	0.82	0.84***	0.84**
Observations	25,042	26,785	26,785	25,042	26,785	26,785	25,042	26,785	26,785

Notes: Columns (1) through (6) display the correlation between the outcome listed in the leftmost column and either math (columns 1-3) or ELA (columns 4-6) scores. Columns (7) through (9) display the square root of the adjusted R2 from a regression of the variable in the leftmost column on cubic polynomials in math and ELA. Estimates in columns (3), (6), an (9) reweight the paper sample to match the online sample on observable characteristics. Weighting is based on the estimated propensity score. Each cell shows relationship between test score(s) in time t and a later outcome in time t+2, with the exception of AP tests, which are for the time period labeled. Tests of statistical significance are based on the difference between the paper and online samples and are conducted using a clustered percentile bootstrap at the school level.

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

Table 3: Online test penalties for PARCC and MCAS 2.0

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Math						
Online PARCC	-0.08***	-0.07***	-0.07***			-0.07***
	(0.01)	(0.01)	(0.01)			(0.01)
Online MCAS		-0.03***	-0.03***	-0.01	-0.02	-0.01
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
N	1,999,649	2,641,125	2,641,125	766,198	766,198	2,641,125
Panel B. ELA						
Online PARCC	-0.21***	-0.21***	-0.20***			-0.21***
	(0.01)	(0.01)	(0.01)			(0.01)
Online MCAS		-0.02**	-0.02**	-0.02	-0.02	-0.02
		(0.01)	(0.01)	(0.01)	(0.02)	(0.01)
N	1,999,010	2,639,755	2,639,755	765,719	765,719	2,639,755
Grade-Year FE	Y		Y		Y	
MCAS Only				Y	Y	

Notes: each regression includes school-by-grade fixed effects, controls for student race/ethnicity, FRL 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.