Center for Education Policy Research
HARVARD UNIVERSITY

STRATEGIC DATA PROJECT

SDP FELLOWSHIP
CAPSTONE REPORT 2016

Is There a PARCC Mode Effect?

Matthew Duque, Baltimore County Public Schools

Executive Summary

While the majority of state-mandated testing in K–8 was administered online in the 2015–2016 school year, there is some evidence that results from online tests are not comparable to those from traditional paper and pencil tests. Using data from a large school district, the present study examined the extent to which student performance on the first year of the Partnership for Assessment of Readiness for College and Careers (PARCC) exam was related to test mode. Controlling for student and school characteristics, results indicate that students who tested on paper scored substantially higher than students who tested online, suggesting that the online format tests more than content knowledge.

Strategic Data Project Fellowship Capstone Reports

Strategic Data Project (SDP) Fellows author capstone reports to reflect the work that they led in their education agencies during the two-year program. The reports demonstrate both the impact fellows make and the role of the SDP network in supporting their growth as data strategists. Additionally, they provide recommendations to their host agency and may serve as guides to other agencies, future fellows and researchers seeking to do similar work. *The views or opinions expressed in this report are those of the authors and do not necessarily reflect the views or position of the Center for Education Policy Research at Harvard University.*

In the 2015–2016 school year, the majority of state-mandated testing in K–8 was administered online (Ed Tech Strategies, 2015). Online testing offers several advantages over traditional paper and pencil tests, including new ways to assess student understanding, time and cost savings in scoring and score reporting, and increased security. However, some studies have found online exam scores to be incomparable to paper versions of the same test (Choi & Tinkler, 2002; Coon, McLeod, & Thissen, 2002). Potential mode differences in mandated state K–8 test results would bias the plethora of schooling decisions that rely on summative test scores, including student grade promotion and retention, course placement, and school accountability measures.

There are three categories of reasons why test scores might not be comparable between test modes: presentation characteristics, response requirements, and general administration characteristics (Bennett, 2003). Presentation characteristics include the number of items that fit on a screen versus a page of paper, as well as differences in the size of font used between the modes. The second category, response requirements, includes any differences in requirements for navigating the test and recording answers between a traditional paper and pencil test and a computer exam. The third category of mode differences comprises general administration characteristics, such as whether the test is adaptive or fixed form and whether the timing of each section and the overall test are similar. Research generally indicates that "the more complicated it is to present or take the test on computer, the greater the possibility of mode effects" (Pommerich, 2004, p. 3). For example, tests that do not require any navigation and only ask multiple choice questions are more likely to be comparable across modes, while those that require scrolling through long reading passages and written responses are less likely to be comparable across modes.

The Present Study

The present study examined the extent to which student performance on the Partnership for Assessment of Readiness for College and Careers (PARCC) was related to test mode in the first year of administration. In the 2015–2016 school year, seven states and the District of Columbia were members of the PARCC consortium. Although PARCC found no test mode differences at the item level in the 2013–2014 pilot year, a student-level analysis of score differences in the first year of testing may indicate different results. While the majority of research on test mode effects focuses on pilot tests or low- or no-stakes tests, the present study is one of the first to examine differences in test mode results of a state-mandated K–8 assessment in its first year of implementation.

This study uses data from Baltimore County Public Schools (BCPS), a large school district in which not all schools were equipped to test online. Over 46,000 students in Grades 3–8 in 106 elementary and 28 middle schools in BCPS were given the PARCC exam in two test administrations during the 2014–2015 school year. Test mode was determined on a school-by-school basis, according to each school's ratio of students to computers. In math, 53% of students tested online; in English/language arts (ELA), 29% of students tested online.[1] Table 1 shows that students who tested online were more likely to be Black or Hispanic and to qualify for free and reduced meals (FARMS); on average, students who tested online also had lower prior achievement.[2]

[1] Thresholds for testing online were lower in math than in ELA because PARCC's online math calculator allowed the district to forego the purchase of handheld calculators in schools that tested online.
[2] These unexpected differences are likely a result of the recent infusion of technology into the district's historically under-resourced Title I schools.

Table 1

Student and School Characteristics of Paper and Online Test Takers

	Math		ELA	
	Paper	**Online**	**Paper**	**Online**
Asian	7.4	5.4***	7.2	4.7***
Black	35.3	41.1***	36.8	40.8***
Hispanic/Latino	7.4	8.5***	6.9	9.0***
Two or More Races	4.6	4.6	4.4	4.8
White	44.9	39.9***	44.2	40.2***
Female	49.4	48.4*	49.4	48.8
FARMs	45.0	56.9***	45.4	60.5***
Special Education	10.8	12.7	10.2	11.9***
ELLs	1.9	2.4**	1.3	1.7***
Gifted	22.1	22.5	28.3	21.3***
Prior Achievement	0.09	-0.07***	0.05	-0.10***
N	19,594	21,788	32,854	13,377

Note. Only includes students who took both tests (PBA and EOY) in the same mode. *p<0.05. **p<0.01. ***p<.001.

Hierarchical linear modeling (HLM) was used to compare differences in students' PARCC scale scores based on test mode, with students nested in schools. HLM accounts for two levels of variation in the outcome—variation within schools and variation between schools. (See appendix for more information on methods.) Student and average school demographics and same-subject prior achievement were included as covariates to control for the non-random assignment of schools to online testing.

Results

Results indicate that there was a statistically significant mode effect in all subject–grade combinations. Table 2 shows that, after controlling for student demographics and prior achievement, students who took PARCC on paper scored substantially higher than students who took the exam electronically. On average, students who tested online scored between 3 and 11 percentile points lower than their peers who tested on paper in math, and between 11 and 18

percentile points lower in ELA. Within ELA, the effect was larger on the writing portion of the test than on the reading portion. Further, the paper advantage was larger in middle grades than primary grades, potentially due to different test response requirements. No consistent interaction effects were found between test mode and prior student achievement or student demographics.

Table 2

Estimated Differences in PARCC Scores by Test Mode, in Standard Deviations

	Math		ELA					
	Overall		Overall		Reading		Writing	
Grade	Without Controls	With Controls	Without Controls	With Controls	Without Controls	With Controls	Without Controls	With Controls
3	-0.03	0.05*	0.24***	0.28***	0.14***	0.18***	0.37***	0.40***
4	0.11***	0.11***	0.39***	0.34***	0.27***	0.22***	0.49***	0.44***
5	0.01	0.06*	0.26***	0.25***	0.16***	0.15***	0.38***	0.38***
6	0.70***	0.23***	0.85***	0.44***	0.78***	0.37***	0.84***	0.48***
7	0.70***	0.24***	0.83***	0.35***	0.70***	0.23***	0.93***	0.44***
8	0.83***	0.23***	0.77***	0.33***	0.66***	0.25***	0.85***	0.45***

Note. Positive differences indicate higher scores for paper test takers; estimates with controls are calculated using a multilevel model.
*p<.05. **p<.01. ***p<.001.

Discussion

The test mode effect found in this study suggests that the online PARCC exam measured more than subject matter. Two possible explanations for a mode effect are students' lack of experience with computer tests and the response requirements of the electronic version of the test. Two elements of the present study favor the latter explanation. First, all BCPS students in Grades 3–8 were administered a non-PARCC formative assessment on computers prior to the PARCC test, suggesting that they had at least some experience with computer tests. Second, the variation in mode effects by subject and school level—Grades 3–5 versus Grades 6–8—align with differences in PARCC's response requirements by subject and grade.

A mode effect on a multi-state, state-mandated summative assessment has substantial implications. In states and districts that tested completely online, there was no observed mode effect; however, there still exists a theoretical mode effect that may bias results away from a true score. Biased assessment scores can effectively invalidate school accountability ratings as well as any schooling decisions that rely on these data, including students' course placement. States are moving towards testing all students online but not all schools are equipped to do so. In the meantime, there is a solution to address mode effects. Test producers should examine mode comparability and, when a mode effect exists, adjust the scale scores of each mode to eliminate any mode disadvantage.

References

Bennett, R. E. (2003). *Online assessment and the comparability of score meaning* (ETS Research Memorandum RM-03-05). Princeton, NJ: ETS.

Choi, S. W., & Tinkler, T. (2002). *Evaluating comparability of paper-and-pencil and computer-based assessment in a K–12 setting*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Coon, C., McLeod, L., & Thissen, D. (2002). *NCCATS update: Comparability results of paper and computer forms of the North Carolina end-of-grade tests* (RTI Project No. 08486.001). Raleigh, NC: North Carolina Department of Public Instruction.

EdTech Strategies. (2015). *Pencils down: The shift to online and computer-based testing*. Retrieved from https://www.edtechstrategies.com/wp-content/uploads/2015/11/PencilsDownK-8_EdTech-StrategiesLLC.pdf

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment, 2*(6).

Appendix:

Methods

In 2014–2015, the PARCC exam was given in two administrations, a Performance Based Assessment (PBA) in early spring and an End of Year (EOY) assessment in late spring. In this analysis, we have only included students who took both administrations in the same mode, which was over 99.5% of students in both math and ELA.

Hierarchical linear modeling (HLM) was used to compare differences in students' PARCC scale scores based on test mode, with students nested in schools. HLM accounts for two levels of variation in the outcome—variation within schools and variation between schools. The following equation was modeled:

$$PARCC_{ij} = \gamma_{00} + \gamma_{00} MeanPriorAch_j + \gamma_{01} MeanX_j + \gamma_{10} PriorAch_{ij} + \gamma_{20} X_{ij} + \varepsilon_{ij} + r_{0j}$$

where $PARCC_{ij}$ is the standardized PARCC scale score of student *i* in school *j*; $MeanPriorAch_j$ is the average prior achievement score of school *j*; $MeanX_j$ is a factor of average school-level student characteristics, including race/ethnicity, free and reduced-price meal (FARM) status, English language learner (ELL) status, special education status, and gifted status; $PriorAch_{ij}$ is the same-subject, same-year winter MAP score of student *i* in school *j*; X_{ij} is a factor of student characteristics, including race/ethnicity, FARM status, ELL status, special education status, and gifted status; and ε_{ij} and r_{0j} are the student-level and school-level error terms, respectively. Student- and school-level demographics and prior achievement are included to control for the non-random selection of schools into the online test mode.

Since PARCC scores were not vertically scaled across grades, scores were standardized by subject and grade within the district, and the above equation was estimated separately by

grade and subject. Separate models that include interaction terms between test mode and student characteristics were also investigated to examine potential differential effects of test mode.

BCPS students in Grades 3–8 took the computer-adaptive MAP test in both the fall and winter. The winter test was utilized as the control for prior achievement due to its temporal proximity to the spring PARCC exam. This proximity minimizes any potential correlation between school value-added and PARCC test mode. It should also be noted that the use of a computer-based test as a control for prior achievement likely downwardly biased the estimates in the present study by essentially removing any potential effect of inexperience with computer testing.