MARYLAND STATE DEPARTMENT OF EDUCATION
PREPARING WORLD CLASS STUDENTS

Jack R. Smith, Ph.D.
Interim State Superintendent of Schools

200 West Baltimore Street • Baltimore, MD 21201 • 410-767-0100 • 410-333-6442 TTY/TDD • msde.maryland.gov

TO: Members of the Maryland State Board of Education

FROM: Jack R. Smith, Ph.D. JRS/CLW

DATE: March 21, 2016

SUBJECT: Study of PARCC Results by Mode of Delivery (Mode Effect)

PURPOSE:

To continue the discussion from the February 23, 2016, State Board meeting concerning an analysis of the 2014-2015 PARCC results by mode of delivery: online vs. paper.

BACKGROUND:

For the initial administration of the PARCC tests, 876,787 tests were scored in grades three through eight and ten in English/Language Arts (ELA) and grades three through eight mathematics, Algebra I, and Algebra II. Tests were administered using two different modes of delivery: 713,672 online computer-based tests (80%) and 163,115 paper-based tests (20%). As part of the internal validation process, student performance on both modes of delivery was studied to ensure that the mode of delivery itself, did not cause an effect on student performance. The findings were shared with the State Board. The State Board then requested a continuation of the discussion to include more specificity on the possible causes of the observed mode effect as well as the expected usage of online testing for the 2015-2016 school year and beyond.

EXECUTIVE SUMMARY:

Maryland has over twenty-five years experience assessing students with computers; eight years assessing students online and including extended responses. Over the years, multiple studies have been conducted showing that the mode of delivery of the assessment had no impact on overall student performance.

The question then is what is the difference between Maryland's prior assessments and the PARCC tests? In the past, paper was the standard. Computer forms mimicked the paper forms both in the items themselves and the presentation of the items. Where constructed responses were required of students, they were expected to be brief, and the text box size on the computer forms were similar in size to what was presented on paper. The PARCC tests were designed with the use of computers now being the standard. The online versions of the tests include items that cannot be rendered on paper. There are technologically enhanced items and tools along with multimedia presentations including both visual and audio stimulus for students to engage. The PARCC tests also include writing which exceeds what has been expected of Maryland students in the past. The technologically enhanced items along with the online tools were studied and were found to have no impact on overall student performance. Items requiring students to write extended responses are the primary source of the difference in student performance between modes of delivery.

Research on the PARCC tests by Steve Graham suggests a cause for the discrepancy in student performance between the two modes of delivery is due to instructional readiness. The world now depends on computers for communication needs. For students to best demonstrate their mastery of the writing process, students need experience using computers as a tool throughout the writing process and not simply to create a clean final draft. Graham's research indicates that students that have experience using computers outperform those that rely on paper; however, students that rely on using paper tools may perform better on paper tests than on a computer tests. Graham's findings are consistent with what is observed in Maryland's data.

ACTION:

For information purposes only. No action required.

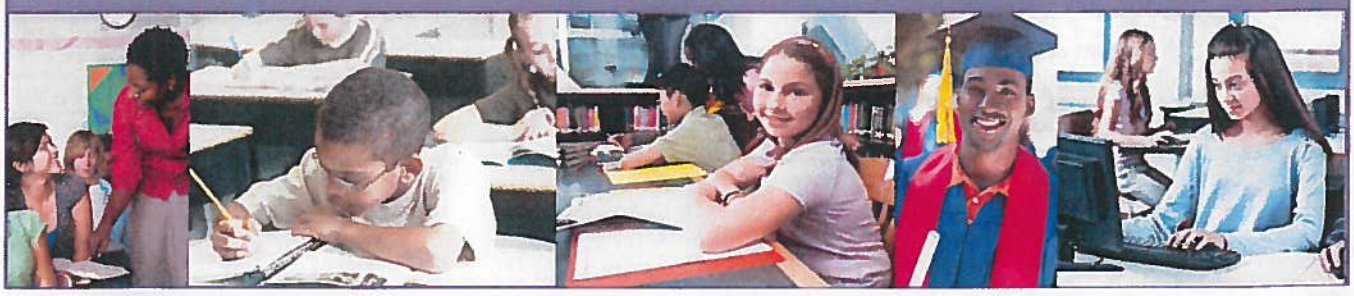
Attachments: Presentation: Study of PARCC Results by Mode of Delivery…A Deeper Dive
Final Technical Report for 2015 Administration of the PARCC
Here's How the Method of Testing Can Change Student Scores

PARCC

Final Technical Report for 2015 Administration

Educational Testing Service
Pearson
Measured Progress

February 22, 2016



The high school report noted that the PARCC assessment program incorporates a number of accessibility features and test accommodations for students with disabilities and for English language learners. Furthermore, the PARCC assessments included items designed to accommodate the needs of students with disabilities.

9.5.3 Mode Comparability Study

The PARCC (Operational) Mode Comparability Study was conducted using the 2015 operational data to support both computer-based testing (CBT) and paper-based testing (PBT) modes of administration of the PARCC assessments (Liu, Brown, Chen, Ali, Hou, & Costanzo, 2016).

For the spring 2015 operational administration, schools and districts within each state selected the mode of test administration. The resulting CBT and PBT test-taking groups were therefore not randomly equivalent. To improve the overall comparability of the CBT and PBT groups, propensity score matching, based on test-taker demographic information, was used. Then item-level analyses (e.g., *p-values*, and differential item functioning) and test-level analyses (e.g., test characteristic curves) were conducted.

Item-level analyses showed that there were negligible to small differences in terms of *p-values* and average item scores across modes for the majority of items in mathematics and ELA/L. Prose Constructed Response (PCR) task traits in ELA/L had larger *p-value* effect sizes than other items, all favoring PBT. A very small percentage of items was identified as functioning differently (with C-level DIF) in the two modes. Many items ELA/L PCR task traits were also found to have B-level (DIF), favoring PBT.

Additionally, the item response theory (IRT) difficulty and discrimination parameters estimated separately within mode were highly correlated. For grade levels with a lower correlations between modes, removing items with outlier parameter estimates provided substantial improvement in the correlation. As well, the overall the differences between common test characteristic curves (TCCs) of different modes were small and within 0.5 raw score points, except for ELA/L grade 9 and Geometry where TCC differences exceeded the differences that matter criterion in regions of the theta scale where large percentages of students were located. When comparing the performance on the common items, the effect sizes ranged from negligible to small for most of the tests evaluated. The directions of effect sizes were not consistent across subject/grade levels.

Additional analyses were conducted on students from one of the states that provided prior state assessment scores. Summary statistics of these students' prior state assessment scores suggested CBT and PBT samples from propensity score matching (PSM) were not comparable in their prior achievement. Therefore, poststratification weights based on prior state assessment score were used to calculate PBT students' PARCC scale score to minimize the impact of noncomparability of prior achievement across modes. The scale score differences were largely reduced for mathematics grade 5, 7 and Algebra I after weighting. Small effect sizes, in favor of PBT, were found for Geometry and ELA/L grade 9 and a negligible effect size was found for ELA/L grade 7 after poststratification weighting.

The PARCC (Operational) Mode Comparability Study found evidence that the score comparability was not consistent across all content domains and grade levels. As noted in the study, only one state provided previous year's achievement data, therefore, the CBT and PBT groups were matched based on

only demographic data. Furthermore, the additional analyses based on the one state that provided prior achievement data indicated that the CBT and PBT matched groups were not comparable in terms of their prior achievement. Thus, caution should be taken when interpreting the results of the Mode Comparability Study.

9.5.4 Device Comparability Study

In addition to the PARCC (Operational) Mode Comparability Study, the comparability across digital devices (e.g., tablet versus non-tablet) was evaluated using the 2015 operational data (Steedle, McBride, Johnson, & Keng, 2015).

PARCC allows students to take its assessments on a variety of digital devices, such as desktops, laptops, and tablets. It is therefore important to evaluate comparability across digital devices by investigating whether test items were of similar difficulty, whether psychometric properties of test scores were similar, and whether overall test score interpretation was similar across traditional (i.e., desktops and laptops) and non-traditional (i.e., tablet) computing devices. For the 2015 Device Comparability Study, any student who took one of the study forms on a tablet or non-tablet device were eligible for inclusion in the study. Students were matched on demographic information to create tablet and non-tablet samples that were considered randomly equivalent.

The 2015 Device Comparability Study found evidence of comparability between test scores from tablets and non-tablet devices. The item p-values and IRT difficulty estimates were similar across tablets and non-tablet devices. A small number of items were flagged for device effects, and nearly all of them were part of high school mathematics assessments. The raw score and scale score distributions indicated similar overall performance on both components (PBA and EOY) of the 2015 PARCC assessments. Additionally, IRT true-score equating indicated that students who tested on non-tablet devices would be expected to score similarly had they taken the same PARCC assessment on tablets.

9.6 Evidence Based on Response Processes

As noted in the AERA, APA, and NCME Standards (2014), additional support for a particular score interpretation or use can be provided by theoretical and empirical evidence indicating that test takers are using the intended response processes when responding to the items in a test. This type of evidence may be gathered from interacting with test takers in order to understand what processes underlie their item responses. Evidence may also be derived from feedback provided by test proctors/teachers involved in the administration of the test and raters involved in the scoring of constructed response items. Evidence may also be gathered by evaluating the correct and incorrect responses to short constructed response items (e.g., items requiring a few words to respond) or by evaluating the response patterns to multi-part items.

9.7 Interpretations of Test Scores

The PARCC ELA/L and mathematics scores are expressed as scale scores (both total scores and claim scores), along with performance levels to describe how well students met the academic standards for

Tab 1 Attachment 2

THE CONVERSATION

Here's how the method of testing can change student scores

March 1, 2016 6.20am EST

What's the best tool for taking tests? Gerald R. Ford School of Public Policy, University of Michigan, CC BY-ND

Steve Graham
Professor of Leadership and Innovation, Arizona State University

Students who recently took the Partnership for Assessment of Readiness for College and Careers (PARCC) scored lower when they took the test on a computer than when they used paper and pencil.

This might not matter much if the results of these tests played a minimal role. But they do not. Test scores are used for accountability purposes at the federal, state and local level. In some states, test scores play a role in student graduation and the evaluation of teachers and principals.

The question is, does the method of test taking actually influence test results?

I have been researching factors that influence test performance when students write essays. Such essays are written with paper or pencil or on a computer. Based on research that I coauthored in 2011, the answer to this question is yes. But there are several caveats.

In contrast to the findings from the PARCC test, we found that students writing on a computer scored higher than students writing with paper and pencil. This finding did not apply to all students, though. Students with little experience using a computer to write had higher scores when writing by hand.

Computer versus pencil-and-paper tests

In the last five years, two partnerships of U.S. states funded by the federal program Race to the Top were tasked with developing assessments for determining if students were on track

or ready for college and the world of work.

The consortia developed computer-based assessments that, among other things, would make scoring easier, sharing results faster and conducting assessments cheaper. Many, but not all states, agreed to use these tests to assess students' academic progress in multiple grades across the school years.

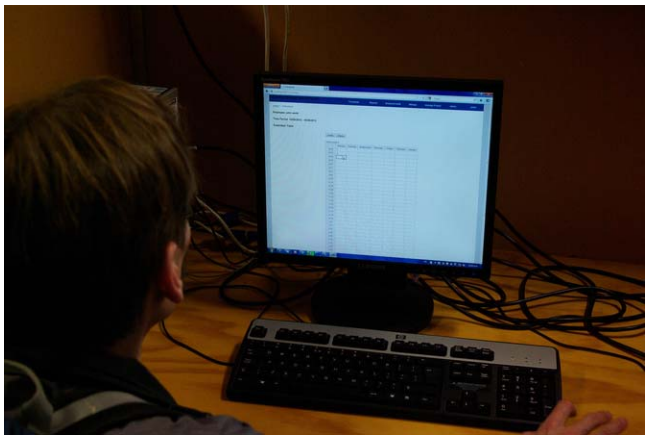
For tests developed by one of the consortia, Partnership for Assessment of Readiness for College and Careers (PARCC), students obtained higher scores in English/language arts on the paper pencil version versus the computer one.

By contrast, I obtained very different results in my review of seven scientific studies of factors that influence test results. Students' writing performance on computer assessments was 21 percentile points higher when compared to students who wrote via paper and pencil.

But then, another review I conducted of 18 scientific studies found the same 21 percentile advantage for writing when students used computer for writing in the classroom.

Computer-based assessments

So why are there differences between PARCC tests results and the finding from scientific studies I reviewed? A likely explanation involves students' experience with the method of testing.



Computers can underestimate writing achievements. Samuel Mann, CC BY

My review of four scientific studies showed that students with little experience using computers as an assessment tool scored 18 percentile points lower than when they composed their essays using paper and pencil.

In other words, a student's mastery of the method of testing matters. For students with little experience, computer assessments underestimate their writing achievement.

To get a sense of how method of testing can influence writing performance, imagine you are asked to write something for a test using a Chinese typewriter. This is a very complex writing tool designed to create 6,000 characters. Top typing speeds are 11 characters per minute.

Even if you reach this benchmark, you will have no hope of typing fast enough to get all your

thoughts down on paper before some of your ideas slip from memory. If you are not proficient with this typewriter, then the problem is even worse. As you hunt for the next character, your memory is taxed even further, resulting in even more ideas being lost.

As this example illustrates the method of testing can interfere with a students' performance. If a student is not adequately familiar with the testing tool or it is cumbersome to use, time and energy most be devoted to using it.

This is time and energy that can profitably be devoted to answering test questions.

Pencil-and-paper assessments

These kinds of problems are not limited to tests taken on a typewriter or computer, they can occur for paper-and-pencil tests too.

Students handwriting is not always fast enough for them to record all of their ideas before some of them slip from memory. This is a problem even for college students.

In a study with University of London undergraduates, handwriting fluency accounted for 40 percent of the variance in their scores on a timed-essay writing test.

Legibility of response can influence results on a pen-and-paper test. Dennis S. Hurd, CC BY-NC-ND

With paper-and-pencil tests, there is an additional complicating factor. Scores on handwritten tests can be influenced by the legibility of the response. Test responses that are less legible can drop scores by 35 percentile points compared to the same response that is written neatly and legibly.

Making the matter even more complicated, a typed paper is scored more harshly than the same handwritten paper.

In a review of five scientific studies, I found that the score for a typed version of a handwritten text dropped by 18 percentile points. According to teachers involved in these studies, spelling and grammar errors were more visible in typed versus the handwritten version of the same paper.

So, method of testing makes a difference in the following ways: if students are not adept at taking a test on a computer, they score higher on the same paper-and-pencil test. If they are adept with a computer, they score higher on the computer test. Students performance is further moderated by handwriting fluency and legibility on paper-and-pencil tests and the number of spelling and grammar errors on computer tests.

Why use digital tools

What testing methods should schools use? Should computer-based assessments be abandoned, in view of recent PARCC results?

In the best of all possible worlds, students should be allowed to use the method of testing they are most proficient with when taking tests. However, this is unlikely to happen as it adds another level of complexity and costs to test taking. So, one alternative for groups like PARCC is to statistically adjust scores to reflect the differences between test taking modes.

Abandoning computer-based tests would be a mistake. These assessments have the potential to move schools from 19th-century writing tools to 21st-century tools.

As high-stakes assessments go increasingly digital, schools will make word processing and other digital composing tools a common staple. Studies have shown that students who use such tools over time become better writers than those who continue to write with paper and pencil.

At the end of the day, testing must produce something positive. Better writing tools in the classroom would be a step in the right direction.

Standardized testing
Partnership for Assessment of Readiness for College and Careers PARCC
Digital tools

Tab 1

Study of PARCC Results by Mode of Delivery (Mode Effect)… *A Deeper Dive*

Maryland State Board of Education Update

March 21, 2016

Dr. Henry R. Johnson, Interim Deputy State Superintendent

Dr. Douglas Strader, Director of Assessment

2014-2015 Administration

- 876,787 tests were scored in:
  - Grades 3-8 English Language Arts (ELA) and English 10
  - Grades 3-8 Mathematics, Algebra I, and Algebra II

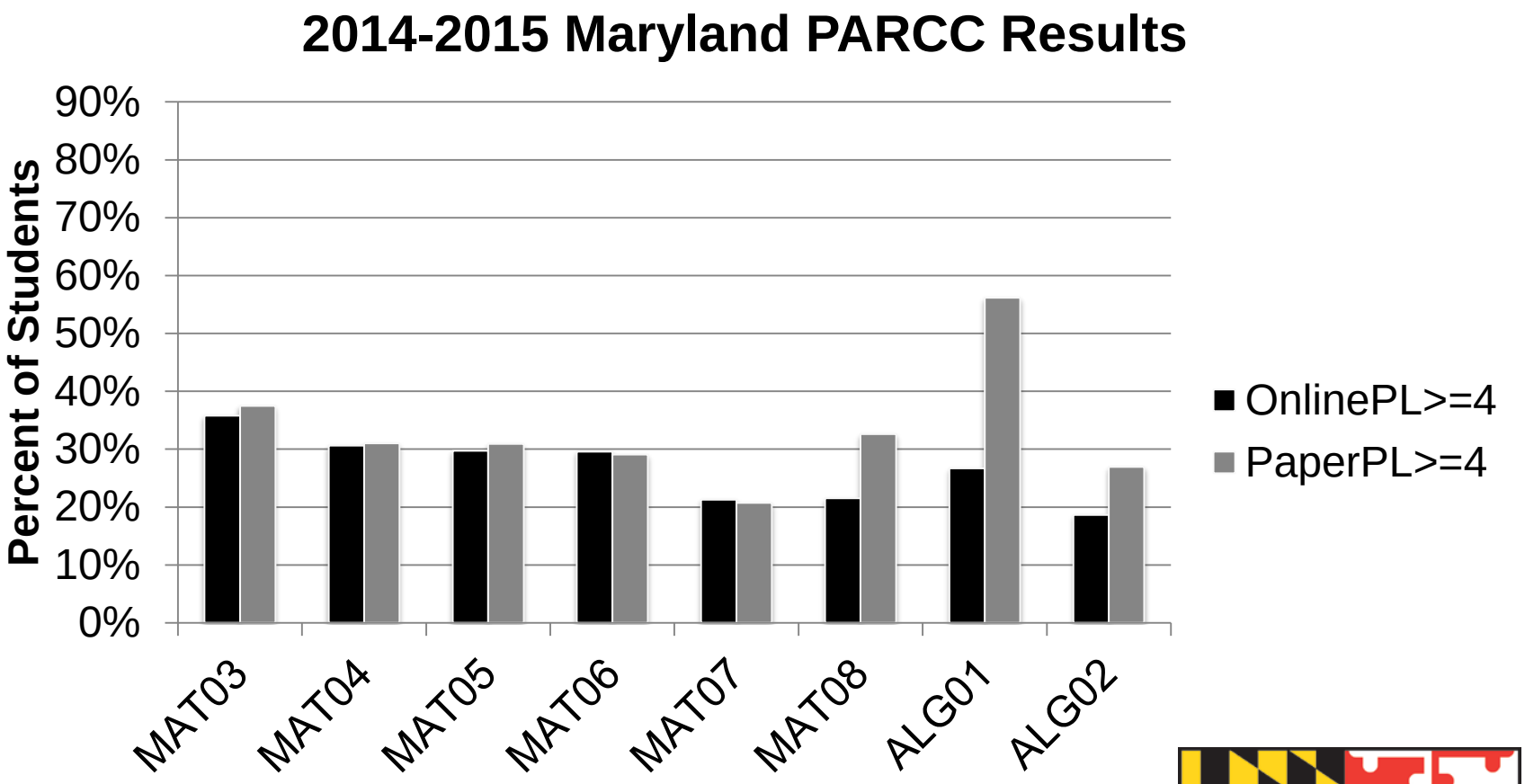
- Tests were administered using 2 different modes of delivery
  - Online Computer-based Tests: 713,672 (80%)
  - Paper Tests: 163,115 (20%)

2015-2016 and beyond

- Pretest files for this 2015-2016 administration indicate that over **92% of tests** will be completed online (10% increase from last year)

- 2016-2017 and beyond… All students will be assessed online. The only exceptions will be made for students needing specific accommodations and situations where the school infrastructure will not support online testing.

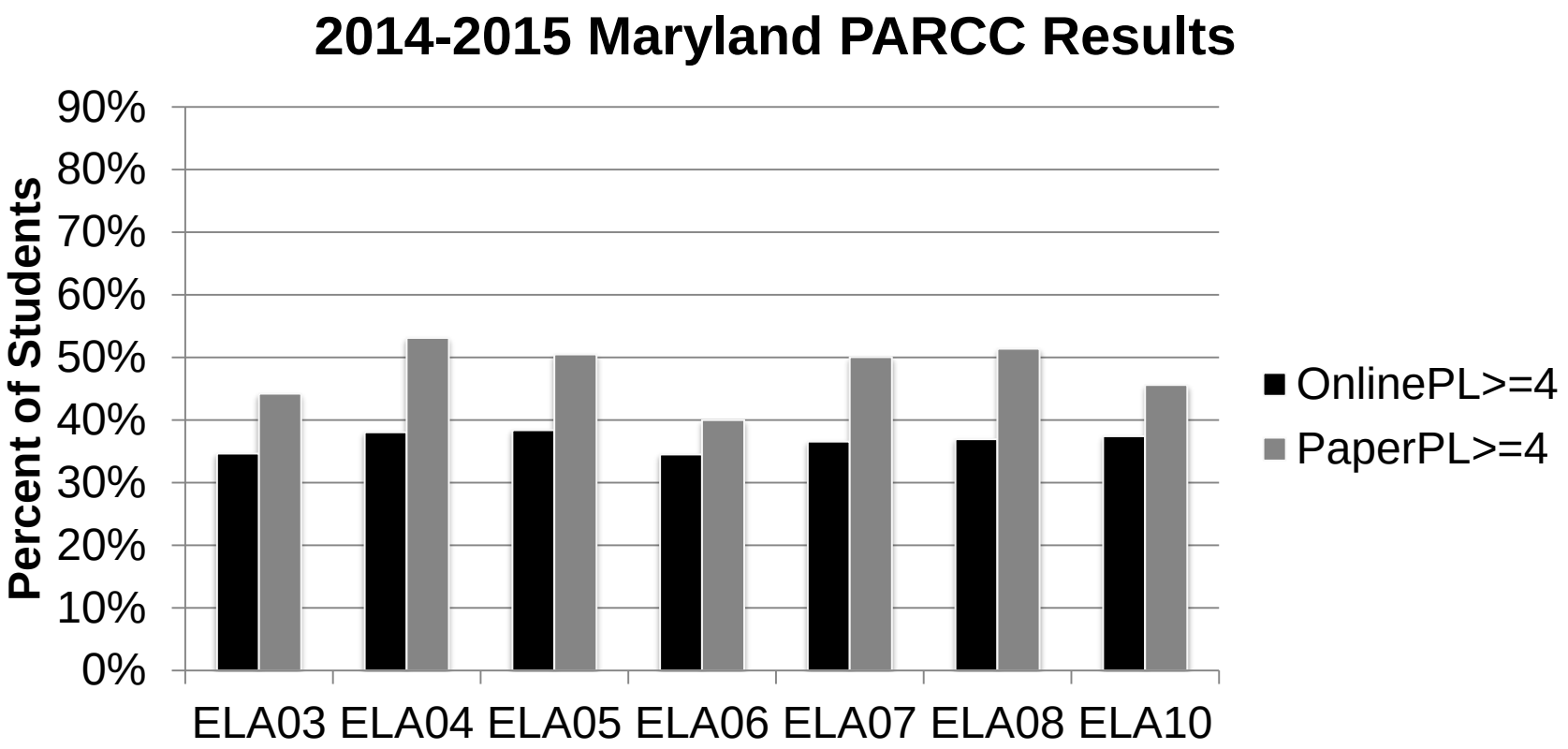
PARCC Results by Mode, Content/Test, and Performance Level



Performance Level 4 and 5 are used for College and Career Readiness Determination

MARYLAND STATE DEPARTMENT OF EDUCATION
PREPARING WORLD CLASS STUDENTS

PARCC Results by Mode, Content/Test, and Performance Level



Performance Level 4 and 5 are used for College and Career Readiness Determination

MARYLAND STATE DEPARTMENT OF EDUCATION
PREPARING WORLD CLASS STUDENTS

Findings shared during February BOE meeting…

- Comparison of Mode by performance level illustrates that students that took the test on paper tended to outperform the online test-takers on the ELA tests and higher level math tests
- Greater percentage of higher performing students in Maryland took the paper form
- No evidence of any particular student group impacted more than the population as a whole

Findings con't

- Items requiring extended responses most greatly impacted by mode favoring paper
- There is no evidence of any technical issues in the development, administration, scoring or reporting of the results

Maryland's History with Use of Computers

- Over 25 years experience assessing with computers
  - Selected response items with the introduction of the Computer Adaptive Version of the Maryland Functional Tests in 1989
  - Extended Constructed Response items with the introduction of the MSA program, MSA Science and the HSA program in 2007
- During the 2014 School year, roughly 60% of Maryland assessments were taken using computers

What's different between Maryland's prior assessments and PARCC?

Maryland's assessment prior to PARCC

- Paper was the standard. Computer forms mimicked paper forms. Same items – same presentation.
- Constructed Response items were relatively brief in length. Text box size on computer forms were similar in size to what was presented on paper

What's different between Maryland's prior assessments and PARCC?

PARCC

- Computer is now the standard
- Introduction of Technologically Enhanced Items (TEIs) and tools
- Introduction of Multimedia in presenting information
- Writing expectations far surpass Maryland's prior assessments

Understanding the Student Experience

Focusing on ELA PARCC Practice Tests

http://parcconline.org/assessments/practice-tests

PARCC Released Items with Student Responses

https://prc.parcconline.org/assessments/parcc-released-items

Possible Reasons for Mode Effect…

1. Technical issues with the test itself in the development, administration, scoring, and/or reporting of results
2. The population of students that took each mode of delivery varied
3. Readiness - students were not equally prepared to engage both modes of delivery

1. Technical issues with the test

Sampling of what was studied…

- Equating forms – linking item model used for equating forms
- Item comparisons across forms
- Item rendering
- Scoring of TEI and other online items

2. Differing populations

- When analyzing how the students performed on the MSA/HSA the prior year
  - PARCC Paper – 81% of students scored Proficient or Advanced the prior year
  - PARCC Online – 76% of student scored Proficient or Advanced the prior year
- Paper population consists of greater population of high performing students accounting for an average of 40% of discrepancy across all tests

3. Readiness

- Item type with greatest mode effect: Extended Constructed Response (ECR) items

- Two areas of readiness contributing to mode effect
  - Interfacing with the technology/assessment platform
  - Instructional readiness

3. Readiness con't

Interfacing with the technology/assessment platform

- Inconclusive
  - Typing time
  - Online platform tools (i.e. equation editor)
- Anecdotal
  - "Fill the box" – the box on paper used for capturing student extended responses was visually much larger than the computer box
  - Computer box increased for 2015-2016

3. Readiness con't

Instructional readiness for assessment of new standards

- Lack of understanding of expectations
- Referencing both multi-media and text in extended responses
- Using technology tools in the writing process

Recent research assessing writing

Comments by Steven Graham…

- If students are adept to writing using a computer, they score higher on the same paper-and-pencil test
- For students with little experience writing using computers, computer assessments can underestimate their writing achievement
- Abandoning computer-based test would be a mistake. These assessment have the potential to move schools from the 19th century writing tools to the 21st century tools

Steven Graham: Here's How the Method of Testing Can Change Student Scores (March 1, 2016)

Recent research assessing writing

As high-stakes assessments go increasingly digital, schools will make word processing and other digital composing tools a common staple. Studies have shown that students who use such tools over time become better writers than those who continue to write with paper and pencil.

Steven Graham: Here's How the Method of Testing Can Change Student Scores (March 1, 2016)

Moving forward…

- Online is the new standard. The PARCC assessments were developed to be online tests. Performance Levels were set using online forms only
- Maryland is transitioning to entirely online by the 2016-17 school year
- Newness/readiness issues need to be further studied and performance expectations made clear

Comments/Questions?