Smarter Balanced "Tests of the Test" Successful:

Field Test Provides Clear Path Forward

Prepared by Nancy Doorey on
behalf of Smarter Balanced

October 2014

Between March and June of 2014, the Smarter Balanced Assessment Consortium conducted a field test of its new online assessment system. Thirteen participating states provided the results of surveys given to students and adults involved in the Field Test. Overall, more than 70% of test coordinators in each of seven states indicated that the Field Test had gone either as well as or better than expected, and most students (an average of 67% across 5 states) found the test interface "easy" or "very easy" to use. Several important lessons emerged to inform final preparations this fall and winter. In addition, the surveys indicate that states need to continue their work to help teachers align instruction with the increased rigor that college and career readiness standards require.

Last spring the Smarter Balanced Assessment Consortium, a group of 21 Governing States and the U.S. Virgin Islands, completed a 12-week Field Test of new, "next-generation" assessments in English language arts/literacy and mathematics for students in grades 3 through 8 and high school. More than 4.2 million students across 16,549 schools participated in the Field Test, making it the largest online assessment and largest field test of a new assessment ever—larger than the National Assessment of Education Progress (NAEP), which samples students across all 50 states plus the District of Columbia and Department of Defense schools, and larger than the Programme for International Student Assessment (PISA), which samples students across 65 participating countries.

The complexity of this effort would be difficult to overstate. The pieces of the system, which were being developed by 10 main contractors and many other contributors, had to come together to create a seamless, integrated system, with little room for error. Adding to the challenge, this project was managed by a group of states that had never before worked together on this scale.

In addition to building the assessments and the open source delivery and scoring platform, the Smarter Balanced states, some of which have been using online state tests for more than a decade and others for which this would be their first foray, also coordinated efforts over the past four years to upgrade the technology infrastructure across their districts and schools. In some cases, state legislatures allocated significant funds to augment local budgets. As this report will describe, however, states' readiness as of spring 2014 varied significantly across and, presumably, within states, as did the readiness of adults to

administer them. The Field Test provided important information to guide states' and districts' remaining preparations for the spring 2015 administration when results will be used for accountability purposes.

The Field Test served a variety of purposes. Its primary purpose was to "test" the computer-delivered questions and embedded tools to ensure that they function properly, are clear, and meet criteria for inclusion in the spring 2015 state summative assessments required under the No Child Left Behind Act. The Consortium developed and tested a much larger number of items than needed for the secure assessments so that a representative subset of them could be placed into a non-secure item bank for the creation of interim assessments that faithfully mirror the rigor and item types of the summative assessments.

The Field Test was also an important trial run of the entire assessment delivery system, the readiness of state, district, and school personnel to administer the tests, and the readiness of students to take them. To gather information about these additional factors directly from their students, educators, and technology personnel, many of the Consortium states surveyed participants as part of the Field Test.

This report is based on a review of the survey results from 13 of the 22 members. These states collectively educate approximately 40% of all students in the Consortium. In total, feedback was reviewed from 19,600 students and 4,946 adults (administrators, classroom teachers and proctors, test coordinators, and other adults closely involved in the administration of the Field Test).

Each of the 13 states developed its own survey questions and survey distribution processes. While this was consistent with the Consortium's emphasis on state flexibility and control, it hindered this review. This report is not based on a scientific sample, nor is it a scientific analysis of consistently worded survey questions across states; rather, this review reflects the author's findings regarding major themes across the results from these 13 states. The number of states that provided data on a topic and the number of student or adult respondents have been noted.

This report is organized around six major topics that were addressed by two or more of the 13 states that provided survey data:

A. **The technology readiness of states, districts, and schools;**
B. The **readiness of test administrators and proctors** to properly deliver the assessments;
C. The **student test interface** and embedded tools, supports, and accommodations;
D. The **functioning of new item types**;
E. The **rigor of the assessments and degree to which items and tasks reflected what students had learned class (instructional alignment);** and
F. The functioning of the Smarter Balanced **test delivery system and help desk.**

It is important to note that the Field Test did not—and could not—serve as a test of one of the major features of the spring 2015 assessments: the management of the delivery and scoring of the assessments by state agencies and their chosen contractors. The Field Test was centrally managed by the Consortium. An early policy decision by the member states, however, was to have the operational assessments managed by states. Each state, therefore, will oversee test administration, help desk services, scoring, and reporting of results with the assistance of their chosen service providers, and may opt to use one or more of the open source system components developed by the Consortium and tested in this trial run of the system.

In addition to summarizing the major themes of the survey responses, this report also includes comments from Consortium leadership and state assessment directors regarding the steps being taken to address the needs that were identified.

A. The technology readiness of states, districts, and schools

The Race to the Top Assessment Program, which provided grants to groups of states that wanted to create shared next-generation assessment systems, required that the assessments be computer-based in order to allow a broader range of complex skills to be measured than is possible through paper-and-pencil tests. Some states had been

administering online state assessments for more than a decade, while others needed to rapidly ramp up their technology infrastructure of computers, servers, bandwidth, and technology support personnel in order to deliver them. The Consortium provided tools to help states and districts identify technology infrastructure improvements needed prior to the spring 2015 operational assessments.

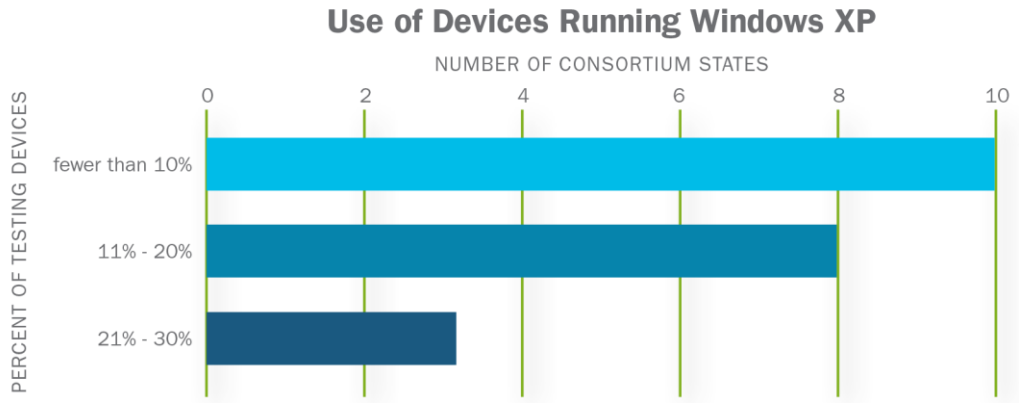
Overall, test coordinators felt the Field Test went well, with 70% or more of them in each of seven states responding it had gone either as well as or better than expected.

Overall, test coordinators felt the Field Test went well, with 70% or more of them in each of seven states responding it had gone either as well as or better than expected. The rate at which technology problems were encountered varied greatly across states, however. At the extremes, 73% of test administrators in one state reported never needing to contact the help desk for assistance, while in another only 26% were so fortunate.

One concern identified back in 2011 was the large number of computers in schools that used the Windows XP operating system, which was due to be removed from active support by Microsoft in April 2014. The July 2012 snapshot of technology readiness across PARCC and Smarter Balanced states[1] estimated that 56.1% of the devices in schools that would most likely be used for testing used Windows XP. Both consortia strongly urged states and districts to update their equipment and operating systems.

By the time of the Smarter Balanced Field Test, only 13% of devices used by students were running the Windows XP operating system: 10 of the 21 Smarter Balanced states[2] had fewer than 10% of devices running Windows XP, another eight had 11% to 20% devices using XP, and the remaining three states had between 21% and 30% of devices using Windows XP.

[1] *A Summary Report for the July 15, 2012 Data Snapshot*, 2012.
[2] This count includes the U.S. Virgin Islands



While these data raise concern for those states with higher percentages remaining, they may also paint an overly rosy picture for states and districts that had only a sample of schools participate: the schools that volunteered for the Field Test may have been those that had completed most of the needed upgrades and may not be representative of all schools.

Another indicator of technology readiness is the number and cause of help desk contacts during both the preparation period, when student accounts were set up and the system was being tested, and during student testing. Of the nearly 78,000 help desk contacts across all Consortium states, more than 80% of the problems arose prior to student testing. During the Field Test preparation and registration period, 35% of help desk calls involved correction or resetting of usernames and/or passwords and another 21% were requests for clarification of prior communications.

A technology challenge cited in the surveys was interruption of Internet connectivity, particularly for devices using wireless connections. The frequency of this problem was not quantified, but each state now has baseline information with which to prepare for the first operational assessments in spring 2015.

As Frank Gerdeman, the Assistant Director for Secondary and Adult Education Division at the Vermont Agency of Education, pointed out, issues of inadequate Internet connectivity and

bandwidth have larger educational implications: “The assessment doesn’t require any more bandwidth than what should already be in place to support instruction.”

The state that faced the largest challenge, by volume, was California, which had more than 3.1 million students participate in the Field Test. “This was a massive effort that required a great deal of planning and communication, but, despite all of the pre-test anxiety, there were no statewide breakdowns and the majority of districts reported a positive experience,” stated Diane Hernandez, Director of Assessment Development and Administration at the California Department of Education. “We’re excited by how well it went and how quickly our schools and districts are making the transition to online testing.”

> “The assessment doesn’t require any more bandwidth than what should already be in place to support instruction.”
>
> —Frank Gerdeman, Assistant Director for Secondary and Adult Education Division at the Vermont Agency of Education

B. **The readiness of test administrators and proctors to properly deliver the assessments**

Overall, 70% of 2,569 test administrators and coordinators surveyed across five states indicated that the test administrator training materials were either “helpful” or “very helpful.” Those who were not satisfied described the materials as much too dense and lengthy, needing to be placed into shorter modules.

The state that had the highest satisfaction with the training materials had held in-person training sessions for district leads. In some states, however, test coordinators said the materials arrived too close to their selected field testing window to allow for in-person training sessions.

Smarter Balanced is taking several steps to address these needs. First, the Test Administration Manual has been revised, broken into smaller modules, and provided to states a full seven months before test administration. Each state can now customize the

materials, adding state-specific information, and conduct training sessions with district personnel prior to next spring.

Five test administration training modules have been released to member states, and four more will be made available this fall. They cover topics such as an overview of the entire system, test registration, and the universal tools and online features. These modules are provided as slide decks so that they can be readily customized to reflect the specific implementation plans for the state and/or district.

A new State Procedures Manual was also developed to assist state agencies as they take responsibility for the delivery, scoring, and reporting of the assessments, with the assistance of their selected contractors.

Another survey finding with significant implications for the spring 2015 administration is that many test administrators seem to have not been aware of one of the test administration tools available to all students—breaks. A frequent comment from teachers and proctors across several states was that the testing sessions were too long and that students became too tired to concentrate. Breaks are a universal tool that allows test administrators to provide a rest period when students appear to be fatigued, although breaks of more than 20 minutes will prevent the student from returning to items already seen. It is now up to districts and schools to use the insights gained from the Field Test to develop appropriate schedules for the spring 2015 administration.

C. The student test interface and embedded tools, supports, and accommodations

When delivering any assessment, it is important to ensure that the mode of delivery does not create challenges for students, impeding the accurate measurement of the desired skill, process, or knowledge. For computer-based testing, students need to be familiar with the testing interface and how to perform tasks such as clicking, scrolling, and moving to the next item.

Students also need to be familiar with each item and response type and each tool that may be attached to an item, such as a calculator, ruler, highlighter, or zoom.[3]

In addition to universal tools such as these, the Smarter Balanced testing platform includes designated supports and accommodations which are available only to students with identified needs. The Smarter Balanced designated supports include test direction translations for students with limited English skills, text-to-speech for students with reading-related disabilities or blind students who do not yet read Braille, and color contrast for students with attention difficulties or specific visual impairments. The system also includes digitally embedded accommodations for students with documented disabilities, such as video of American Sign Language translation, closed captioning for deaf or hard-of-hearing students, and refreshable braille embossers for visually impaired students.

"In the past, what we were able to afford was a multiple choice test with very few accommodations or supports," explained Angela Hemingway, the Director of Assessment and Accountability for Idaho. "Working with a consortium of states allowed us to create a better assessment with an incredible set of accommodations and accessibility (designated) supports."

The expanded set of tools and supports also created a challenge in that students needed to be familiar with them prior to testing. In order to provide opportunities for students to become familiar with them, the Consortium developed two online tools. The online Practice Tests for each tested grade level were released in the spring of 2013 to allow students to become familiar with the variety of item types, including performance tasks, and some of the embedded tools and supports. In February of 2014,

"Working with a consortium of states allowed us to create a better assessment with an incredible set of accommodations and accessibility (designated) supports."

—Angela Hemingway, Director of Assessment and Accountability for Idaho

[3] The Smarter Balanced assessment delivery platform includes a number of embedded tools, although member states may turn off specific tools if their use is in conflict with state restrictions.

the Consortium released Training Tests to provide an opportunity to learn to use the test interface and all of the embedded universal tools, designated supports, and accommodations. Both of these tools can be accessed in or out of school.

While an average of 67% of responding students across five states found the test interface "easy" or "very easy" to use, all students will need to be familiar with it by this spring in order to obtain valid summative assessment scores.

The Consortium recommended that all students have at least one opportunity to use each of these tools. Based on survey responses from more than 4,300 students across three states, however, approximately 25% had used neither or only one prior to the Field Test, and in one state that rate increased to one in three students. While an average of 67% of responding students across five states found the test interface "easy" or "very easy" to use, all students will need to be familiar with it by this spring in order to obtain valid summative assessment scores.

The use of computer-based features seemed to also positively impact student engagement. "I liked that we could highlight and strikethrough words and/or choices to an answer. It helped me a lot when finding evidence for the ELA test," responded one student.

The Practice Test and Training Tests are also valuable for those who administer the tests, according to California's Hernandez. "Through our pre- and post-Field Test surveys and focus groups, we learned that when the students and the teachers had taken both the Practice Test and Training Test in advance, they were more confident and prepared going into the Field Test."

"We anticipate that when the tests are given for accountability, this participation gap in the use of these tools will close," projected Joe Willhoft, the Executive Director of the Smarter Balanced Assessment Consortium.

D. Functioning of new item types

The Smarter Balanced assessments include some item and task types that have not been used previously in most state assessments. Some require students to use the mouse or

trackpad to select and highlight text, drag-and-drop text or graphic elements, or manipulate points on a graph. These item types appear to have caused little problem for the large majority of students, although educators raised concerns about those without access to technology at home. Interestingly, the youngest students reported the greatest ease with navigating the items and entering responses.

The Performance Tasks, extended multi-part tasks that required 90–120 minutes to complete, were also new item types that involved the use of a script for test administrators to lead the class through a warm-up classroom activity of approximately 30 minutes, and in English language arts/literacy typically required extended reading and writing/keyboarding.

The classroom activity was felt to be helpful to students by 74% of test administrators across two states, but only slightly more than half of students agreed. Teachers and other test proctors commented that high school students in particular did not seem to benefit from or see the need for the activity. The Consortium will be reviewing the high school level classroom activities and will provide additional information to teachers and test proctors regarding their role in ensuring a common foundation for students prior to the tasks.

A prevalent concern related to the Performance Tasks was the volume of keyboarding required, particularly for young students and those without access to computers at home. Several testing coordinators commented that their school districts will be putting into place plans to improve the keyboarding skills of students. Some may opt to have younger students with inadequate keyboarding skills use the paper-and-pencil version that will be available—subject to state approval—for the first three years of testing. Gerdeman, however, wants to see schools using these types of tasks and technologies as part of regular instruction: “If students are using technology appropriately in their learning, taking these assessments won’t be an issue.”

E. **The rigor of the assessments and degree to which items and tasks reflected what students had learned in class (instructional alignment)**

Students across all grade levels commented on the rigor of these assessments as compared to previous state assessments, describing them as "challenging" or "really hard," and that it "took more thought to answer questions." A tenth grader found it "hard" because "if you didn't know it [the answer], you couldn't guess" like on multiple choice tests. A 6th grader commented, "It's the first test I've ever taken where I actually learned something while taking it."

> "It's the first test I've ever taken where I actually learned something while taking it."
>
> —6th grade student

The proportion of students who found the assessments to be very difficult increased with grade level: in one state with a large number of student responses, only 14% of students at grades 3-5 found the tests to be "very difficult." At the high school level, an average of 46% of students across three states reported the assessments to be "very difficult," with the mathematics sections rated as most difficult.

For the youngest students, the ELA/literacy segments were rated as more difficult than the mathematics sections, and teachers commented on the amount of writing (keyboarding) being very challenging for some younger students and students without technology access outside of school.

While many students commented that they enjoyed the more interactive nature of the online tests and item types, some did not enjoy the increased difficulty that came with them. "I did not really like that there were not enough multiple choice questions. I felt that if there were more multiple choice questions, the test would have been easier," commented one secondary student. Another succinctly stated a complaint raised by many concerning the use of extended reading passages and writing tasks: "It was soooooooo long."

Three states, representing the East, Midwest and West, asked students about the extent to which the test questions were about things they had learned in class. Again, great variability

was seen across states, likely reflecting the intensity of teacher training to date on the Common Core, but perhaps also reflecting the degree to which the Common Core was a shift from their prior state standards. In one state, only 10% of students reported the tests to be "very well" aligned to instruction, and at the other extreme, 35% of students in another state reported this to be the case.

Across four states, older students found the test questions to address things they had learned in class less often than younger students, which likely contributed to the perception of difficulty described above. Roughly nine out of 10 students at grades 3-5, two out of three students at grades 6-8, and only one out of three students at grades 9 -11 found the assessments to be "somewhat well" to "very well" aligned to instruction.

This variation in instructional alignment could also be seen in comments from teachers who administered the assessments, particularly concerning the Performance Tasks. While some commented that the tasks were easy to explain and administer because they closely resembled their instructional activities and expectations, others expressed the opinion that the tasks and questions were "above grade level," "way too long," or "frustrating for students."

Whether the states that provided survey data are representative of the entire Consortium is unknown but, assuming the assessments themselves are well aligned to the Common Core, this limited snapshot seems to indicate that there is still a tremendous amount of work to be done to deeply align classroom instruction with the standards.

To help teachers align instruction, the Consortium recently launched a Digital Library that contains exemplar instructional modules in English language arts/literacy and mathematics at each grade level, as well as professional learning and instructional materials contributed by teachers.[4] This shared resource bank is expected to grow significantly over time. Ultimately, in most states, districts are responsible for establishing the curriculum that guides teachers on what, when, and how teachers provide instruction to students. In

[44] The Digital Library is available to member states at a cost of $4.80 per student as part of a package of support resources that also includes the interim assessment system.

addition, it is the responsibility of district and school leaders to ensure that teachers have access to the training and materials needed to align their instruction.

"The Smarter Balanced tests are a better measure of what we *should* expect kids to know and be able to do," observed Jan Martin, the Administrator of Assessment for the South Dakota Department of Education. "Those districts that have taken the most advantage of the materials provided didn't see as much misalignment."

> "The Smarter Balanced tests are a better measure of what we *should* expect kids to know and be able to do."
>
> —Jan Martin, Administrator of Assessment for the South Dakota Department of Education

F. **The functioning of the Smarter Balanced test delivery system and help desk.**

As described above, 70% or more of test administrators in each of seven states indicated that the Field Test had gone either as well as or better than expected. With a peak of approximately 184,000 simultaneous test takers, this is better than some predicted four years ago, but also may not be representative of how well the spring 2015 administration will go, given that a) in 15 of the 21 member states only a subset of schools participated in the Field Test and may have been those with the best technology infrastructure and support, and b) many students took either the math or English language arts/literacy assessment, but not both, so the overall scheduling and system load will be greater next spring.

The most common test delivery system problems reported on surveys were:

- Unexpected log-off, which may have occurred due to lost wireless connectivity;
- Computers/servers freezing;
- Difficulty logging in;
- Test sessions timing out during long reading sections;
- Poor text-to-speech voice quality (this appeared to be the case for just one of the voice package options);

- Laptop battery life insufficient to get through a day of testing; and
- Tablets needing to be reset.

While several of these problems will be addressed locally over the coming months, test administrators also requested some test platform enhancements including an incremental rewind on video and audio files and mouse-over definitions of embedded tools. AIR, the developer of the open source delivery platform, will add the incremental rewind feature by the end of 2014; they will not be adding mouse-over definitions of tools, but this and other enhancements can be made by others to the open source code over time.

Based on the summary help desk report, the technical issue that may have impacted students most was the occasional inability to move to the next question, but this appears to have occurred in just 1 out of every 26,000 test administrations.

"The biggest surprise from the Field Test," stated Angela Hemingway, "was that the test platform and interface worked so well, across the various student devices and operating systems." She now feels confident that districts can move forward with acquiring the devices of their choice, knowing that they will deliver the assessments appropriately.

"The biggest surprise from the Field Test was that the test platform and interface worked so well, across the various student devices and operating systems."

—Angela Hemingway, Director of Assessment and Accountability for Idaho

Security breaches appear to have occurred much less often than skeptics had predicted, and the strategies developed by the Consortium to identify and stop them appear to have been effective. The externally contracted help desk documented nine test security breaches, although an unknown number were reported directly to states. Most were identified through the coordinated monitoring of social media sites and involved students taking photographs of test questions and posting them on Twitter or other social media sites. The incentive to breach security will increase significantly as states begin to use the test results for student, school, and possibly educator accountability purposes, although lessons learned during the

Field Test and the use of adaptive testing on the end-of-year component should help to curtail it.

Conclusion

The development of this assessment system by a newly formed consortium of 21states plus the U.S. Virgin Islands was a herculean effort that, as evidenced by the Field Test survey results from 13 participating states, was implemented extremely well.

As states take over responsibility for the delivery of these assessments—and as schools and districts work on final preparations for their first operational year—five themes from these survey results can help inform those state and district plans.

1. States need to customize the Test Administrator Manual and Training Modules based on their implementation choices and get those materials out to districts as soon as possible to allow sufficient time for thorough preparation, training of test proctors, and testing of the technology infrastructure at the district and school levels.
2. Schools and districts need to continue to update their technology infrastructure and ensure sufficient Internet connectivity/bandwidth in all locations that will be used for testing.
3. Students with weak keyboarding or word processing skills need opportunities to strengthen them, whether in or out of school.
4. In order to give students the full benefit of the untimed nature of these assessments and the optional breaks, schools will need to thoughtfully develop their test administration schedules.
5. Students should have an opportunity to try out the test so that the final results describe students' knowledge and skill rather than their familiarity with the test format. Teachers have several resources available to help students become familiar with the format of the test, including a practice test, training test, and the optional interim assessments.

Another theme from the survey results warrants mentioning, even though it is not related to the Field Test itself. It is clear from the responses of both teachers and students that, at least in some states, efforts to help teachers align instruction with the Common Core State Standards need to be significantly intensified, particularly in the upper grades. Students cannot be expected to perform well if the tests address skills and knowledge that they have not yet been taught.

Perhaps the most critical questions regarding these new assessments—whether they do a better job of measuring important skills and knowledge and result in improved instruction and students' readiness for college and careers—are critical topics for future studies. But Vermont's Gerdeman is optimistic: "While summative assessments are just one of several types of tests used by educators to monitor and improve learning," he explains, "the Smarter Balanced summative assessments, which will incorporate a much richer array of item types and accessibility features, represent a quantum leap from where we were."

Tony Alpert, the Chief Operating Officer for the Consortium, feels that the Field Test was "a dramatic success." There were some bumps, he acknowledges, and more will likely occur during the first operational year when all eligible students are tested and states use contracted service providers for test delivery, help desk, and scoring.

"This first year of implementation brings several new challenges, but the greatest challenge for Consortium states," Willhoft predicts, "will be getting over the anxiety." While Consortium staff will no longer manage assessment delivery, states and district personnel will not be entirely on their own. "We look forward to collaborating with them," Alpert adds, "to make it a success."

About the Author

Nancy Doorey is an education consultant who has been deeply involved in educational reform for more than 20 years, serving as a teacher, state and district policymaker, program director, and currently as a consultant in the areas of assessment, policy, and leadership. She is the lead author of the widely used summary of the six multi-state assessment consortia, now in its fifth edition, "Coming Together to Raise Achievement: New Assessments for the Common Core State Standards," produced by the K-12 Center at ETS.