NNSTOY
TEACHERSLEADING
POLICY • PRACTICE • ADVOCACY

Still on the Right Trajectory

State Teachers of the Year Compare Former and New State Assessments

National Network of State Teachers of the Year

Catherine McClellan, Ph.D.
Jilliam Joe, Ph.D.
Katherine Bassett, M.Ed.

December 2016

NNSTOY
TEACHERSLEADING
POLICY • PRACTICE • ADVOCACY

We at the National Network of State Teachers of the Year (NNSTOY) are most pleased to share with you the latest in our series of research reports.

In this report we continue our focus on the important issue of assessing our students' learning through standardized, summative assessments. Utilizing research-based methodologies and practices, including Evidence Centered Design, Webb's Depth of Knowledge, and survey instruments designed for this study, we convened a panel of outstanding educators to examine three assessment instruments. The study panel was composed of State and National Teachers of the Year and Finalists for State Teacher of the Year.

The panel examined three assessments: the grade 5 assessments given by the states of Nevada and Oregon and the grade 5 assessment developed by the Smarter Balanced consortium, to which the states have switched.

The study is a continuation of our *The Right Trajectory* research study released in November 2015. In this follow-up study, we focus on the Western part of our country as so many of our Western states have adopted the Smarter Balanced assessment.

Working with our study partners, EducationCounsel on the policy side and Clowder Consulting on the science end, we are eager to share our results. We found that participating teachers viewed state transitions to the new consortia assessments as a positive move forward. In other words, in using these assessments, we are still on the right trajectory.

The teachers identified areas for continuous improvement, such as the need to challenge our highest-performing students, the desire to minimize test length and burden while still achieving accurate measurement, and the difficulty of striking a balance between asking all students to stretch and not overwhelming them. Still, these same teachers believe that the new tests are a significant step in the right direction, and they are determined that educators and policymakers should focus together on the work ahead. They want to transform teaching and learning so that all students have the opportunity to master the knowledge and skills necessary for success in college, career and life.

At NNSTOY, we believe that educators should always be at the table when education policy is being crafted, debated or modified. As professionals, we know the most about what is likely to directly impact students and the work in the classroom, both positively and negatively. We are excited to share this paper with you and look forward to working with you in bringing the voice of educators to the policy process

With warm regards,

Katherine Bassett

Acknowledgements

NNSTOY wishes to thank the following individuals and groups for their support and contributions to this project:

Our Partners

Our partner in this work, EducationCounsel, specifically Mr. Scott Palmer, Ms. Bethany Little, Mr. Nick Spiva, and Ms. Sandi Jacobs, for their commitment to learning what a group of outstanding educators think about the trajectory that we are taking in moving to new state assessments. Their guidance, policy expertise and assistance with access to the assessments studied was invaluable, as was their overall collaboration.

Our science partner in this work, Clowder Consulting. Dr. Catherine McClellan is a consummate psychometrician and research scientist. Her vast knowledge of survey science, research methodology, and analytic ability made this research study possible. Dr. Jilliam Joe is a gifted facilitator of focus groups, and her analytic capabilities made unpacking data understandable and clear for lay people.

We thank both sets of partners for their patience, dedication and collaboration in this lengthy process.

Assessment Providers

Allowing an outside agency access to confidential assessment material is a serious undertaking. We are most grateful to the states of Nevada and Oregon for allowing us access to their prior state student assessments. We are equally grateful to Smarter Balanced for giving us access to their assessments. We protected the confidentiality of these assessments diligently and appreciate your allowing us access to them. Without this access, there would be no study.

Our Funders

We were fortunate to have generous funding with which to conduct this study, supplied by the Rockefeller Philanthropy Advisors, the Bill and Melinda Gates Foundation, and Bloomberg Philanthropies. Without this funding, this study would not have taken place. We are most grateful.

Our Reviewers

We sincerely thank the following for making the time to conduct an external review of this report: Mr. Chris Minnich, Executive Director of the Council of Chief State School Officers; Dr. Rebecca Snyder, Pennsylvania State Teacher of the Year 2009 and Past President, NNSTOY; Dr. Joshua Starr, Chief Executive Officer, PDK.

The Panelists

Finally, we could not have asked for a more prepared and committed set of educators with whom to do this work. Each panelist made certain to be well prepared for the work of the study. Each is an exemplary educator who brought intense knowledge, skill and ability to the table. Each entered into this work without preconceived ideas or opinions about the assessments. Each is a shining example of the best in education in our country and we are grateful for their participation.

Table of Contents

Still on the Right Trajectory: State Teachers of the Year Compare Former and New State Assessments

Executive Summary

Still on the Right Trajectory follows *The Right Trajectory* (2015) with additional insight on the value new state assessments add to the measurement of student outcomes. A panel of the best teachers in the country convened to give voice to critical questions about the quality of former and new state assessments, with particular attention paid to the new consortium test under study. These front-line experts believe that despite challenges still to be overcome, Smarter Balanced is an improvement on the former assessments and represents movement in the right direction for students and for education in their states.

What do great teachers think of the new assessments compared with the previous ones?

As part of state transitions to college- and career-ready (CCR) standards, including the Common Core State Standards in more than 40 states (NGA & CCSSO, 2010), states are for the first time administering new summative assessments aligned to those standards and aiming for a higher bar in assessment quality. For a majority of states, this means the "consortia assessments" – the Partnership for Assessment of Readiness for College and Careers (PARCC) or Smarter Balanced Assessment Consortium (Smarter Balanced). In this supplement to the original *Right Trajectory* study, we evaluate only the Smarter Balanced consortium assessment.

Assessment of student learning has always been an important part of education, but in recent years the use of assessment data to inform everything from instruction to accountability to policy decisions has made test quality a topic of much discussion. As the National Network of State Teachers of the Year (NNSTOY), we are deeply interested in understanding what excellent teachers – given the opportunity to closely examine new and an additional set of former assessments side by side – would think about Smarter Balanced, and informing the field accordingly. Simply put: *Does the new assessment still serve as a better reflection of what great teachers are doing in their classrooms? Does it still reflect higher quality than former state tests? Does the assessment still represent movement in the right direction?*

To answer these questions, we assembled a group of former State Teachers of the Year (STOYs) from multiple states, each of whom has been recognized at local and state levels for their teaching excellence. The panel reviewed the 5th grade Smarter Balanced and two prior state assessments: OAKS from Oregon and the Nevada state assessments (both states currently use Smarter Balanced).

What we found is clear: There was consensus across participating teachers that the new consortium assessment – Smarter Balanced – is an improvement and still the right trajectory. When compared with the prior state assessments examined in the study, Smarter Balanced illustrates where we should be headed in summative assessment.

Outstanding teachers can be powerful champions for assessment. As those closest to the process of preparing students for and administering new assessments, teachers often have the most

trusted perspective on the transition for students, parents and other educators. Their voices and support are essential if these new initiatives are to be successful. Several significant results from the study are highlighted below.

1. **The new consortium assessment remains a better reflection of the range of reading and math knowledge and skills that all 5th grade students should master.** Teachers in our study spent time meticulously examining Smarter Balanced and the former state assessments. They rated the items on the cognitive challenge required to respond to each. And while no summative assessment can capture the full range of knowledge and skills reflected in CCR teaching and learning, there was clear consensus among the teachers that the consortium assessment better reflected and measured those expectations for 5th grade students, including higher-order skills.

   For example, when asked whether they agreed with the statement: "This test measures an appropriately broad sampling of the ELA/Math knowledge and skills in instruction in an excellent 5th grade classroom," 67% of participating 5th grade teachers agreed or strongly agreed when referring to the consortium test, but less than 30%, on average, agreed when referring to the former state tests. One teacher explained: "Smarter Balanced does a better job than the other two assessments; however, the most important and authentic ELA skills are difficult to assess in this controlled type of setting."

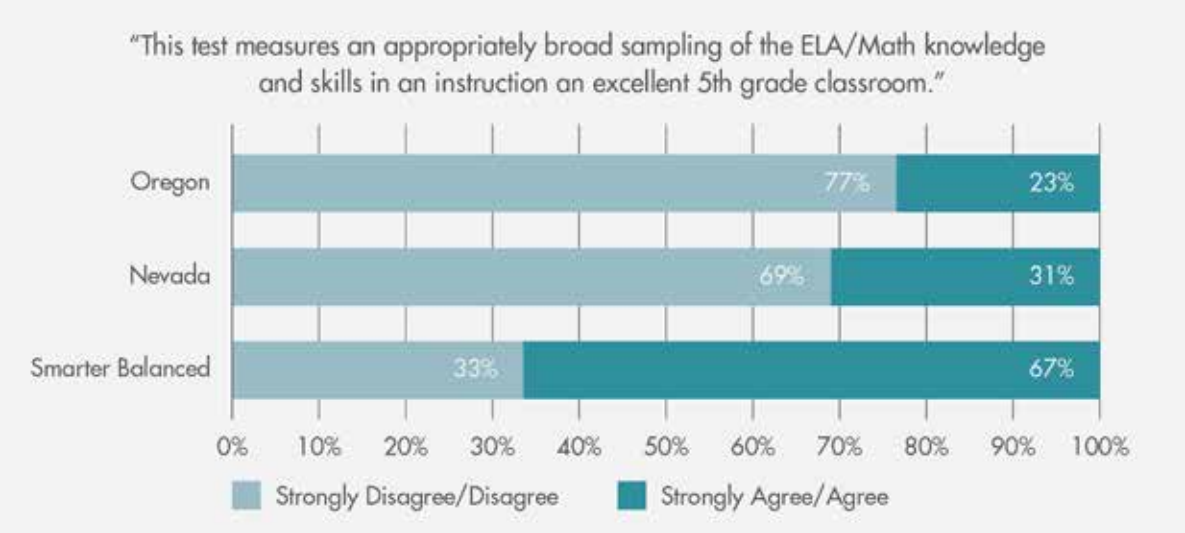


Figure 1. Percent agreement with the statement: "This test measures an appropriately broad sampling of the ELA/Math knowledge and skills in instruction in an excellent 5th grade classroom." Detail may not add to total due to rounding.

2. **The new consortium assessment is designed to include items that better reflect a full range of cognitive complexity in a balanced way at the 5th grade level.** Teachers found that items on the new consortium test required a variety of levels of cognitive demand, whereas prior assessments were characterized as lacking questions that demanded higher levels of cognitive complexity from students. One teacher commented: "When I scored the test in terms of the depth of knowledge, … the balance [on] Smarter Balanced to me was astonishing actually." When asked whether they agreed with the statement: "This test strikes a balance between the number of items that require recall responses and responses that require higher-level cognitive skills," 85% endorsed it for the 5th grade consortium test, but only 12%, on average, did so for the former state tests.

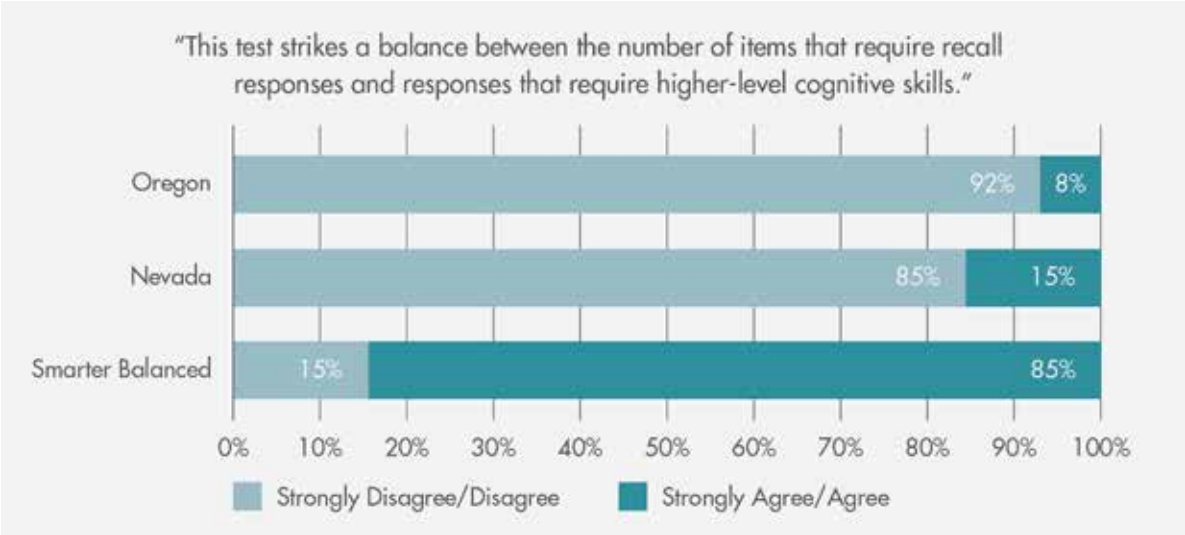


Figure 2. Percent agreement with the statement: "This test strikes a balance between the number of items that require recall responses and responses that require higher-level cognitive skills." Detail may not add to total due to rounding.

3. **The new consortium assessment better aligns with the kinds of strong instructional practices these expert teachers believe should be used in the classroom at the 5th grade level, and thereby better supports great teaching and learning throughout the school year.** The consortium assessment was perceived as a better reflection of the teaching and learning practices that occur in our very best classrooms. No standardized test captures all the activities of a classroom, but the most important skills and knowledge were represented on the consortium test. In addition, questions were asked in ways better aligned to the instructional practices of excellent classrooms than the previous assessments. As one teacher noted: "I think there are real pockets of excellence in the tests that you can see. Especially in some of the more complex questions and in some of the more in-depth questions, particularly with [Smarter Balanced]." Another told us: "I do believe that the [Smarter Balanced] tests are aligned for this. The [other] tests are moderate level tests with limited higher level critical thinking."

   These teachers found the new assessment more representative of meaningful instruction, both in content and delivery, in well-taught classrooms. All of the teachers agreed or strongly agreed "preparing students for this test would require meaningful lessons and learning, beyond skill and drill practice," but only 35%, on average, agreed or strongly agreed with the statements for the prior state tests.

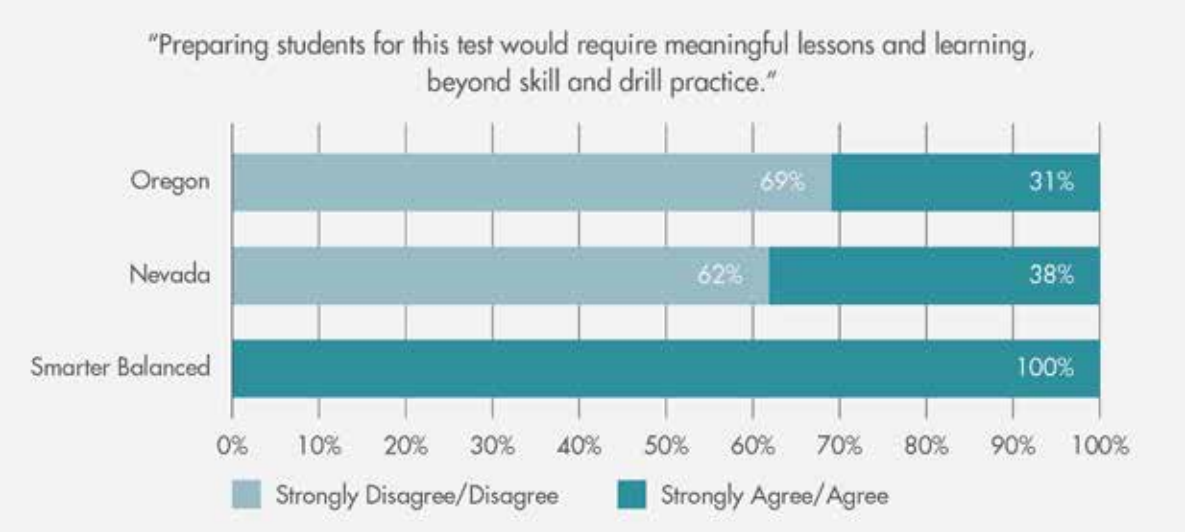


Figure 3. Percent agreement with the statement: "Preparing students for this test would require meaningful lessons and learning, beyond skill and drill practice." Detail may not add to total due to rounding.

4. **While the new consortium assessment is still more rigorous and demanding, it is grade-level appropriate, even more so than prior state tests.** The decision by states to increase the rigor of standards means that the expectations of new assessments aligned to those CCR standards also would be higher. It is important, however, that the assessment remains developmentally appropriate to the tested grade level. A strong majority of the teachers found the depth and range of content on the new test to be appropriate for 5th grade students. There was variation between state assessments in teachers' opinions of the appropriateness of the range of content; 62% strongly agreed or agreed the range was appropriate across the Nevada assessment items, compared with the 46% who strongly agreed or agreed across the Oregon assessment items.

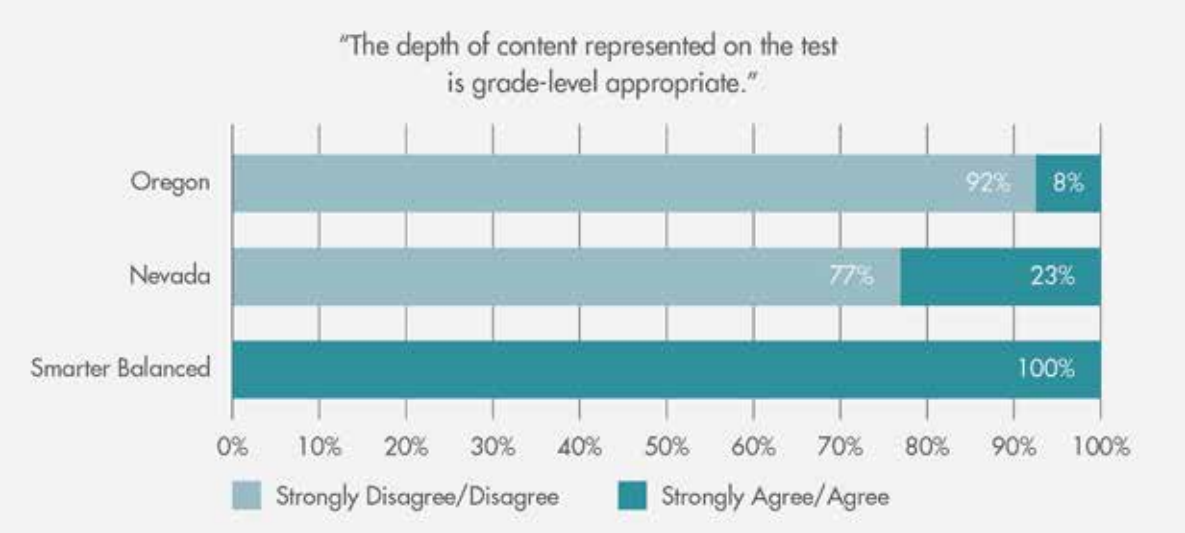


Figure 4. Percent agreement with the statement: "The depth of content represented on the test is grade-level appropriate." Detail may not add to total due to rounding.

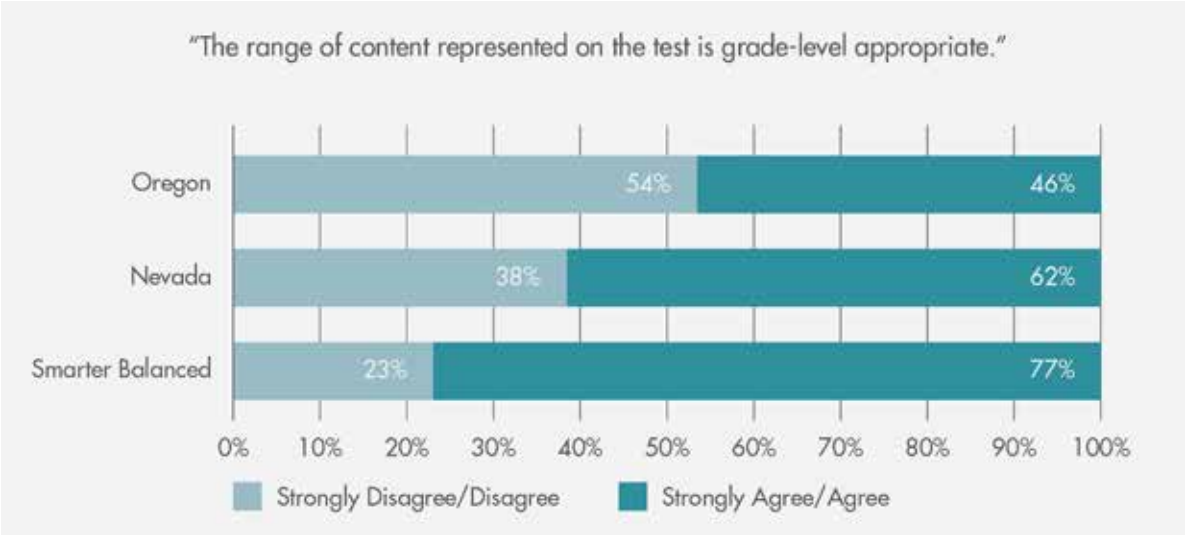


Figure 5. Percent agreement with statement: "The range of content represented on the test is grade-level appropriate." Detail may not add to total due to rounding.

Summary: The Transition to the Smarter Balanced Assessment is Still Worth it

The data gathered from our best teachers are compelling and continue to indicate that transitioning to the Smarter Balanced assessment is still worth it. Our teachers emphasized a need for alignment between assessments and excellent classroom instruction. The Smarter Balanced assessment remains on the right trajectory toward meeting that goal. Teachers also acknowledged the assessment not only reflects what they do in the classroom, but it also has the potential to inform improvements to their practice and help move the teaching profession forward. One teacher expressed this idea in this way:

> "I think one big takeaway for me is that [as] teachers, we've been brutalized by assessment in certain areas. I think that this really gives us [a chance] to look at what an assessment can be and what it can do and how it can really be part of your classroom so that teaching to the test wouldn't be a negative. If the test was really intelligently designed, it should be what you're doing [in the classroom]."

Overview of Study

A team convened by the National Network of State Teachers of the Year set a research direction for examining the differences between former and new state summative assessments (in this study, Smarter Balanced consortium's end-of-year assessments) through excellent teachers' perspectives. Do educators view the new consortium assessments as high-quality measures of important knowledge, skills and abilities taught in their classrooms? Using evidence-centered design as a framework (Mislevy, Almond, & Lukas, 2003), the team identified several claims that proponents of the new assessments would hope to have supported by educators. Three primary claims resulted from this exercise, in Figure 6.

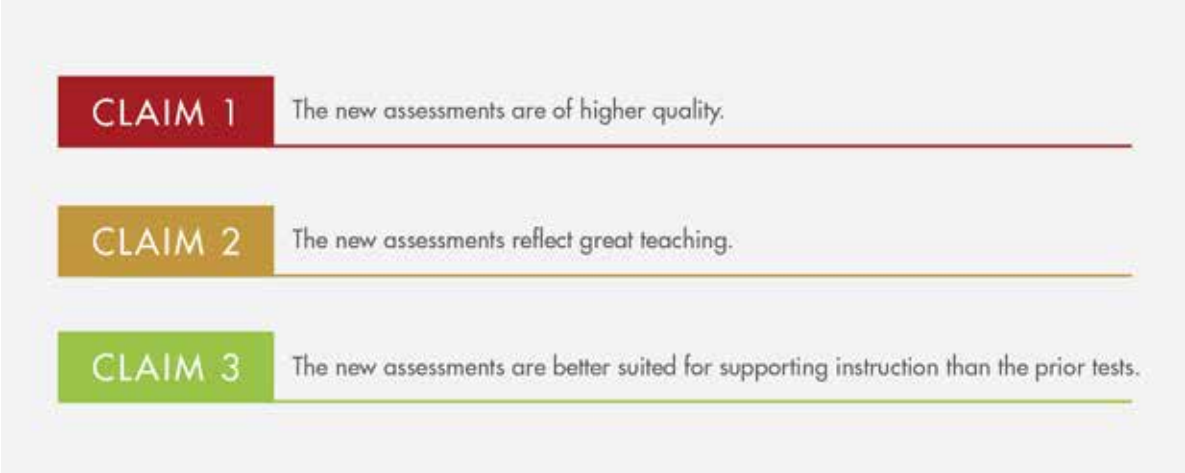


Figure 6. Claims made about the Common Core Assessments.

A small-scale study was designed in 2015 to gather evidence against which to evaluate these claims through an in-depth evaluation and comparison of former and new state assessments by teachers. The research team organized the study around five key questions:

1. Do the new consortia assessments better reflect the range of knowledge and skills that all students should know?
2. Are the new consortia assessments designed to better reflect the full range of cognitive complexity in a balanced way?
3. Do the new consortia assessments better align with the strong instructional practices these teachers use in the classroom, and thereby better support great teaching and learning throughout the school year?
4. Do the new consortia assessments provide information relevant to a wide range of performers?
5. While the new consortia assessments are more rigorous and demanding, are they grade-level appropriate, and more or less so than prior state tests?

Together these five areas would provide a picture of assessment quality from the perspective of teaching and learning. If the new assessments are to have greater efficacy than the former assessments, they must address each area. Indeed, a common criticism of former K-12 state assessments is their failure to measure the kinds of outcomes teachers deem important. We intentionally designed the study so that teachers would have an authentic opportunity to evaluate both prior state assessment forms at grade 5 and new consortia assessment forms on their own merits. We used a neutral alignment tool (described below) and designed a rubric (described in the Appendix) focused on general assessment quality issues, rather than any particular set of learning standards. The results of this study are detailed in *The Right Trajectory* (2015). The current study was designed to explore these research questions with a different panel of teachers given an additional set of former state assessments and Smarter Balanced assessment at grade 5. The findings of this investigation are described in this report.

Methodology

The study comprised an in-depth review and alignment of two former state assessments and one new state assessment. The panel examined two former assessments and new 5th grade Smarter Balanced assessment over two days. The study plan is described below.

The review was conducted using Norman Webb's Depth of Knowledge (DOK; Webb, 1997) framework. The DOK framework guided participants' orientation to each of the assessments used in this study. In preparation for this alignment work, each panel participated in an online webinar exposing them to DOK. In addition, each panelist was asked to prepare for the study panels by examining their own state's standards in Math and English Language Arts (ELA).

The DOK levels are intended to be neutral to the content standards that underlie a particular set of items. For this reason, Webb's DOK is widely accepted as a useful framework for classifying the cognitive demand required of students on assessment items and tasks. There are four DOK levels, each with increasing complexity or cognitive demand, as shown in Figure 7.

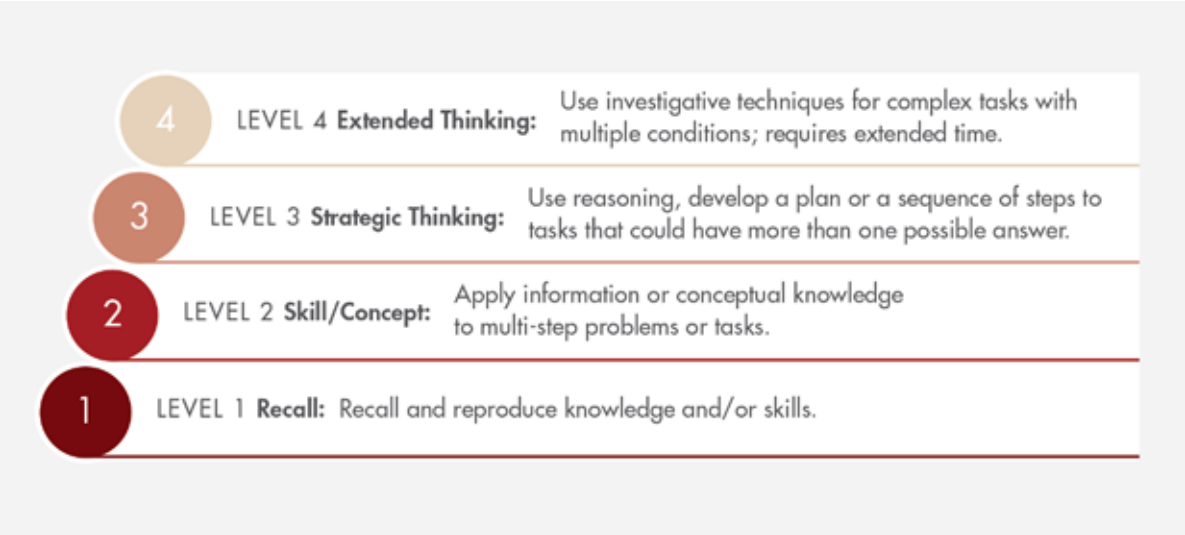


Figure 7. Webb Depth of Knowledge Levels (Webb, 2005).

A mixed-methods approach defined the second phase of the study. Mixed-methods designs employ qualitative and quantitative data collection techniques to allow for both depth and breadth in an investigation. Given the types of questions explored within this study and the relatively narrow pool of participants targeted (those who were selected as State Teachers of the Year or were Finalists), this design seemed the most appropriate. The quantitative component included a 58-item survey that was designed to capture teachers' perceptions of the quality of former and new state assessments. It also included a short, 8-item pre-post measure of teachers' attitudes toward tests and test items. The surveys were followed by a whole-group discussion to elicit additional information about findings we thought would be useful to explore.

State K-12 Assessments and Survey Instruments

Two former 5th grade state ELA and Math assessments and one new consortium summative assessment at grade 5 were reviewed for the study. These assessments were reviewed by a panel of expert teachers. They reviewed the 5th grade Oregon Assessment of Knowledge and Skills (OAKS), Nevada's state assessment, and the Smarter Balanced consortium test[1]. The former state assessments were all viewed in paper format. The new state assessment is computer-based. Appendix A includes more details about the reviewed assessments.

Fifth grade was chosen as the grade level to review since it is on the cusp between elementary and middle school, making assessments at that grade relevant to elementary and middle school teachers and students alike. In addition, using 5th grade allowed us to draw from a vertical range of educators with knowledge of the content that students would be expected to know at this grade level, both above and below it, to give us a wider spectrum of educators from which to choose.

Two online survey instruments were developed for this study. The Attitudes Toward Tests survey was designed by the research team to capture teachers' perceptions about tests and item types. Educators can hold preferences for how best to measure student knowledge and skills. We thought it important to understand what these preferences were for participants prior to and after engaging with the assessments. Generally, we found that the panelists preferred tests with constructed-response items over selected-response items, and that the review process improved their ratings of the types of skills selected-response items could assess. However, none of the shifts in the survey responses were large and thus seem unlikely to have impacted the results meaningfully.

The Survey of Assessment Quality was developed to evaluate the five key areas of quality of the assessments listed above. These items addressed the appropriateness and rigor of the items for low-, mid- and high-performing students; the content; performance levels; balance; and grade appropriateness of the items in each of the assessments overall. In addition, a background questionnaire was created to gather relevant demographic and background information about participants. All instruments underwent several reviews prior to their final use. The surveys are provided in Appendix B.

Participants

We convened 13 outstanding educators for the study, each a State Teacher of the Year or Finalist recognized for excellence in classroom practice. The panel was designed to represent diversity along several measures:

- **Content area**. We selected panelists with rich teaching experience in either Math or ELA;
- **Grade level**. We focused on 5th grade assessments as a transition point between elementary and middle school. We included teachers with familiarity of the 5th grade content through vertical grade-level alignment.

1 OAKS and Smarter Balanced are adaptive tests, but teachers only reviewed one linear form based on a student at the 60th percentile of the proficiency distribution at 5th grade.

- **States.** We included teachers from each of the states whose assessments we examined, and we sought geographic diversity. The group of participants included teachers from Colo., Idaho, Minn., N.J., Nev., N.Y., Ore, Penn., S.D. and Wash.
- **Race/ethnicity and gender.** We sought to reflect the racial/ethnic and gender makeup of the general teaching population to the extent possible;
- **School setting.** We worked to bring together panelists from a variety of school settings, e.g. rural, suburban, urban.

There were three or more teachers representing the state in which the prior assessments were administered on each panel. Several teachers were included from a state that is not using the new assessment reviewed in the study. In terms of content area, we ensured there was an equal balance of Math and ELA teachers. We were careful to select teachers possessing familiarity with 5th grade instruction for the 5th grade assessment evaluation. More detailed demographic data on the panelists is presented in Appendix C. For taking part in this study, participants were given a stipend for their time and reimbursed for expenses incurred for travel, lodging and food. No other compensation was provided. All of the teachers in the study were members of the National Network of State Teachers of the Year, either State Teachers of the Year or Finalists for State Teacher of the Year in their respective state. These teachers are a group whose instructional practices may be presumed to represent "excellence" in the classroom, and their knowledge of teaching formed the basis of their judgements about the assessments.

Data Collection

The review process drew on participating teachers' existing areas of expertise: how well the assessments reflect the kind of teaching and learning that they want to see in the classroom. The panel met in Las Vegas, Nev. for two days of onsite activities in early August. We employed a four-step data gathering process. Each of these steps are described briefly in the sections that follow.

1. Training and orientation (including the Attitudes Toward Tests survey)
2. Webb DOK alignment
3. Assessment review using the Survey of Assessment Quality
4. Focus group discussion

Cognitive Demand of Assessment Items

Before arrival, participating teachers received pre-reads and other materials to jumpstart their understanding of the process. Participants used Webb DOK, a commonly used framework, to evaluate the assessment items. Webb DOK provided the educators with a vocabulary and reference point for understanding content complexity in assessments and other educational tools (e.g., curriculum units and lesson plans). They viewed a one-hour online training session on Webb DOK levels facilitated by one of the lead researchers.

Upon arrival at the study site, participants were given an introduction to the study and the research team. Each signed an informed consent and completed the Attitudes Toward Tests online survey and a demographics background questionnaire. Data collection started with a brief review of

Webb DOK, led by the researcher who provided the initial training. Next, the participants were given an orientation to the assessments they would be reviewing. During the orientation, participants were encouraged to work through the items as if they were a typical well-prepared 5th grade student, not necessarily the kind of student who happened to be in their individual classrooms. The definition used for the study does not include many students in 5th grade, and panelists discussed why it is not possible to do this type study with every 5th grade student in mind. To improve the consistency of the response data, this definition provided a common lens through which to evaluate the cognitive demand associated with a particular assessment item.

Who is the 5th grade student for this study?

- The "5th grade student who is at grade level" for this study is a student who has been well-served by the education system in your state.
- Think about a student who is not exactly typical but who has:
    - been well taught and prepared,
    - isn't special needs (since we are excluding such students from this study), and
    - had acceptable opportunities to learn and be taught before arriving in the 5th grade.
- Not the best student you ever taught, but not one strongly disadvantaged by his or her circumstances either.

Teachers participated in a consensus discussion around the DOK levels. It was important to us that the educators demonstrated consistency in their interpretation and application of DOK levels. Publicly available items for 4th through 7th grade assessments in ELA, Math and Social Studies from other state assessments (Kentucky Department of Education, 2007; Southern Nevada Regional Professional Development Program, 2009) were presented to the group and levels assigned by the teachers. The facilitator then led participants through a discussion of why a particular DOK level was selected for an item and why the adjacent levels were not, with the goal of achieving internal consistency in the panel. The panel reviewed six sample items. DOK ratings were quite similar even at the beginning of this process.

After the consensus activity, the order in which the assessment was reviewed was randomly assigned to participants. The order of review was different for panelists to mitigate fatigue effects on the data. For example, one group of participants started with the OAKS, another group started with the Smarter Balanced assessment, and the third group started with the Nevada Assessment. Participants were given a tutorial on how to access and navigate the computer-based Smarter Balanced Assessment.

Paper copies of the other state assessments were distributed. Participants were given approximately two hours to complete their review of each assessment. Depending on his background, a participant focused his review on either the ELA or the Math section. Each item from each assessment was assigned a DOK level of 1 to 4. If participants were not certain about which DOK level the item belonged to, they were instructed to indicate, "I don't know." Ratings were entered into a spreadsheet and submitted to the research team at the end of the day.

Many of the assessments in this study have used Webb DOK or other evaluations of cognitive complexity completed in studies where that was the focus of the work. The goal of this activity in our study was to assure that the participants engaged deeply and carefully with the assessment items, and that they had a common framework and language when discussing the items with each other. The primary focus of this study was the responses to the Evaluation of Assessment Quality survey and the discussion that followed for each panel. Given this, we present only a brief

summary of the Webb DOK results here.

Findings

The major findings from the DOK ratings assigned by the participants were that the former state ELA assessments largely comprised Level 1 and 2 items. These items require students to recall and apply information and conceptual knowledge. There was a marginal increase in the cognitive complexity seen in the consortium test; items were generally in the Level 2 and 3 range. Level 2 and 3 items require students to apply information and conceptual knowledge, as well as logic and reasoning. On the math assessments, teachers judged that the majority of former state and new assessment items are written at DOK Level 1 and 2. Students are generally expected to recall and apply information and conceptual knowledge. The new state assessment did not provide the expected increase in cognitive complexity. These findings differ from the previous study, where teachers noted a slight increase in cognitive complexity for consortium math assessment. We anticipated some variation in DOK results, as the purpose of the activity was to help teachers engage with the assessment items, not train them as DOK experts or have them classify items as DOK experts would.

Evaluation of Assessment Quality

Once teachers aligned the assessment items with the DOK levels, they moved on to evaluate the quality of each assessment more holistically. Prior to completing the Survey of Assessment Quality, participants were given a brief orientation to the next set of activities. They were reminded to consider the "well-served 5th grade student" and only 5th grade Math and ELA instruction—not other content areas—when evaluating the quality of the assessments. Some time was devoted to discussing formative assessment. Several of the survey items address formative assessment practices as they relate to the summative state assessment content. It was important to clarify that the test items were not developed for the purpose of formative assessment. Our interest was in determining the degree to which the content (i.e., concepts and topics) of the items might be useful for supporting and developing teachers' formative assessments. Participants were allowed to reference their ratings from the DOK alignment exercise as they completed their evaluation of assessment quality.

After participants completed the Survey of Assessment Quality, they were given a break to allow the research team time to review the survey results and the participants time to rest. Items with interesting or unclear responses were selected for clarification during the whole-group discussion. The discussion was recorded using audio equipment, with participants' permission, and for analysis. The protocol used for the discussions is located in Appendix E along with some specific questions used to guide the discussions for each panel.

Results

Recall that the five focus questions of the study, against which the original claims were to be tested, were:

1. Does the new consortium assessment better reflect the range of knowledge and skills that all

students should know?

2. Is the new consortium assessment designed to better reflect the full range of cognitive complexity in a balanced way?
3. Does the new consortium assessment better align with the strong instructional practices these teachers use in the classroom, and thereby better support great teaching and learning throughout the school year?
4. Does the new consortium assessment provide information relevant to a wide range of performers?
5. While the new consortium assessment is more rigorous and demanding, is it grade-level appropriate, and more or less so than prior state tests?

This section highlights some of the most pertinent findings from the Survey of Assessment Quality. Qualitative data from the follow-up discussions are integrated with our summary of the survey data, where appropriate for clarification and illumination. Note that in some cases, the response categories have been combined to simplify the visual presentation in charts.

Question 1: Range of important knowledge and skills

Panelists were asked the extent to which they agreed with the statement: "The distribution of content on the test is representative of excellent 5th grade instruction." The results are shown in Figure 8. The new 5th grade assessment better reflects the range of reading and math knowledge and skills that all students should master. There were mixed results between the state assessments. A larger percentage of teachers strongly agreed or agreed that the distribution of content on the Nevada assessment is more representative of excellent 5th grade content than the OAKS.

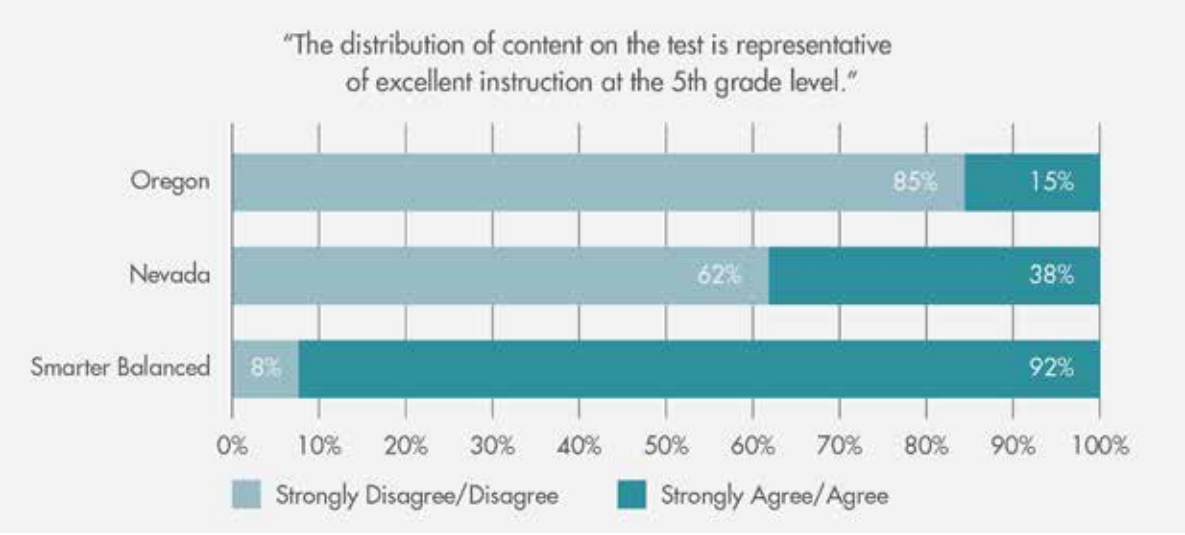


Figure 8. Percent agreement with statement: "The distribution of content on the test is representative of excellent 5th grade instruction." Detail may not add to total due to rounding.

Teachers were asked to rate their agreement with the statement, "This test measures the most important knowledge and skills to be taught in an excellent 5th grade Math/ELA classroom" for all three tests in their panel. The results are shown in Figure 9. A majority of teachers, 77%, strongly agreed or agreed that the Smarter Balanced assessment measures the most important knowledge and skills to be taught in an excellent 5th grade Math/ELA classroom. In contrast, an average of

38% of the teachers strongly agreed the former state assessments measured the most important 5th grade knowledge and skills. Teachers rated the Nevada state assessment higher than the OAKS.

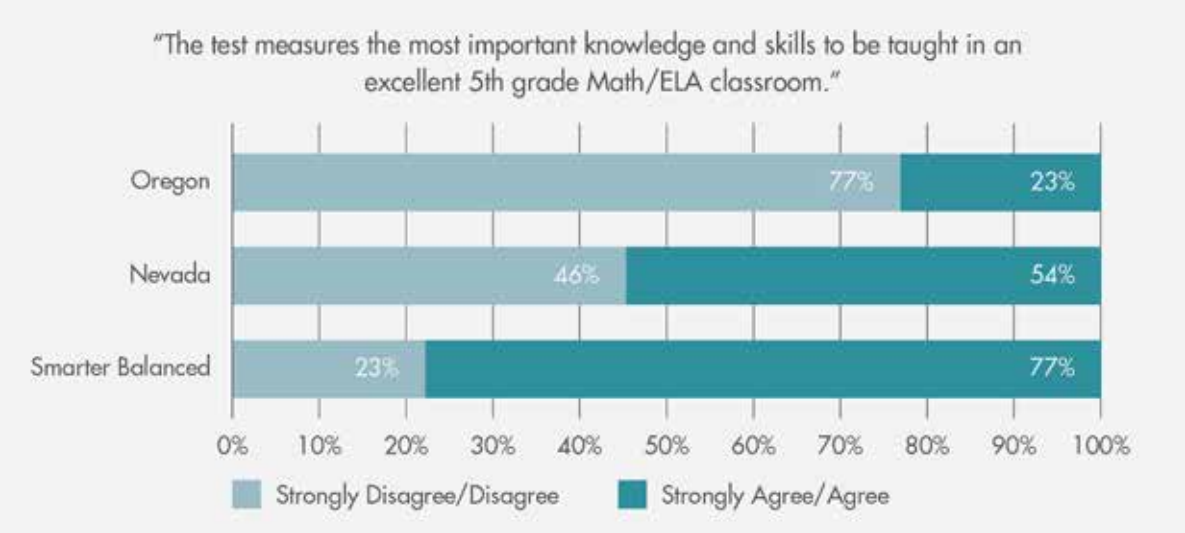


Figure 9. Percent agreement with statement: "This test measures the most important knowledge and skills to be taught in an excellent 5th grade Math/ELA classroom." Detail may not add to total due to rounding.

The next concern we addressed is whether the consortium test measures the full range of cognitive complexity that is important to teachers of 5th graders. While no assessment appears to have tapped into the full range of cognitive complexity perfectly, the data suggest that the balance achieved on the 5th grade consortium assessment, in particular, is clearly an improvement on the former state assessments.

Question 2: Assesses full range of cognitive complexity in a balanced way

Teachers tended to agree or strongly agree that the 5th grade Smarter Balanced tests balance the number of items that require recall responses with those that require the application of higher-level cognitive skills. They thought the opposite was true for the former assessments reviewed. Teachers also tended to disagree or strongly disagree that the former assessments balanced recall and higher-level cognitive items. As reflected in their DOK ratings, for example, the former assessments emphasized lower-level skills rather than the kinds of skills that would require strategic or extended thinking. The data are shown in Figure 10.

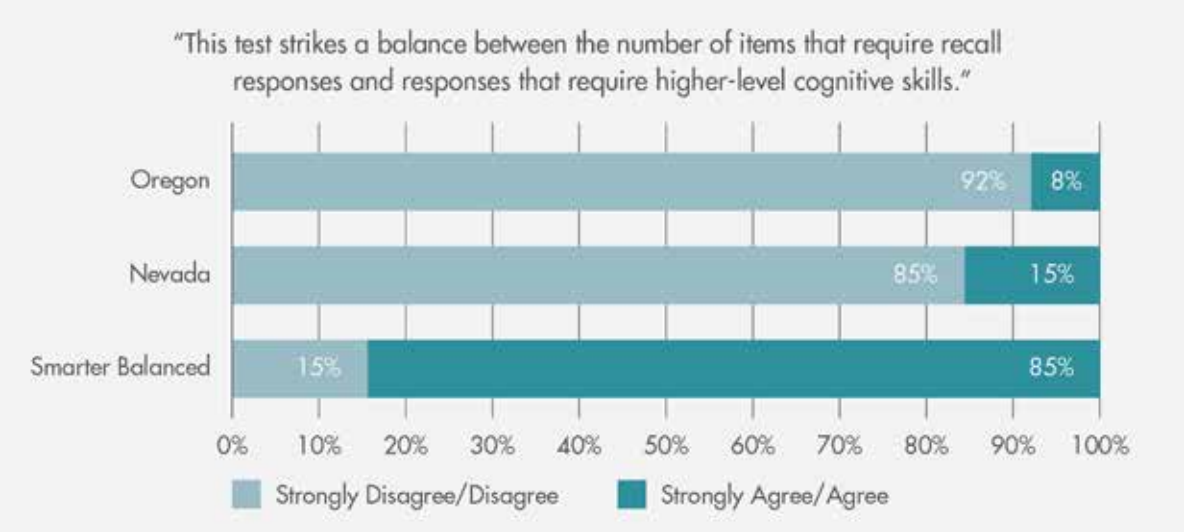


Figure 10. Percent agreement with the statement: "This test strikes a balance between the number of items that require recall responses and responses that require higher-level cognitive skills." Detail may not add to total due to rounding.

Another set of items asked about test questions that required students know and use different types of cognitive processes and placed a variety of levels of cognitive demand for response. The panelists' responses are shown in Figures 11 and 12.

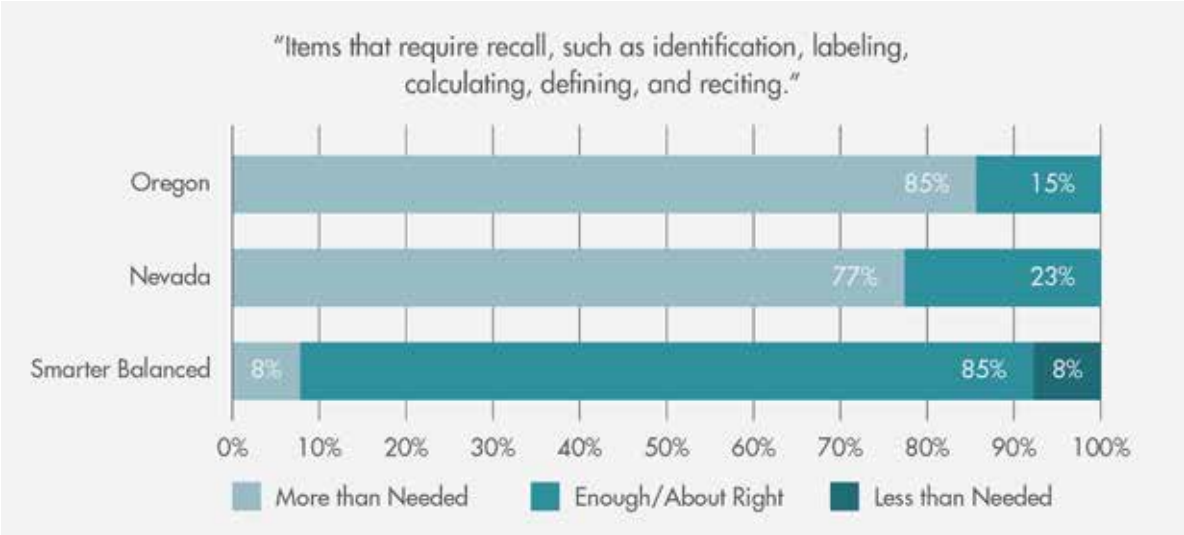


Figure 11. Percent of teachers who indicated the number of 5th grade "Items that require recall, such as identification, labeling, calculating, defining, and reciting" was "more than needed," "about right/enough," or "less than needed." Detail may not add to total due to rounding.

Most of the assessments, both former and consortium, were rated as having either more than needed or enough/about the right number of items at this level of cognitive demand. This description is consistent with items typically aligned with Webb DOK Level 1.

In Figure 12, the response data for the next type of test item is shown. Again, most of the assessments are rated as having enough/about the right number of items at this level of cognitive demand, with the consortium assessment receiving the highest ratings in this category. Several also received ratings indicating that there are fewer items than needed of this type as well (e.g., Nevada).

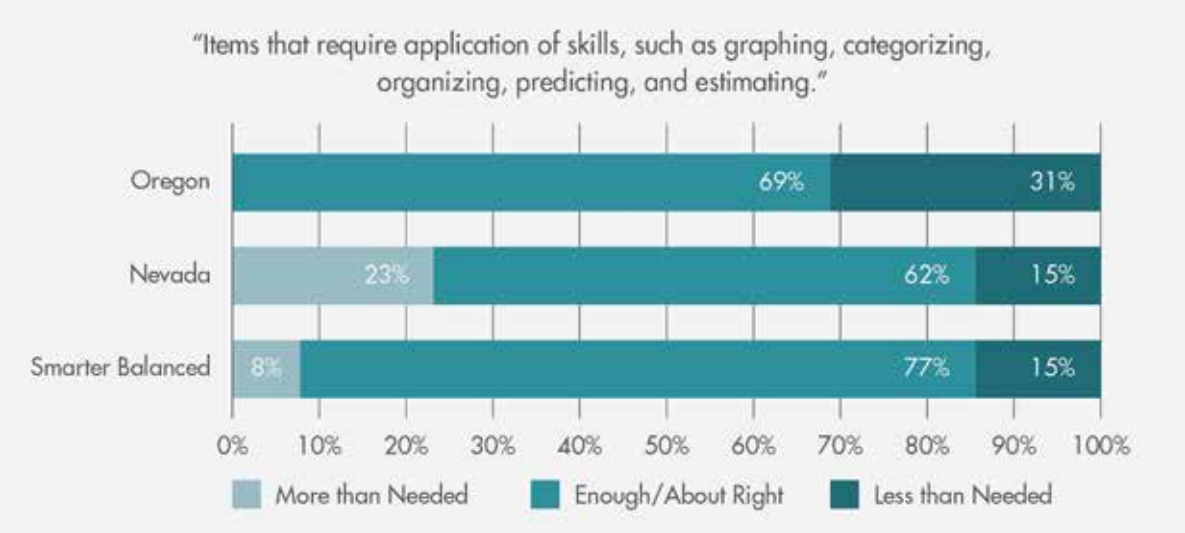


Figure 12. Percent of teachers who indicated the number of 5th grade "Items that require application of skills, such as graphing, categorizing, organizing, predicting, and estimating," was "more than needed," "about right/enough," or "less than needed." Detail may not add to total due to rounding.

There was clear consensus that the 5th grade Smarter Balanced does a better job of measuring higher-level cognitive skills than the former assessments. For example, when asked to rate whether each assessment had enough items that "require students to demonstrate strategic and extended thinking such as investigation, analysis, and design," teachers typically viewed the consortium test as having about the right amount or enough items of this type (Figure 13). These findings are consistent with the results from *The Right Trajectory* (2015).

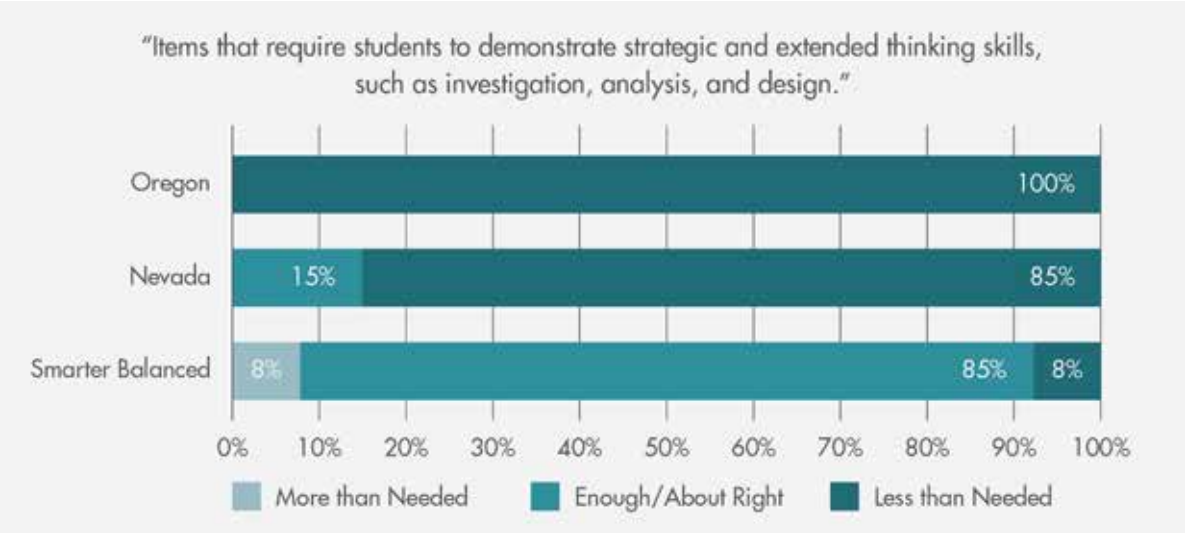


Figure 13. Percent of teachers who indicated the number of 5th grade "Items that require students to demonstrate strategic and extended thinking such as investigation, analysis, and design" was "more than needed," "about right/enough," or "less than needed." Detail may not add to total due to rounding.

In contrast, the former assessments were perceived as having gaps in their measurement of the kinds of deep learning that take place in 5th grade or teacher participants' classrooms. An average of 92% of the teachers indicated the former assessments contained fewer items than needed that "require students to demonstrate strategic and extended thinking such as investigation, analysis, and design." Only the 5th grade Smarter Balanced assessment was considered by any panelist to have more items than needed of this type (the lightest green-colored bars in Figure 13).

Question 3: Instructional practices and support for great teaching and learning throughout the school year

There was strong consensus that the consortium assessment measured excellent 5th grade instruction. The views concerning the former assessments, although varied, indicate the former assessments were not representative of excellent 5th grade instruction.

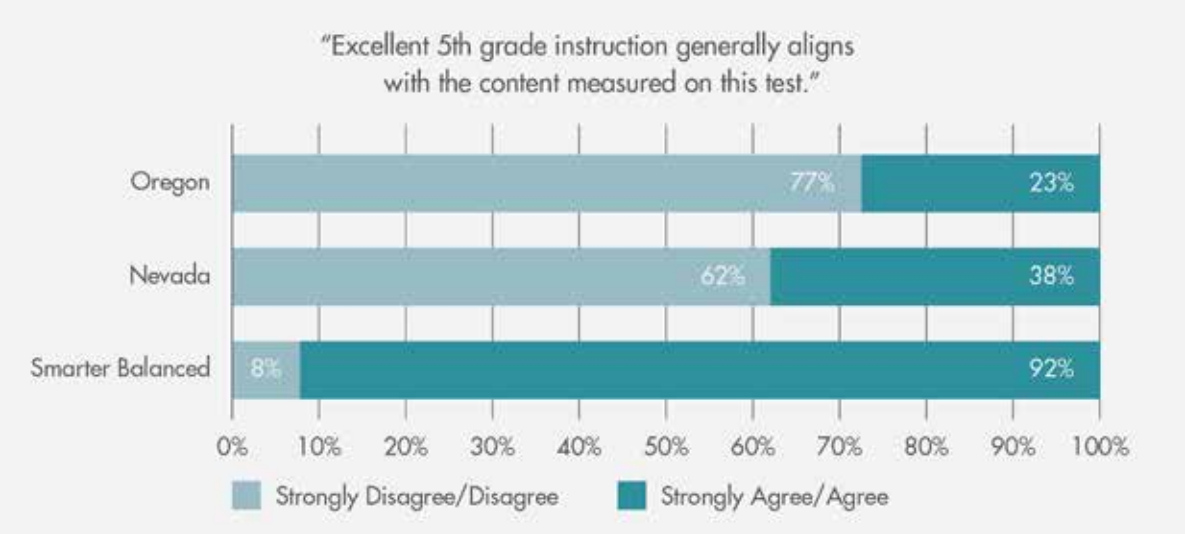


Figure 14. Percent agreement with statement: "Excellent 5th grade instruction generally aligns with the content measured on this test." Detail may not add to total due to rounding.

In addition, the Smarter Balanced 5th grade test measures the learning outcomes that participant teachers would set for student learning in 5th grade classes. As shown in Figure 15, approximately 92% of teachers strongly agreed or agreed with this statement in regards to the consortium test. In comparison, 15% and 46% of teachers strongly agreed or agreed with this statement in regards to the OAKS and Nevada assessments, respectively.

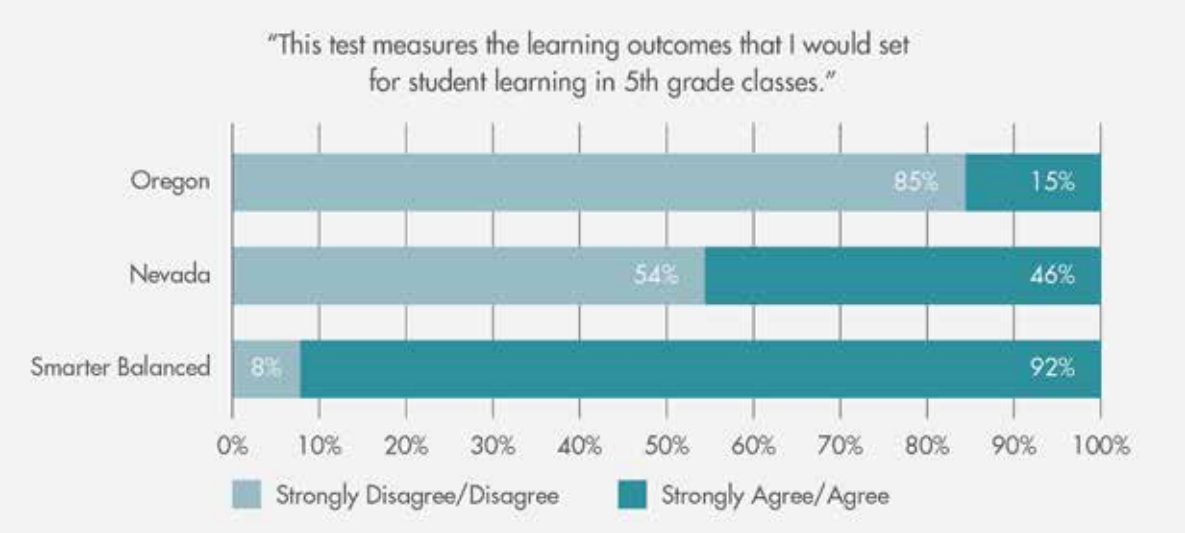


Figure 15. Percent agreement with statement: "This test measures the learning outcomes that I would set for student learning in 5th grade classes." Detail may not add to total due to rounding.

The data also show that 92% of the teachers strongly agreed or agreed with this statement when evaluating the consortium test: "One criterion for a high-quality assessment is that the assessment is designed to measure whether underlying concepts have been taught and learned, rather than reflecting mostly test-taking skills or reflecting out-of-school experiences. This test meets that criterion." The responses are summarized for the three assessments in Figure 16.

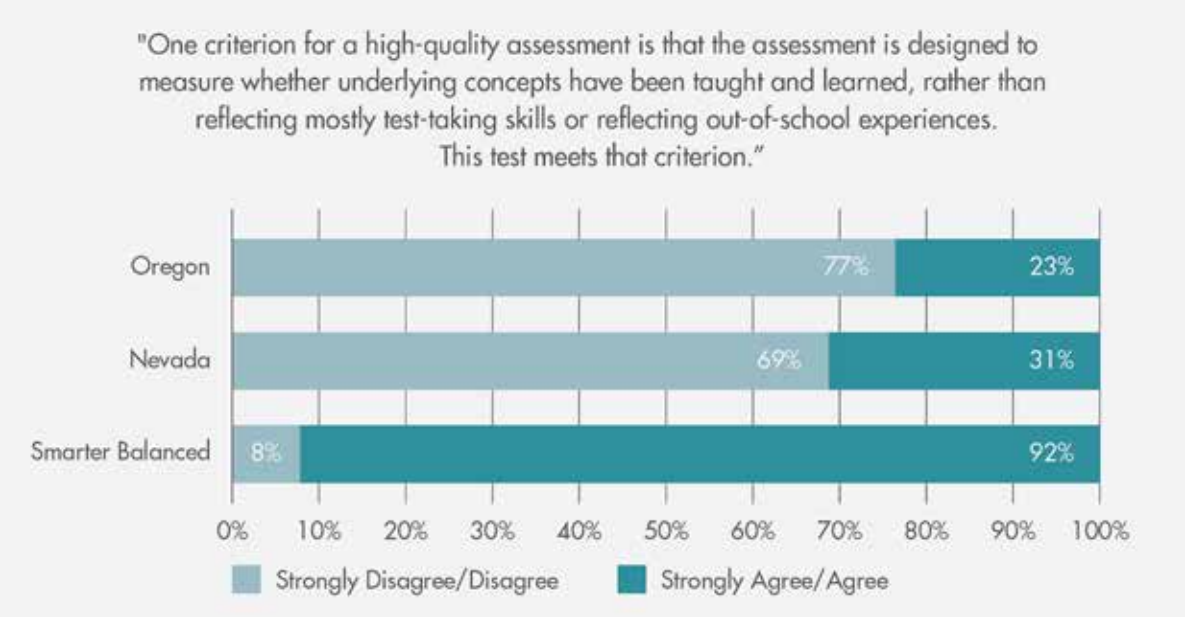


Figure 16. Percent agreement with statement: "One criterion for a high-quality assessment is that the assessment is designed to measure whether underlying concepts have been taught and learned, rather than reflecting mostly test-taking skills or reflecting out-of-school experiences. This test meets that criterion." Detail may not add to total due to rounding.

Participating teachers found the new assessment more representative of meaningful instruction in well-taught classrooms, both in content and delivery. For the consortium assessment, all of the 5th grade teachers agreed or strongly agreed "preparing students for this test would require meaningful lessons and learning, beyond skill and drill practice," but, on average, only 35% of the 5th grade teachers agreed or strongly agreed with the statements for the prior state tests (Figure 17).

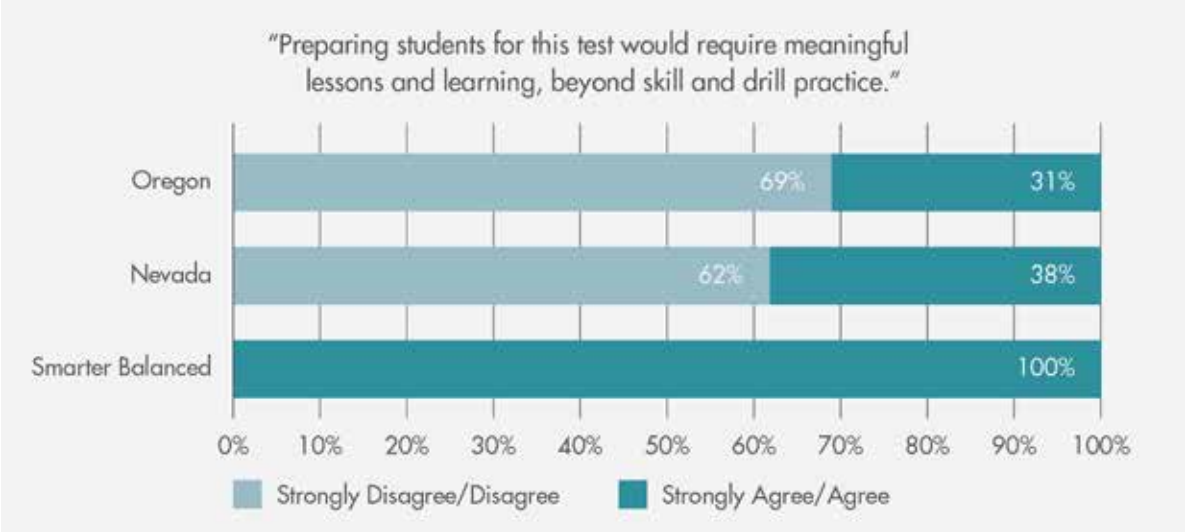


Figure 17. Percent agreement with the statement: "Preparing students for this test would require meaningful lessons and learning, beyond skill and drill practice." Detail may not add to total due to rounding.

When the topic of teaching to the test was raised, one teacher summarized the group's general perception:

> "I think by teaching the test, like that has such a negative connotation to it, but you know, think about it… the problem is some of our tests are so benign and they're not authentic enough that yeah, you don't want to teach that."

Another teacher followed that statement with this:

> "If I was teaching to these tests [referencing former state assessments] I kind of feel bad about that. Because I'm only really addressing DOK 1 and DOK 2, **disproportionately** to 3 and 4. But the Smarter Balanced, because the DOKs are addressed with 30%, 30%, 30%, 10%—I can confidently say, yeah, I feel really good about teaching to this test, because I'm getting past knowledge and comprehension. I'm getting to critical thinking, getting to application, getting to analysis because the standards of the test say that I should. I shouldn't live in DOK 1 and 2. Like, I'm forced to be an excellent educator if I'm really teaching to the test. Because if I'm not an excellent educator, I can't get to 3 and 4 and then then my practice needs to improve. So, the question about will this test help me improve my practice? With Smarter Balanced, absolutely. Because if I put those questions in front of students, there's no option but to do a good job teaching."

Question 4: Information relevant to a wide range of performers

The Smarter Balanced assessments generally provide information that is relevant to mid- and high-performing students, which is consistent with the previous report's findings. Specifically, when asked if low-, mid- and high-performing students would perform well on the assessment, participating teachers indicated that for both the former assessments and the consortium assessments low-performing students would generally not perform as well as mid- and high-performing students. One 5th grade teacher noted the following:

> "To do well at the Smarter Balanced, my honest assessment is you really have to know your stuff and you're going to [be] pushed to think in new ways and to really apply knowledge that—in a way that you need to have mastery of it to do well."

Teachers were asked if there were less than, enough, or more than the number of items that would surface information about 5th grade students at higher ability levels to inform instructional strategies. The majority indicated that the consortium assessment had enough of those items, as shown in Figure 18.

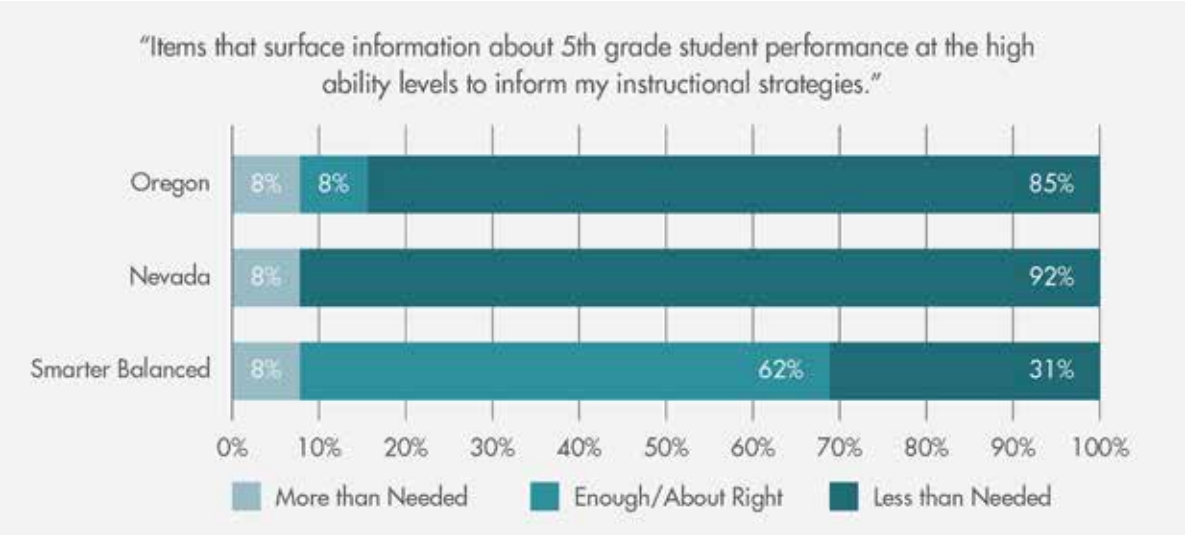


Figure 18. Percent agreement with the statement: "Items that surface information about 5th grade student performance at the high ability levels to inform my instructional strategies." Detail may not add to total due to rounding.

Question 5: Grade appropriateness

Finally, evidence shows that that the new consortium test measures the learning outcomes that the teachers believe are appropriate for student learning in 5th grade classes. One concern heard frequently is that the consortia assessments may be too challenging for students, who may find them overwhelming or confusing. Assessments should always be fair to the candidates sitting them and at an appropriate level of cognitive demand. Survey questions were included to evaluate the participants' perceptions of the former state and the new consortium assessments once they had reviewed them to get at this issue.

When asked to rate their agreement with the statement: "This test is less cognitively demanding than is warranted for the 5th grade level," 73% of the participating teachers either strongly agreed or agreed across the two former state assessments. The rigor and challenge of the assessments did not meet the rigor of 5th grade instruction. A much smaller percentage, 15%, strongly agreed or agreed with that statement for the Smarter Balanced Assessment.

A strong majority of the teachers found the depth and range of content on the Smarter Balanced test to be appropriate for 5th grade students, as shown in Figures 19 and 20. There was variation in teachers' opinions of the appropriateness of the range of content on the former state assessments: 62% strongly agreed or agreed the range was appropriate across the Nevada assessment items compared to the 46% who strongly agreed or agreed across the Oregon assessment items.

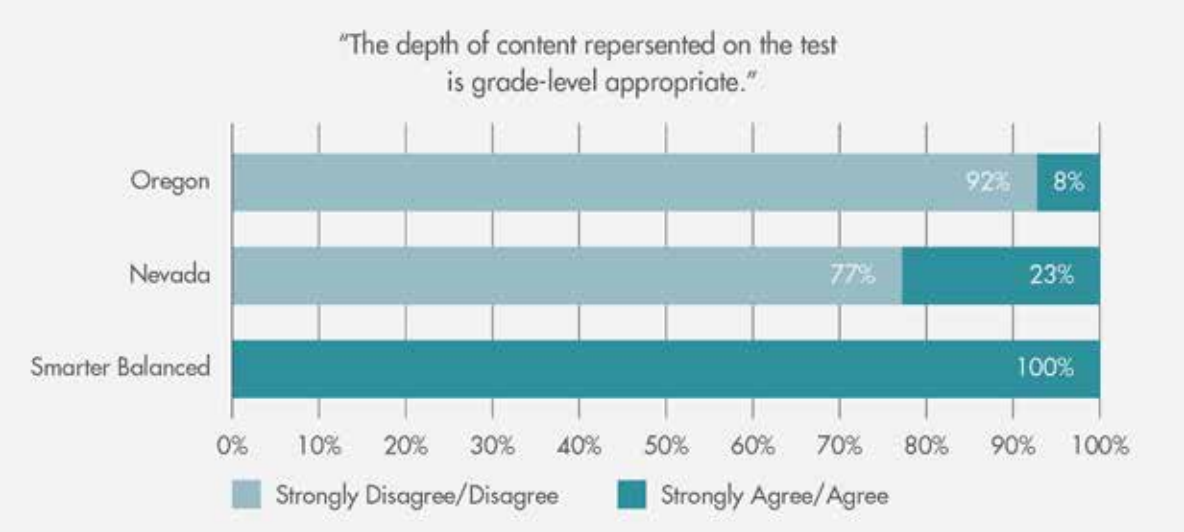


Figure 19. Percent agreement with the statement: "The depth of content represented on the test is grade-level appropriate" for 5th grade assessments. Detail may not add to total due to rounding.

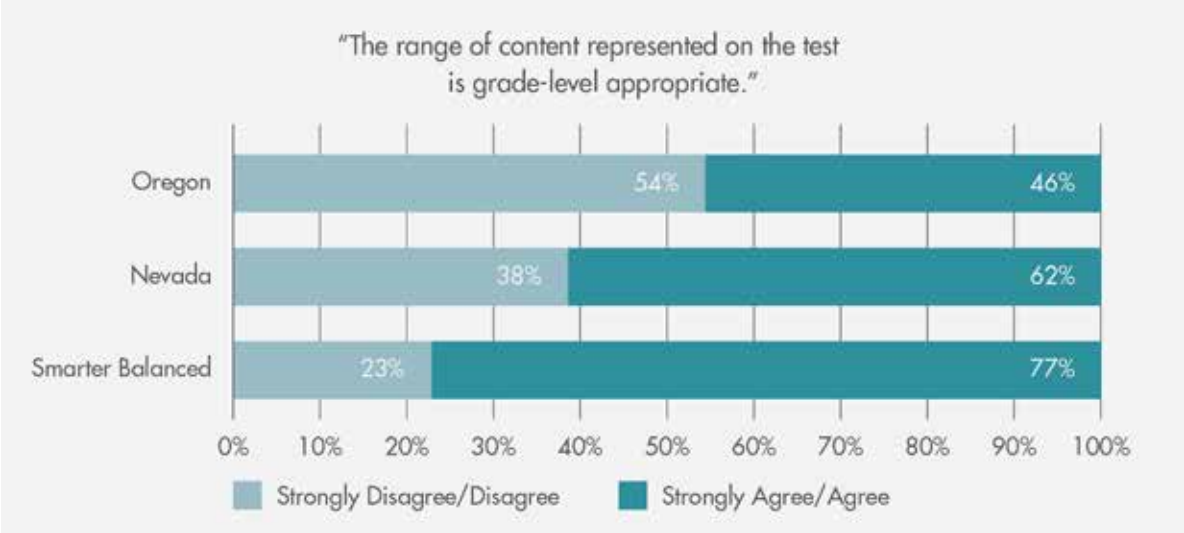


Figure 20. Percent agreement with statement: "The range of content represented on the test is grade-level appropriate" for 5th grade assessments. Detail may not add to total due to rounding.

Consistently throughout these items, the 5th grade Smarter Balanced assessments, in particular, have fared well on items evaluating grade-level appropriateness. While these new assessments clearly are seen as rigorous, they are not viewed as too challenging or unfair. They are seen as appropriate for a well-served 5th grade student and aligned with the expectation of an excellent teacher at this level.

Attitudes Toward Tests

Teachers were given an Attitudes Toward Tests survey to measure shifts in their perceptions of tests and test items over the course of the study. As shown in Table 1, the largest differences (.30 of a point or greater) or change in mean scores were for the statements:

- "I prefer tests that are comprised mostly of constructed-response or performance-based items." Teachers agreed less with this statement after evaluating the assessments.
- "Tests that are largely constructed-response/performance based are more appropriate for

the knowledge and skills embedded in my learning outcomes than selected-response tests." Teachers agreed less with this statement after evaluating the assessments.

- "Selected-response items can be used to measure complex thinking skills." Teachers agreed more with this statement after evaluating the assessments.

Table 1. Average Attitudes Toward Tests Results. Detail may not sum to total due to rounding.

Pre-Mean (1 to 4)	Attitudes toward Tests items	Post-Mean (1 to 4)	Pre-Post Difference
2.0	I prefer tests that are comprised mostly of selected-response items	2.0	0.0
1.8	Tests that are largely selected-response are more appropriate for the knowledge and skills embedded in my learning outcomes than constructed-response or performance-based tests.	1.9	0.1
2.8	I prefer tests that are comprised mostly of constructed-response or performance-based items.	2.6	-0.3
3.0	Tests that are largely constructed-response/performance based are more appropriate for the knowledge and skills embedded in my learning outcomes than selected-response tests.	2.7	-0.3
3.3	I prefer tests with some selected-response and some constructed-response items.	3.2	-0.1
3.2	Tests that are comprised of some selected-response items and some constructed-response items are more appropriate for the knowledge and skills embedded in my learning outcomes than multiple-choice tests.	3.3	0.1
2.9	Selected-response tests are simply easier to administer than constructed-response or performance-based tests.	3.0	0.1
2.5	Selected-response items can be used to measure complex thinking skills.	2.8	0.4

Concluding Thoughts

Utilizing the insight and expertise of excellent teachers in answering five key questions, we sought data and evidence to evaluate three claims:

1. The new consortium test remains better suited to supporting instruction than former tests.
2. The new consortium test reflects great teaching.
3. The new consortium test is of higher quality and worth the transition.

To ground this evaluation in the concrete actuality of where states had been and where they are now, we compared the 5th grade consortium assessment to two former state assessments. The results were clear and included insights from our best teachers about ways to adjust our course as we progress.

The findings from our study suggest that the 5th grade consortium assessment indeed is better for teaching and learning than the former assessments. It improves representation of the breadth and depth of content in excellent 5th grade classrooms over former assessments. If any standardized test is to truly support and influence teaching, it's important that the "right" kinds of questions are asked—the kinds of questions that appropriately reflect student knowledge and skills. The Smarter Balanced test does not assess outside the range of what 5th grade students are expected to know and do. In fact, it represents the shift toward better alignment between classroom instruction and standardized testing. The Smarter Balanced test also represents the kind of rigor that teachers think is reflective of their highest goals in teaching and learning. This is the direction they wish education to go in their classrooms, districts, states and jurisdictions, as well as in the nation as a whole.

We find that excellent teachers do want and prefer the new assessment. Teachers found the new assessment to be more closely aligned with their own practices in the classroom. Smarter Balanced had an appropriate mixture of items encompassing all levels of cognitive complexity while still being suitable to the grade level. If anything, the teachers in this study believed that the Smarter Balanced assessment could be extended to provide more information on high-performing students. Teachers support the adoption of content and learning standards that place high expectations on the learning of all students. They see value in assessments that are aligned to them, even in the face of the inevitable challenges of transition. One teacher put it this way:

> "…we want our students to be successful, and if we adopted these learning standards and the tests measures student performance with respect to those learning standards, why would [we] teach anything else?"

With careful implementation, strong support and training for teachers, transparency and effective communication, and patience from all stakeholder communities, the transition to consortia tests will be worthwhile.

References

Delaware Department of Education. (2012). Delaware Comprehensive Assessment System: State Summary Results of the Reading, Mathematics, Science, and Social Studies Assessment. Retrieved from http://www.doe.k12.de.us/ cms/ lib09/DE01922744/Centricity/domain/111/assessment/2012%2 0dcas%20summary%20reports/2012%20 DCAS%20 Summary%20Report.pdf

Guide to Using the 2013 NECAP Reports, (2014). Retrieved from https://reporting.measuredprogress.org/ necappublicri/ documents/1314/Fall/Guide%20to%20U sing%20the%202013%20NECAP%20Reports.pdf

Illinois State Board of Education (1997). Illinois State Learning Standards (archive). Retrieved from http://www.isbe.state. il.us/ils/archive/default.htm

Kentucky Department of Education. (2007). Support Materials for Core Content for Assessment, Version 4.1, Mathematics. Retrieved from: http://education.ky.gov/curriculum/docs/documents/cca_dok_support_808_mathematics.pdf

Kentucky Department of Education. (2007). Support Materials for Core Content for Assessment, Version 4.1, Reading. Retrieved from: https://www.aea267.k12.ia.us/system/assets/uploads/files/2472/reading_samples.pdf#page=6

Mislevy, R. J., Almond, R. G., Lukas, J. F. (2003). A brief introduction to evidence-centered design. Princeton, NJ. Educational Testing Service.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010).Common Core State Standards. Washington, DC: Authors.

New Jersey Department of Education. (2013). New Jersey Assessment of Skills and Knowledge 2012 Technical Report. Retrieved from http://www.nj.gov/education/assessment/es/njask_tech_report12.pdf

No Child Left Behind Act of 2001, P.L. 107-110, 20 U.S.C. § 6319 (2002).

Southern Nevada Regional Professional Development Program. (2009). Retrieved from http://rpdp.net/pdfs/ShopTalk%20 PDF/ShopTalk_Spr_09.pdf

Teacher of the Year [n.d.]. Council of Chief State School Officers. Retrieved from http://www.ccsso.org/ntoy/About_the_ Program.html

Webb, N. L. (2005). Alignment, depth of knowledge, and change. Presented at the Florida Research Association 50th Annual Meeting. Miami, Florida.

Webb, N. L. (2002). An analysis of the alignment between mathematics standards and assessments for three states. Paper presented at the American Educational Research Association Annual Meeting. New Orleans, LA.

Webb, N. (1997). Research Monograph Number 6: "Criteria for alignment of expectations and assessments on mathematics and science education." Washington, D.C.: CCSSO.

Appendix A: Assessment Details Grade 5 Tests

OAKS

The Oregon Assessment of Knowledge and Skills (OAKS) grade 5 assessment was used for the study. It is typically administered as a computer-adaptive test (CAT). However, we elected to not use the CAT version of the test for the purposes of the study. The form used was a linear form based on a student at the 60th percentile of the proficiency distribution at 5th grade. There were a total of 49 selected-response items the comprised 7 reading passages on the ELA assessment. There were 40 selected-response items on the Math assessment.

Nevada State Assessment

The Nevada State Assessment for grade 5 was used for the study. It is typically administered as a paper and pencil test. There were a total of 32 selected-response items the comprised 5 reading passages on the ELA assessment. There were a total of 63 selected- and short-response items on the Math assessment.

Smarter Balanced

The Smarter Balanced consortium assessment was designed to measure the standards set forth by the CCSS. It is typically administered as a computer-adaptive test (CAT). However, we elected to not use the CAT version of the test for the purposes of the study. The form used was a linear form based on a student at the 60th percentile of the proficiency distribution at grade 5. There were 44 selected-response and short- and extended-response items on the ELA assessment that comprised reading and listening passages. There were 40 selected-response, short-response, and non-traditional item types (e.g., hot spot where the student selects response by clicking the appropriate place on the graphic) on the Math assessment used for the study.

Appendix B: Survey Instruments Grade 5 Tests

Attitudes Toward Tests

The Attitudes Toward Tests survey was designed by the research team to capture teachers' perceptions about tests and item types. Educators can hold preferences for how best to measure student knowledge and skills. We thought it important to understand what these preferences were for participants prior to engaging with the assessments and especially to be aware if there were participants with extreme or outlier positions in the study.

For example, teachers might strongly prefer constructed-response items because of a belief that they are better suited for measuring most, if not all, complex knowledge and skills; this belief might be problematic if that teacher were reviewing an assessment comprised of solely forced-choice items. We also wanted to know if these preferences were subject to change after engaging with the assessments. Did their preference change after identifying selected-response items from one or more of the assessments that did a particularly good job of measuring highly complex knowledge or skills?

Teachers' attitudes toward tests were measured using an 8-item survey that was administered twice, once before and once after the panelists reviewed the assessments. The responses were given along a 4-point scale, where a response of '1' meant they strongly disagreed with a statement of preference and '4' meant they strongly agreed with a statement of preference. For example, "I prefer tests that are mostly comprised of constructed-response items." Key terms, such as "constructed-response" and "selected-response" were defined.

Instructions: For each of the following statements, please indicate your level of agreement.

Response scale: 1 (Strongly Disagree), 2 (Disagree), 3 (Agree), 4 (Strongly Agree).

	1	2	3	4
1. I prefer tests that are comprised mostly of selected-response items.	○	○	○	○
2. Tests that are largely selected-response are more appropriate for the knowledge and skills embedded in my learning outcomes than constructed-response or performance-based tests.	○	○	○	○
3. I prefer tests that are comprised mostly of constructed-response or performance-based items.	○	○	○	○
4. Tests that are largely constructed-response/performance based are more appropriate for the knowledge and skills embedded in my learning outcomes than selected-response tests.	○	○	○	○
5. I prefer tests with some selected-response and some constructed-response items.	○	○	○	○
6. Tests that are comprised of some selected-response items and some constructed-response items are more appropriate for the knowledge and skills embedded in my learning outcomes than multiple-choice tests.	○	○	○	○
7. Selected-response tests are simply easier to administer than constructed-response or performance-based tests.	○	○	○	○
8. Selected-response items can be used to measure complex thinking skills.	○	○	○	○

Survey of Assessment Quality

The *Survey of Assessment Quality* was developed to evaluate the five key areas of quality, as defined by the research team, for each test:

1. Does the new consortium assessment better reflect the range of knowledge and skills that all students should know?
2. Is the new consortium assessment designed to better reflect the full range of cognitive complexity in a balanced way?
3. Does the new consortium assessment better align with the strong instructional practices these teachers use in the classroom, and thereby better support great teaching and learning throughout the school year?
4. Does the new consortium assessment provide information relevant to a wide-range of performers?

5. While the new consortium assessment is more rigorous and demanding, is it grade-level appropriate, and more or less so than prior state tests?

The assessment quality survey consisted 58 total items, broken into two major components with different response scales. The first asked participants to evaluate whether, in their judgment as an expert teacher, the assessments had "enough" of the quantity being described in the survey item. The response scale was: "More than needed;" "Enough/About right;" and "Less than needed." The second asked participants to evaluate whether they "agreed" with statements describing the assessments in various ways in the survey item. The response scale was: "Strongly agree;" "Agree;" "Disagree" and "Strongly disagree."

This survey was administered once, after the reviews of all assessments were complete. Panelists responded to each survey item three times, once for each assessment they reviewed. While the participants completed their DOK review of the assessments in a randomly-assigned sequence to reduce any order effect, the survey responses were always in the same order to minimize confusion in responding.

SECTION I:

Instructions: Consider each statement and indicate the level at which there is "enough," 1 (less than needed), 2 (enough/about right) and 3 (more than needed) in the space provided for each test. You may also respond "N/A-I don't know" if you do not feel that you have enough information or are not qualified to judge. Note that for each item there is a Comments box where you may provide feedback on the item or why you gave your response; however, you are not obligated to put anything in the Comments box unless you feel the information is important for us to know.

Response Scale: 1 (less than needed), 2 (enough/about right) or 3 (more than needed)

	Test 1	Test 2	Test 3
1. Items that require recall, such as identification, labeling, calculating, defining, and reciting.			
2. Items that require application of skills, such as graphing, categorizing, organizing, predicting, and estimating.			
3. Items that require students to demonstrate strategic and extended thinking skills, such as investigation, analysis, and design.			
4. Cognitive demand for low-performing 5th grade students			
5. Cognitive demand for mid-performing 5th grade students			
6. Cognitive demand for high-performing 5th grade students			
7. Items that require 5th grade students to demonstrate basic knowledge of concepts.			
8. Items that surface information about 5th grade student performance at the lower ability levels that would be useful to inform my instructional strategies.			
9. Items that low-performing 5th grade students would be expected to get right.			
10. Items that low-performing 5th grade students would be expected to get wrong.			

	Test 1	Test 2	Test 3
11. Items that surface information about 5th grade student performance at the middle ability levels that would be useful to inform my instructional strategies.			
12. Items that mid-performing 5th grade students would be expected to get right.			
13. Items the mid-performing 5th grade students would be expected to get wrong.			
14. Items that surface information about 5th grade student performance at the high ability levels that would be useful to inform my instructional strategies.			
15. Items that high-performing 5th grade students would be expected to get right.			
16. Items that high-performing 5th grade students would be expected to get wrong.			
17. Number of items that require application of skills needed to distinguish mid-performing from low-performing 5th grade students.			
18. Number of items that require complex thinking skills needed to distinguish high-performing from mid-performing 5th grade students.			
19. The number of items that are above 5th grade level.			
20. The number of items that are below 5th grade level.			
21. Items that are likely to authentically engage student interest.			

SECTION II:

Instructions: Consider each statement and indicate your level of agreement, 1 (strongly disagree) to 4 (strongly agree) in the space provided for each test. You may also respond "N/A-I don't know" if you do not feel that you have enough information or are not qualified to judge. Note that for each item there is a Comments box where you may provide feedback on the item or why you gave your response; however, you are not obligated to put anything in the Comments box unless you feel the information is important for us to know.

Response Scale: 1 (strongly disagree), 2 (disagree), 3 (agree), or 4 (strongly agree)

	Test 1	Test 2	Test 3
1. Students are required to integrate a variety of knowledge and skills from a single domain.			
2. Students are required to transfer knowledge from different domains.			
3. Students are required to integrate a variety of knowledge and skills from different domains.			
4. This test provides sufficient opportunity to evaluate students' ability to communicate in writing.			

	Test 1	Test 2	Test 3
5. This test provides sufficient opportunity to evaluate students' ability to show their reasoning when solving a problem or arguing a case.			
6. This test strikes a balance between the number of items that require recall responses and responses that require higher-level cognitive skills.			
7. Students are required to demonstrate complex thinking skills, such as experimentation, analysis, and synthesis.			
8. This test is more cognitively demanding than is warranted for the 5th grade level.			
9. This test is less cognitively demanding than is warranted for the 5th grade level.			
10. Items on this test are consistent with what excellent 5th grade Math/ELA teachers ask their students to know and do.			
11. Preparing students for this test would require meaningful lessons and learning, beyond skill and drill practice.			
12. One criterion for a high-quality assessment is that the assessment allows students to transfer their learning to new situations and problems[2]. This test meets that criterion.			
13. This test measures an appropriately broad sampling of the ELA/ Math knowledge and skills in instruction an excellent 5th grade classroom.			
14. Excellent 5th grade instruction generally aligns with the content measured on this test.			
15. This test measures the most important knowledge and skills to be taught in an excellent 5th grade Math/ELA classroom.			
16. This test measures the learning outcomes that I would set for student learning in 5th grade classes.			
17. Certain item types are emphasized more heavily on the test than is warranted for the grade level.			
18. Certain content areas are emphasized more heavily on the test than is warranted for the grade level.			
19. I would give more emphasis to certain content areas in 5th grade classes than the test does.			
20. The distribution of content on the test is representative of excellent instruction at the 5th-grade level.			
21. The depth of content represented on the test is grade-level appropriate.			
22. The range of content represented on the test is grade-level appropriate.			
23. One criterion for a high-quality assessment is that the assessment is designed to measure whether underlying concepts have been taught and learned, rather than reflecting mostly test-taking skills or reflecting out-of-school experiences. This test meets that criterion.			

2 Darling-Hammond, L., Herman, J., Pellegrino, J., et al. (2013). Criteria for high-quality assessment. Stanford, CA: Stanford Center Opportunity Policy in Education

	Test 1	Test 2	Test 3
24. If I backwards-mapped a 5th grade lesson against items like those on this test, it would help inform my lesson plan and guide me toward high quality instruction.			
25. I would like to use formative assessments built using items from this test in a 5th grade classroom.			
26. The optimal formative assessments that I would give to 5th grade students measure concepts not addressed by this test.			
27. If used for formative assessment, items on this test would help me make decisions about instruction.			
28. Student results from this test would give me valuable information about how students are learning.			
29. The item types on this test are aligned with the skills they appear to be designed to measure.			
30. This test provides a satisfactory balance between selected-response items and constructed response/performance-based items.			
31. Low-performing students would find it easy to get most of the items on this test correct.			
32. Mid-performing students would find it easy to get most of the items on this test correct.			
33. High-performing students would find it easy to get most of the items on this test correct.			
34. Low-performing students would generally perform well on this test.			
35. Mid-performing students would generally perform well on this test.			
36. High-performing students would generally perform well on this test.			
37. Students would likely be authentically engaged in items from this test.			

The percentage of survey items that cover each area is summarized in Table B1 by section. There were two sections of the survey. In Section 1, teachers were asked to indicate the level of "enough" of a particular characteristic each test possessed. For example, for the statement, "Cognitive demand for low-performing students," teachers were asked to indicate if the amount was "less than needed" (1), "enough/about right" (2), or "more than needed" (3). A substantial percentage of this section addressed the appropriateness or rigor of the items for low-, mid- and high-performing students (40%). In Section 2, participants were asked to indicate their level of agreement ("strongly disagree," "disagree," "agree" or "strongly agree") with statements about the content, performance levels, balance, and grade appropriateness of the items in each of the assessments, overall. A larger percentage of this section addressed the representativeness of the knowledge and skills by test items (36%). There were two additional questions, one in each sec-

tion, concerning the likelihood of student interest or engagement each test would inspire (e.g., "Students would likely be authentically engaged in items from this test").

Table B2. Percent Coverage of Key Areas by Section

Key Area	Description	Percent Coverage	
		Section 1	Section 2
KSAs	Represents the full range of knowledge and skills taught in your classes appropriate for this type of assessment.	20%	36%
Cognitive	Assesses deep levels of cognitive ability in a balanced way.	15%	22%
Performance	Is appropriate for a wide range of performance levels.	40%	17%
Teaching	Promotes your most successful classroom teaching practices.	15%	19%
Grade	Grade Appropriate.	10%	6%

Appendix C: Panel Demographics Grade 5 Tests

In this appendix, the details of the panel demographics are provided.

Figure C1. Gender

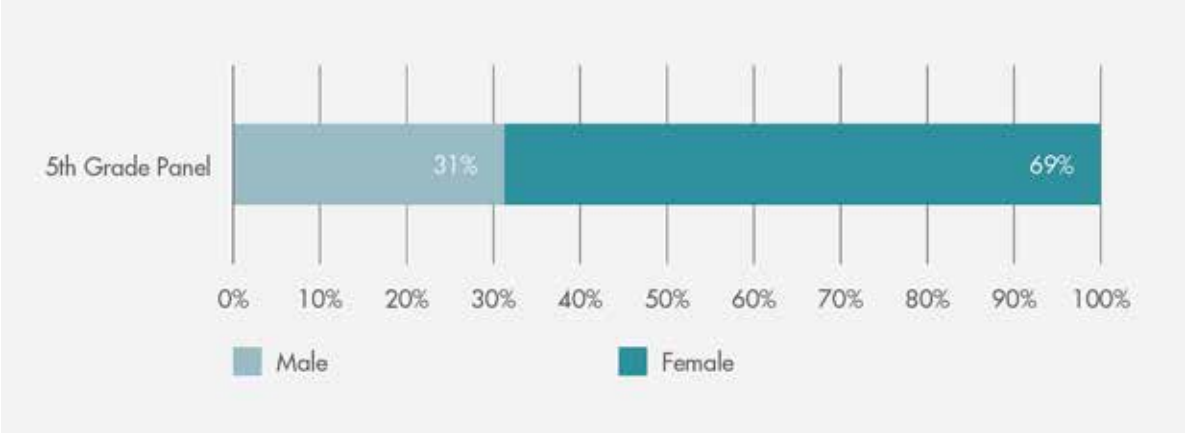


Figure C2. Race/Ethnicity

Detail may not add to total due to rounding.

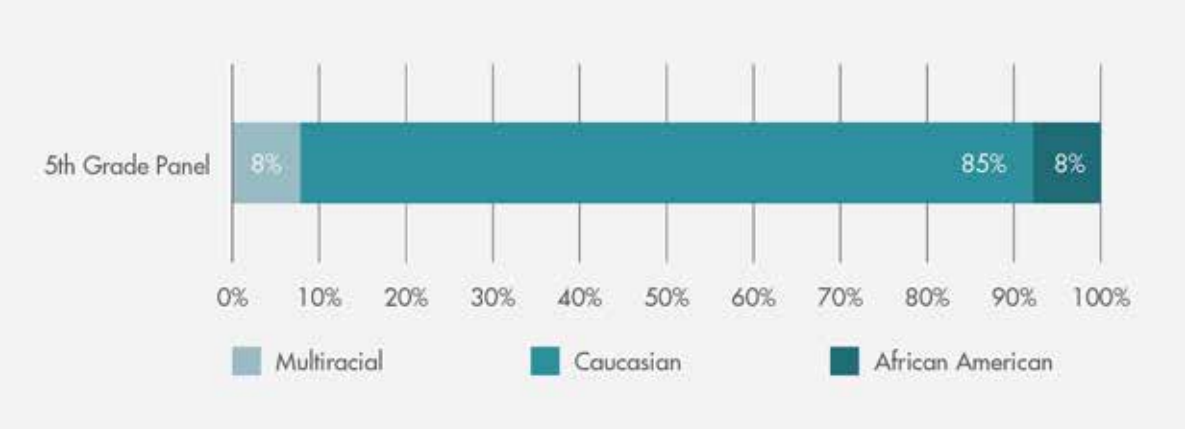


Figure C3. Years of Teaching Experience

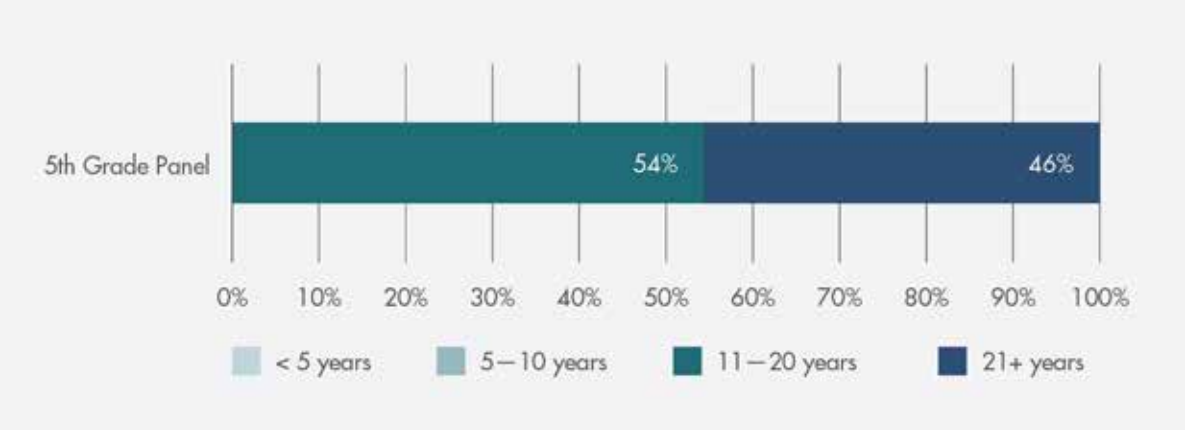
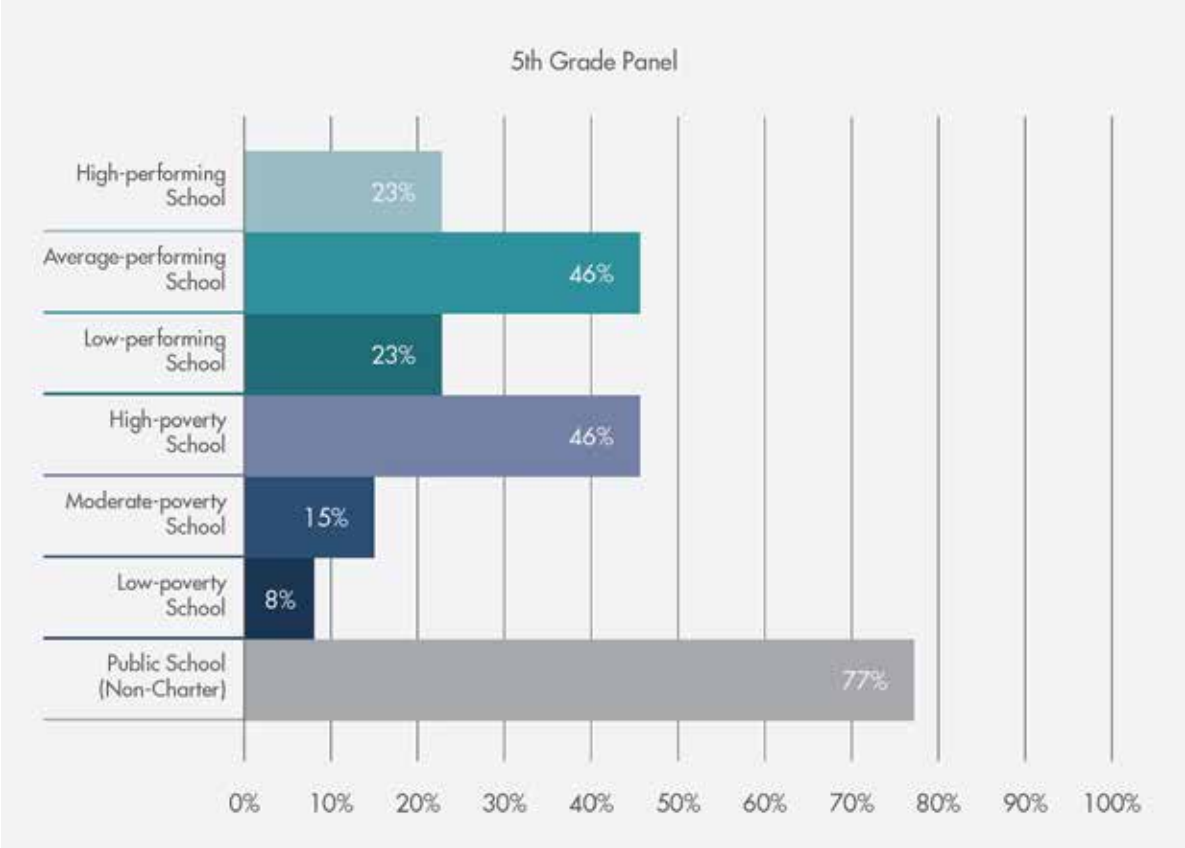


Figure C4. Teaching Contexts



Appendix D: Guiding Questions for Panel Discussions Grade 5 Tests

A set of standard questions was developed based on the survey data, and follow-up prompts were incorporated organically throughout the discussion. The standard questions asked of each panel are listed below.

1. Were there any aspects of the study that may have prejudiced your judgments in favor of one test or another before you started today's survey?
2. Were there any aspects of the study that may have prejudiced your judgments in favor of one test or another while you were completing the survey?

The next set of questions were motivated by the panel's survey data:

3. A majority of you of responded that the former state assessments, OR and NV, did not contain enough items that authentically engage student interest, and Smarter Balanced assessment contained enough of those items. In what ways does Smarter Balanced engage students more authentically than OR and NV?
4. What aspects of this study will you be taking away with you today?
5. What are you going to do with the information that was shared with you during this study?

Appendix E: Survey of Assessment Quality Items Grade 5 Tests

Participants were asked to evaluate whether, in their judgment as an expert teacher, the assessments had "enough" of the quantity being described in the survey item below. The response scale was: "More than needed;" "Enough/About right" and "Less than needed." The results are presented below in Table E1 for the 5th grade panel, in two formats. The percentage of teachers who responded in each category for each assessment is shown. The percentages are shaded so that values of 50% or greater are blue.

In addition, the categories were coded as follows:

- More than needed = 3
- Enough/About right = 2
- Less than needed = 1

These values were averaged and the mean score is shown in Table E1 for each assessment as well.

Table E1. "Amount" Items: Nevada, OAKS, 5th Grade Smarter Balanced

Amount Items	Nevada				OAKS				Smarter Balanced			
	Less Than	Enough	More Than	Mean Score	Less Than	Enough	More Than	Mean Score	Less Than	Enough	More Than	Mean Score
Items that require recall, such as identification, labeling, calculating, defining, and reciting.	0%	23%	77%	2.8	0%	15%	85%	2.8	8%	85%	8%	2.0
Items that require application of skills, such as graphing, categorizing, organizing, predicting, and estimating.	15%	62%	23%	2.1	31%	69%	0%	1.7	15%	77%	8%	1.9
Items that require students to demonstrate strategic and extended thinking skills, such as investigation, analysis, and design.	85%	15%	0%	1.2	100%	0%	0%	1.0	8%	85%	8%	2.0
Cognitive demand for low-performing 5th grade students	15%	54%	31%	2.2	15%	54%	31%	2.2	8%	54%	38%	2.3
Cognitive demand for mid-performing 5th grade students	38%	54%	8%	1.7	54%	38%	8%	1.5	0%	77%	23%	2.2
Cognitive demand for high-performing 5th grade students	100%	0%	0%	1.0	100%	0%	0%	1.0	31%	69%	0%	1.7
Items that require 5th grade students to demonstrate basic knowledge of concepts.	8%	8%	85%	2.8	8%	23%	69%	2.6	15%	77%	8%	1.9
Items that surface information about 5th grade student performance at the lower ability levels to inform my instructional strategies.	17%	33%	50%	2.3	15%	54%	31%	2.2	23%	69%	8%	1.8

Table E1. "Amount" Items: Nevada, OAKS, 5th Grade Smarter Balanced (continued)

Amount Items	Nevada				OAKS				Smarter Balanced			
	Less Than	Enough	More Than	Mean Score	Less Than	Enough	More Than	Mean Score	Less Than	Enough	More Than	Mean Score
Items that low-performing 5th grade students would be expected to get right.	15%	46%	38%	2.2	23%	54%	23%	2.0	54%	38%	8%	1.5
Items that low-performing 5th grade students would be expected to get wrong.	31%	38%	31%	2.0	31%	46%	23%	1.9	15%	38%	46%	2.3
Items that surface information about 5th grade student performance at the middle ability levels to inform my instructional strategies.	62%	38%	0%	1.4	38%	62%	0%	1.6	15%	85%	0%	1.8
Items that mid-performing 5th grade students would be expected to get right.	31%	38%	31%	2.0	31%	54%	15%	1.8	15%	85%	0%	1.8
Items the mid-performing 5th grade students would be expected to get wrong.	38%	54%	8%	1.7	38%	54%	8%	1.7	15%	62%	23%	2.1
Items that surface information about 5th grade student performance at the high ability levels to inform my instructional strategies.	92%	0%	8%	1.2	85%	8%	8%	1.2	31%	62%	8%	1.8
Items that high-performing 5th grade students would be expected to get right.	54%	8%	38%	1.8	62%	0%	38%	1.8	17%	67%	17%	2.0
Items that high-performing 5th grade students would be expected to get wrong.	85%	8%	8%	1.2	92%	0%	8%	1.2	38%	62%	0%	1.6
Number of items that require application of skills needed to distinguish mid-performing from low-performing 5th grade students.	23%	46%	31%	2.1	15%	62%	23%	2.1	15%	77%	8%	1.9
Number of items that require complex thinking skills needed to distinguish high-performing from mid-performing 5th grade students.	77%	23%	0%	1.2	100%	0%	0%	1.0	8%	77%	15%	2.1
The number of items that are above 5th grade-level.	69%	31%	0%	1.3	77%	23%	0%	1.2	23%	62%	15%	1.9
The number of items that are below 5th grade-level.	15%	46%	38%	2.2	8%	46%	46%	2.4	46%	46%	8%	1.6
Items that are likely to authentically engage student interest.	83%	17%	0%	1.2	85%	15%	0%	1.2	31%	69%	0%	1.7

Participants were asked to evaluate whether they "agreed" with statements describing the assessments in various ways in the survey item. The response scale was: "Strongly agree;" "Agree;" "Disagree" and "Strongly disagree." The results are presented below in Table E2 for the 5th grade panel in the same two formats as above and with the same shading protocol. The categories were coded as follows:

- Strongly agree = 4
- Agree = 3
- Disagree = 2
- Strongly disagree = 1

These values were averaged and the mean score is shown in Table E2 for each assessment as well.

Table E2. "Agree" Items: Nevada, OAKS, 5th Grade Smarter Balanced

Agree Items	Nevada					OAKS					Smarter Balanced				
	SD	D	A	SA	Mean Score (1 to 4)	SD	D	A	SA	Mean Score (1 to 4)	SD	D	A	SA	Mean Score (1 to 4)
Students are required to integrate a variety of knowledge and skills from a single domain.	8%	62%	23%	8%	2.3	25%	33%	33%	8%	2.3	0%	23%	46%	31%	3.1
Students are required to transfer knowledge from different domains.	15%	46%	38%	0%	2.2	23%	69%	8%	0%	1.8	0%	8%	77%	15%	3.1
Students are required to integrate a variety of knowledge and skills from different domains.	17%	58%	25%	0%	2.1	25%	75%	0%	0%	1.8	0%	8%	75%	17%	3.1
This test provides sufficient opportunity to evaluate students' ability to communicate in writing.	31%	38%	23%	8%	2.1	62%	38%	0%	0%	1.4	0%	62%	38%	0%	2.4
This test provides sufficient opportunity to evaluate students' ability to show their reasoning when solving a problem or arguing a case.	31%	38%	31%	0%	2.0	38%	62%	0%	0%	1.6	0%	15%	69%	15%	3.0
This test strikes a balance between the number of items that require recall responses and responses that require higher-level cognitive skills.	31%	54%	15%	0%	1.8	62%	31%	8%	0%	1.5	0%	15%	69%	15%	3.0
Students are required to demonstrate complex thinking skills, such as experimentation, analysis, and synthesis.	31%	62%	8%	0%	1.8	46%	54%	0%	0%	1.5	0%	0%	77%	23%	3.2
This test is more cognitively demanding than is warranted for the 5th grade level.	38%	62%	0%	0%	1.6	46%	54%	0%	0%	1.5	0%	62%	38%	0%	2.4
This test is less cognitively demanding than is warranted for the 5th grade level.	0%	31%	62%	8%	2.8	8%	15%	54%	23%	2.9	15%	69%	15%	0%	2.0

Table E2. "Agree" Items: Nevada, OAKS, 5th Grade Smarter Balanced (continued)

Agree Items	Nevada					OAKS					Smarter Balanced				
	SD	D	A	SA	Mean Score (1 to 4)	SD	D	A	SA	Mean Score (1 to 4)	SD	D	A	SA	Mean Score (1 to 4)
Items on this test are consistent with what excellent 5th grade Math/ELA teachers ask their students to know and do.	46%	31%	23%	0%	1.8	54%	31%	15%	0%	1.6	0%	23%	69%	8%	2.8
Preparing students for this test would require meaningful lessons and learning, beyond skill and drill practice.	23%	38%	38%	0%	2.2	23%	46%	31%	0%	2.1	0%	0%	77%	23%	3.2
One criterion for a high-quality assessment is that the assessment allows students to transfer their learning to new situations and problems. This test meets that criterion.	31%	46%	23%	0%	1.9	54%	46%	0%	0%	1.5	0%	23%	62%	15%	2.9
This test measures an appropriately broad sampling of the ELA/Math knowledge and skills in instruction an excellent 5th grade classroom.	23%	46%	31%	0%	2.1	38%	38%	23%	0%	1.8	0%	33%	58%	8%	2.8
Excellent 5th grade instruction generally aligns with the content measured on this test.	15%	46%	38%	0%	2.2	8%	69%	23%	0%	2.2	0%	8%	85%	8%	3.0
This test measures the most important knowledge and skills to be taught in an excellent 5th grade Math/ELA classroom.	23%	23%	54%	0%	2.3	23%	54%	23%	0%	2.0	0%	23%	69%	8%	2.8
This test measures the learning outcomes that I would set for student learning in 5th grade classes.	23%	31%	46%	0%	2.2	31%	54%	15%	0%	1.8	0%	8%	85%	8%	3.0
Certain item types are emphasized more heavily on the test than is warranted for the grade level.	0%	15%	46%	38%	3.2	0%	15%	46%	38%	3.2	0%	69%	31%	0%	2.3
Certain content areas are emphasized more heavily on the test than is warranted for the grade level.	0%	38%	38%	23%	2.8	0%	31%	46%	23%	2.9	0%	85%	15%	0%	2.2
I would give more emphasis to certain content areas in 5th grade classes than the test does.	0%	38%	38%	23%	2.8	0%	8%	62%	31%	3.2	0%	69%	23%	8%	2.4
The distribution of content on the test is representative of excellent instruction at the 5th grade level.	15%	46%	38%	0%	2.2	23%	62%	15%	0%	1.9	0%	8%	85%	8%	3.0
The depth of content represented on the test is grade-level appropriate.	31%	46%	23%	0%	1.9	38%	54%	8%	0%	1.7	0%	0%	92%	8%	3.1
The range of content represented on the test is grade-level appropriate.	23%	15%	62%	0%	2.4	38%	15%	46%	0%	2.1	0%	23%	69%	8%	2.8

Table E2. "Agree" Items: Nevada, OAKS, 5th Grade Smarter Balanced (continued)

Agree Items	Nevada					OAKS					Smarter Balanced				
	SD	D	A	SA	Mean Score (1 to 4)	SD	D	A	SA	Mean Score (1 to 4)	SD	D	A	SA	Mean Score (1 to 4)
One criterion for a high-quality assessment is that the assessment is designed to measure whether underlying concepts have been taught and learned, rather than reflecting mostly test-taking skills or reflecting out-of-school experiences. This test meets that criterion.	15%	54%	31%	0%	2.2	38%	38%	23%	0%	1.8	0%	8%	85%	8%	3.0
If I backwards-mapped a 5th grade lesson against items like those on this test, it would help inform my lesson plan and guide me toward high quality instruction.	31%	38%	31%	0%	2.0	38%	54%	8%	0%	1.7	0%	0%	77%	23%	3.2
I would like to use formative assessments built using items from this test in a 5th grade classroom.	15%	23%	62%	0%	2.5	23%	46%	23%	8%	2.2	0%	0%	62%	38%	3.4
The optimal formative assessments that I would give to 5th grade students measure concepts not addressed by this test.	0%	31%	62%	8%	2.8	0%	0%	92%	8%	3.1	0%	92%	8%	0%	2.1
If used for formative assessment, items on this test would help me make decisions about instruction.	8%	23%	62%	8%	2.7	15%	31%	46%	8%	2.5	0%	0%	85%	15%	3.2
Student results from this test would give me valuable information about how students are learning.	25%	25%	50%	0%	2.3	42%	25%	33%	0%	1.9	0%	8%	83%	8%	3.0
The item types on this test are aligned with the skills they appear to be designed to measure.	0%	46%	54%	0%	2.5	8%	62%	31%	0%	2.2	0%	8%	92%	0%	2.9
This test provides a satisfactory balance between selected-response items and constructed response/performance-based items.	31%	46%	23%	0%	1.9	69%	31%	0%	0%	1.3	0%	31%	54%	15%	2.8
Low-performing students would find it easy to get most of the items on this test correct.	15%	69%	15%	0%	2.0	15%	31%	54%	0%	2.4	46%	54%	0%	0%	1.5
Mid-performing students would find it easy to get most of the items on this test correct.	0%	8%	85%	8%	3.0	0%	8%	77%	15%	3.1	8%	62%	31%	0%	2.2
High-performing students would find it easy to get most of the items on this test correct.	0%	0%	46%	54%	3.5	0%	0%	46%	54%	3.5	8%	15%	69%	8%	2.8

Table E2. "Agree" Items: Nevada, OAKS, 5th Grade Smarter Balanced (continued)

Agree Items	Nevada					OAKS					Smarter Balanced				
	SD	D	A	SA	Mean Score (1 to 4)	SD	D	A	SA	Mean Score (1 to 4)	SD	D	A	SA	Mean Score (1 to 4)
Low-performing students would generally perform well on this test.	23%	54%	23%	0%	2.0	15%	38%	46%	0%	2.3	38%	62%	0%	0%	1.6
Mid-performing students would generally perform well on this test.	0%	0%	92%	8%	3.1	0%	0%	92%	8%	3.1	0%	46%	54%	0%	2.5
High-performing students would generally perform well on this test.	0%	0%	54%	46%	3.5	0%	0%	46%	54%	3.5	0%	0%	69%	31%	3.3
Students would likely be authentically engaged in items from this test.	8%	77%	15%	0%	2.1	15%	77%	8%	0%	1.9	0%	38%	62%	0%	2.6


NNSTOY
TEACHERSLEADING