NNSTOY
TEACHERSLEADING
POLICY • PRACTICE • ADVOCACY

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.

November 2015


NNSTOY
TEACHERSLEADING
POLICY • PRACTICE • ADVOCACY

We at the National Network of State Teachers of the Year (NNSTOY) are delighted to share with you the latest in our series of Research Reports.

In this report, we 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 two panels to examine six assessment instruments. Each study panel was composed of State and National Teachers of the Year and Finalists for State Teacher of the Year. Each panel examined three assessments: two assessments given by states before switching to new state assessments developed by the PARCC and Smarter Balanced assessment consortia and one consortia assessment.

Working with our study partners, EducationCounsel on the policy side and Clowder Consulting on the science end, we are eager to share our findings. In short, participating teachers viewed state movement to the new consortia assessments as a positive step forward, specifically:

1. The new consortia assessments better reflect the range of reading and math knowledge and skills that all students should master.
2. The new consortia assessments include items that better reflect the full range of cognitive complexity in a balanced way.
3. The new consortia assessments better align with the kinds of strong instructional practices these expert teachers believe should be used in the classroom, and thereby better support great teaching and learning throughout the school year.
4. The new consortia assessments provide information relevant to a wide range of performers, particularly moderate and high-performers.
5. While the new consortia assessments are more rigorous and demanding, they are grade-level appropriate, and even more so than prior state tests.

Though they noted areas for continuous improvement, these same teachers also felt that educators and policymakers should focus together on the work ahead – 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, Chief Executive Officer, NNSTOY

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, Ms. Terri Taylor, and Mr. Nick Spiva for their commitment to learning what a group of outstanding educators thought 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.

Our Reviewers

We wish to thank Mr. Joshua Starr, Mr. Michael Petrilli, and Mr. Joshua Parker for their thoughtful and painstaking reviews of this study report. Their contributions and questions were invaluable as we prepared to release these findings.

Assessment Providers

Allowing an outside agency access to confidential assessment material is a serious undertaking. We are most grateful to the States of Delaware, Illinois, New Hampshire, and New Jersey for allowing us access to their prior state student assessments. We are equally grateful to the two consortia, PARCC and 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. Without this funding, this study would not have taken place. We are most grateful.

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 and 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

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

Executive Summary

"The Right Trajectory" brings to the forefront an often-overlooked voice in the debate about new state assessments developed in consortia: that of the best teachers in the country. This research suggests, despite challenges still to overcome, that these front-line experts believe that the new consortia tests are an improvement on the former assessments and so represent movement in the right direction for students and for education in their states.

What do great teachers think of the new assessments compared to 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).

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 former tests side by side – would think about these new consortia assessments, and informing the field accordingly. Simply put: Do the new assessments better reflect what great teachers are doing in their classrooms? Do they reflect higher quality than former state tests? Do these assessments 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 the local and state levels for their teaching excellence. One panel reviewed PARCC and two prior state assessments: ISAT from Illinois, and NJASK from New Jersey (both states currently use PARCC). The second panel reviewed Smarter Balanced and two prior state assessments: DCAS from Delaware, and NECAP from New Hampshire (both states currently use Smarter Balanced). All assessments were for fifth grade reading and math because it is on the cusp between elementary and middle school, making assessments at that grade relevant to elementary and middle teachers and students.

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 directions are to be successful. Several significant results from the study are highlighted below.

What we found is clear: There was consensus across participating teachers that the new consortia assessments – both PARCC and Smarter Balanced – represent an improvement and the right trajectory. They illustrate where we should be headed in summative assessment over the prior state assessments that were examined. In particular, as elaborated in the full report, evidence gathered from participating state teachers of the year support the following related findings[1]:

[1] These findings combine responses from two different participant groups. Each group examined a non-overlapping set of state assessments and one consortium assessment. All participants were asked the same set of survey questions from which these response data were taken.

1. **The new consortia assessments better reflect the range of reading and math knowledge and skills that all students should master.** Teachers in our study spent time meticulously examining the consortia tests 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 consortia assessments better reflected and measured those expectations, 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 fifth grade classroom," 70% of participating teachers agreed or strongly agreed when referring to the consortia tests. Only 33% agreed when referring to the former state tests. There is considerable variation in opinion on specific state tests, with the New Jersey assessment (NJASK) scoring as well as the consortia tests, but as a group the former state tests were not highly rated.

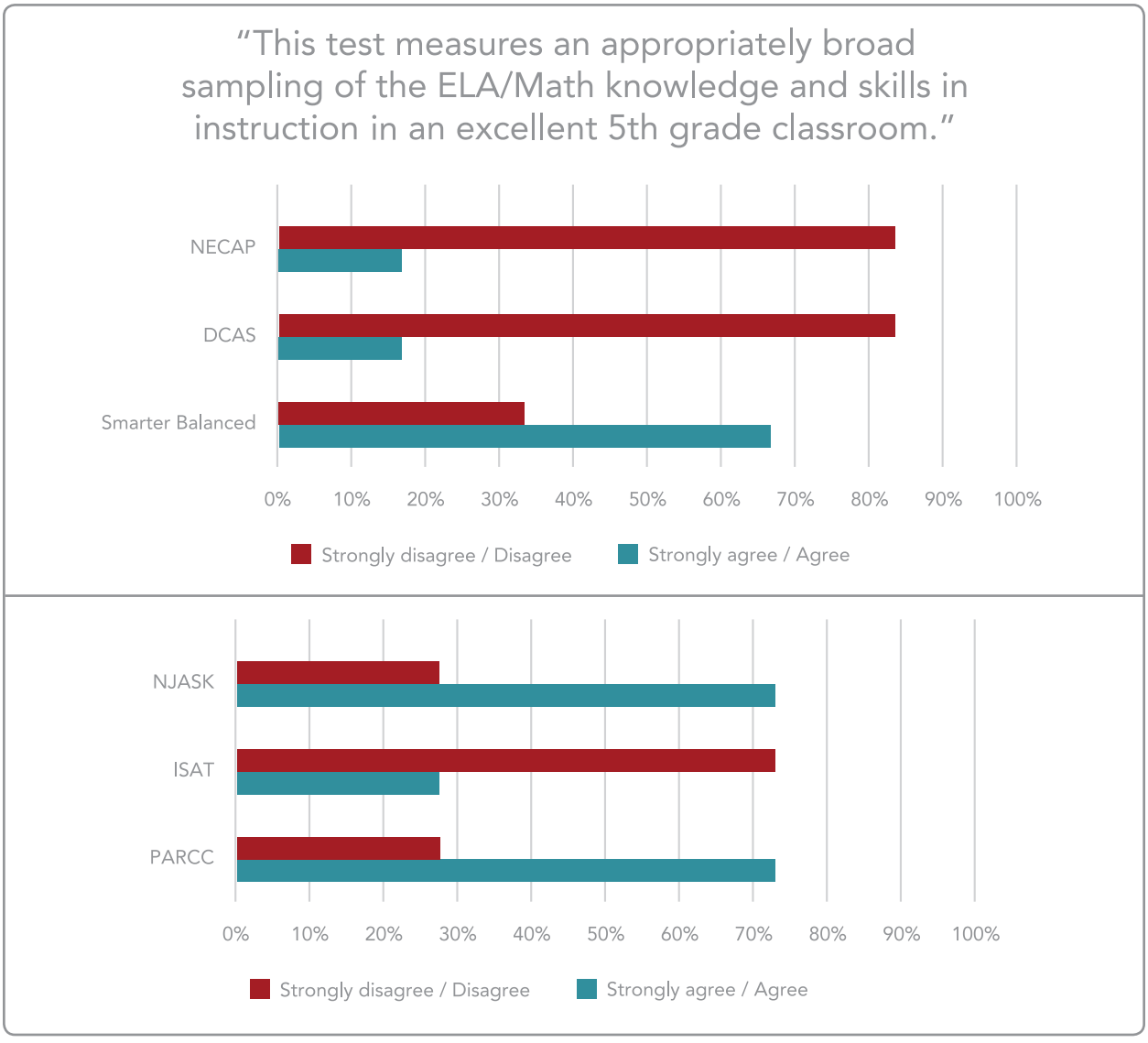


Figure 1: Percent agreement with the statement: "This test measures an appropriately broad sampling of the ELA/Math knowledge and skills in instruction in an excellent fifth grade classroom."

2. **The new consortia assessments are designed to include items that better reflect the full range of cognitive complexity in a balanced way.** Teachers found that items on the new consortia tests 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. When asked whether they agreed that "The distribution of content on the test is representative of excellent instruction at the fifth grade level," 74% endorsed it for the consortia tests, but only 37% did so for the former state tests. Again, the NJASK was rated somewhat higher than the other state assessments.

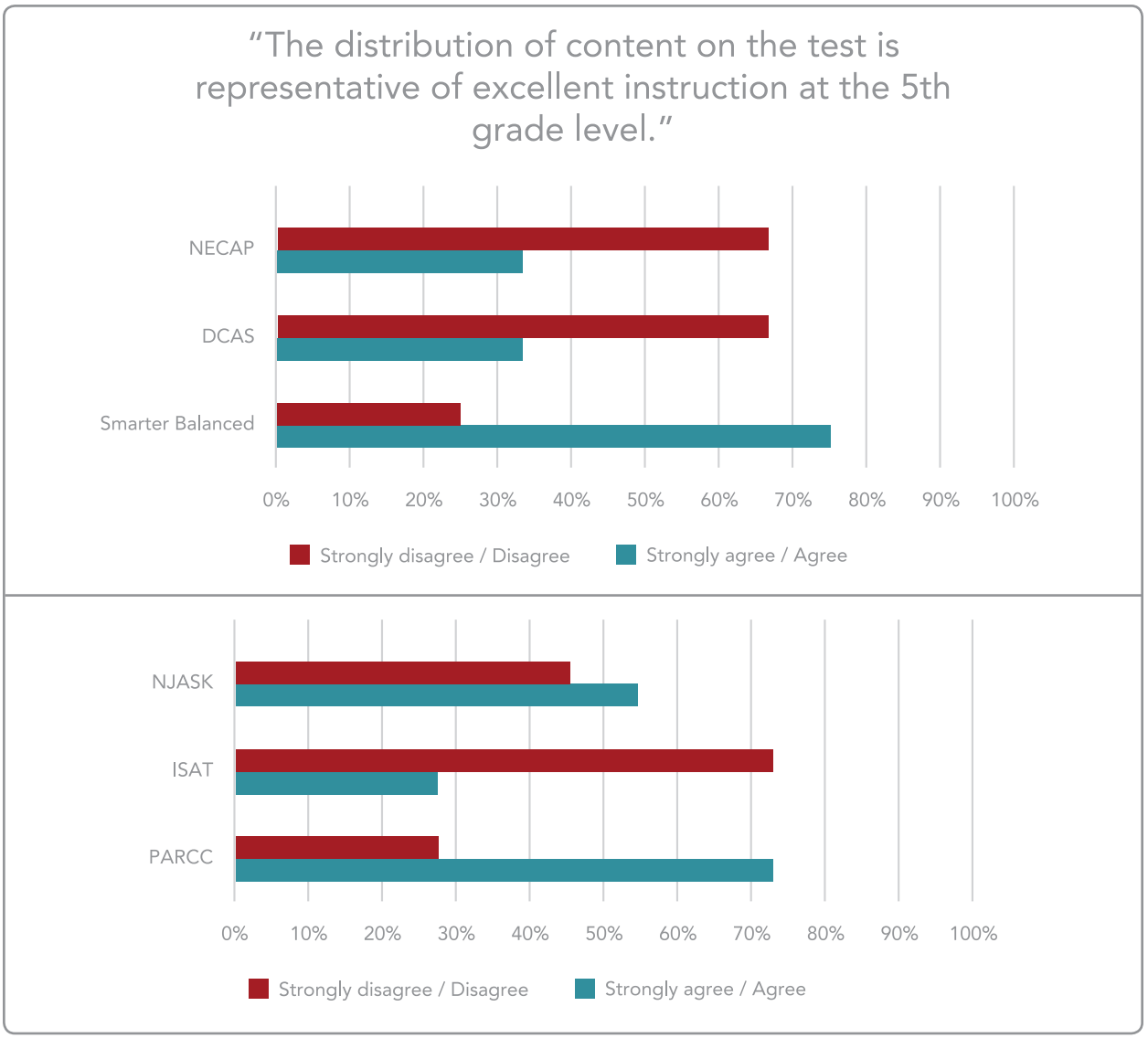


Figure 2: Percent agreement with the statement: "The distribution of content on the test is representative of excellent instruction at the fifth grade level."

3. **The new consortia assessments better align with the kinds of strong instructional practices these expert teachers believe should be used in the classroom, and thereby better support great teaching and learning throughout the school year.** The consortia assessments were 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 consortia tests, and questions were asked in ways that were better aligned to the instructional practices of excellent classrooms than the previous assessments.

These teachers found the new assessments more representative of meaningful instruction, both in content and delivery, in well-taught classrooms. For the consortia assessments, 88% agreed or strongly agreed "preparing students for this test would require meaningful lessons and learning, beyond skill and drill practice," but only 44% agreed or strongly agreed with the statements for the prior state tests. There was variation, however, among the prior state assessments on this item; the NJASK received very high support, at 82% agreeing or strongly agreeing.

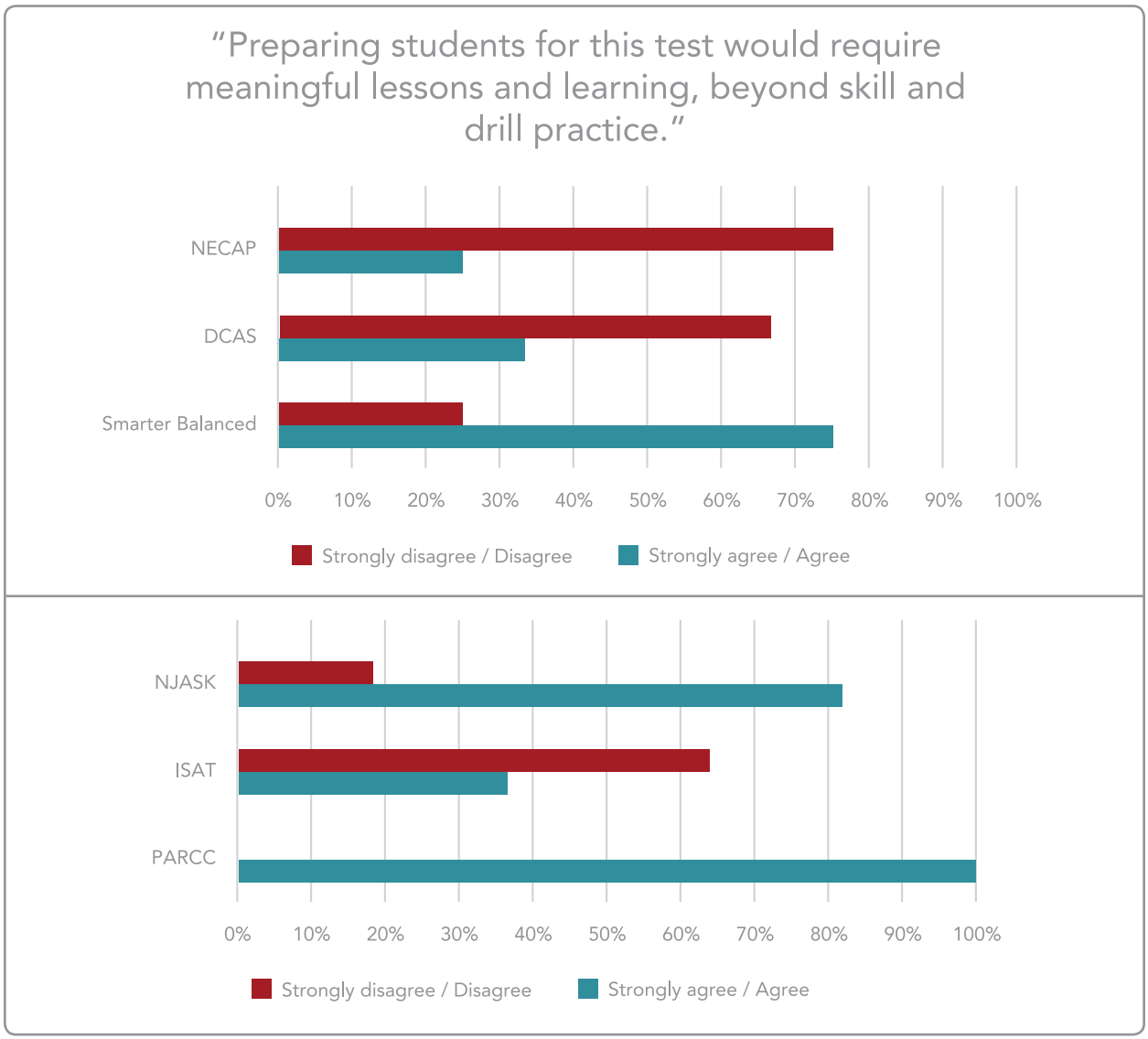


Figure 3: Percent agreement with the statement: "Preparing students for this test would require meaningful lessons and learning, beyond skill and drill practice."

4. **The new consortia assessments provide information relevant to a wide range of performers, particularly moderate and high-performers.** A clear trend that emerged through the project was that the new consortia tests gave moderate and high-performing students opportunities to demonstrate the range and depth of their knowledge and skills. For example, teachers generally thought that the former assessments had fewer items that required complex thinking skills than was needed to distinguish mid-performing and high-performing students, but the new consortia assessments possessed about the right amount or enough of those items.

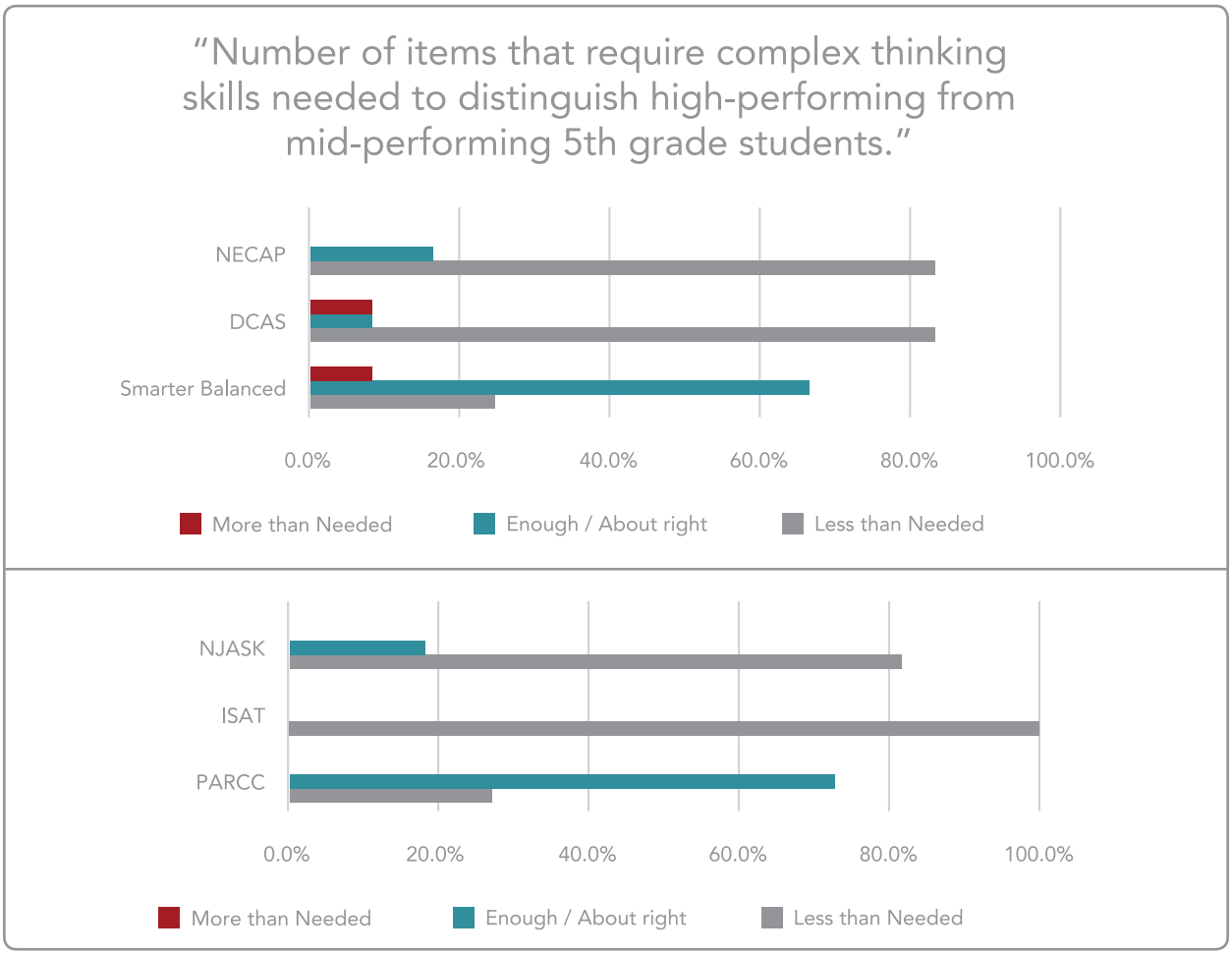


Figure 4. Average percent of teachers who indicated the: "Number of items that require complex thinking skills needed to distinguish high-performing from mid-performing fifth grade students" was more than needed, enough/about right, or less than needed.

One teacher commented, for example:

> *"For my really high-performing students, I actually really do want those high levels of questions there. Because if you're going to give me a [test targeted at] low-level ability [students] for fifth grade, that doesn't give me an overall perception of a summative assessment or where they're at in the ultimate journey. I don't want to give a [former] assessment that tells me that my students have basic skills that aren't going to get them anywhere. I want to give something like an [Consortium Test] that's going to tell me where—if I think they're highly performing, are they performing highly? I actually want the bar to be a high standard. And then it's my job to give formative assessments and other pieces to determine what I need to change for instruction to get them to that higher level.*

When asked if the number of items that would allow them to distinguish between low-level student performance and mid-level student performance was less than needed, about right, or more than needed, teachers' responses differed quite a bit between panels. One panel indicated the number of items that require application of skill to distinguish between low-performing and mid-performing fifth grade students was about right for all three assessments (NJASK, ISAT, and PARCC), as shown in Figure 18. The other panel indicated that the number of items was less than what was needed for the former assessments and about right or enough only for the consortium assessment (Smarter Balanced).

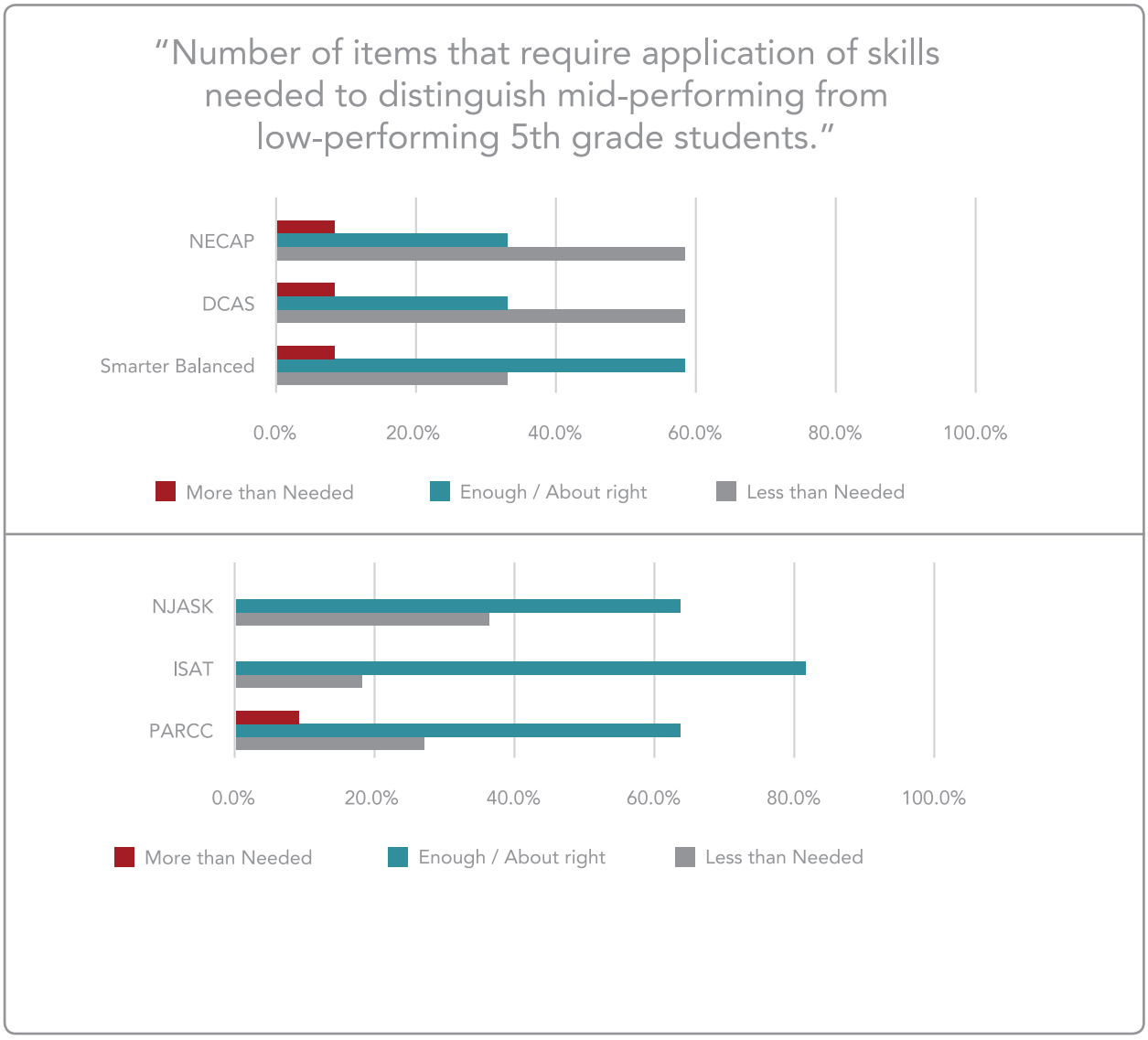


Figure 5. Average percent of teachers who indicated the: "Number of items that require application of skills needed to distinguish mid-performing from low-performing fifth grade students" was more than needed, enough/about right, or less than needed.

5. **While the new consortia assessments are more rigorous and demanding, they are grade-level appropriate, and 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 remain developmentally appropriate to the tested grade level. A strong majority of the teachers found the range and depth of content on the new tests to be appropriate for fifth grade students. On average, 74% of the teachers across both panels agreed or strongly agreed that the depth of content represented on the new consortia tests are grade-level appropriate. Approximately 83% of the teachers typically agreed or strongly agreed that the range of content represented on the new tests is grade- level appropriate.

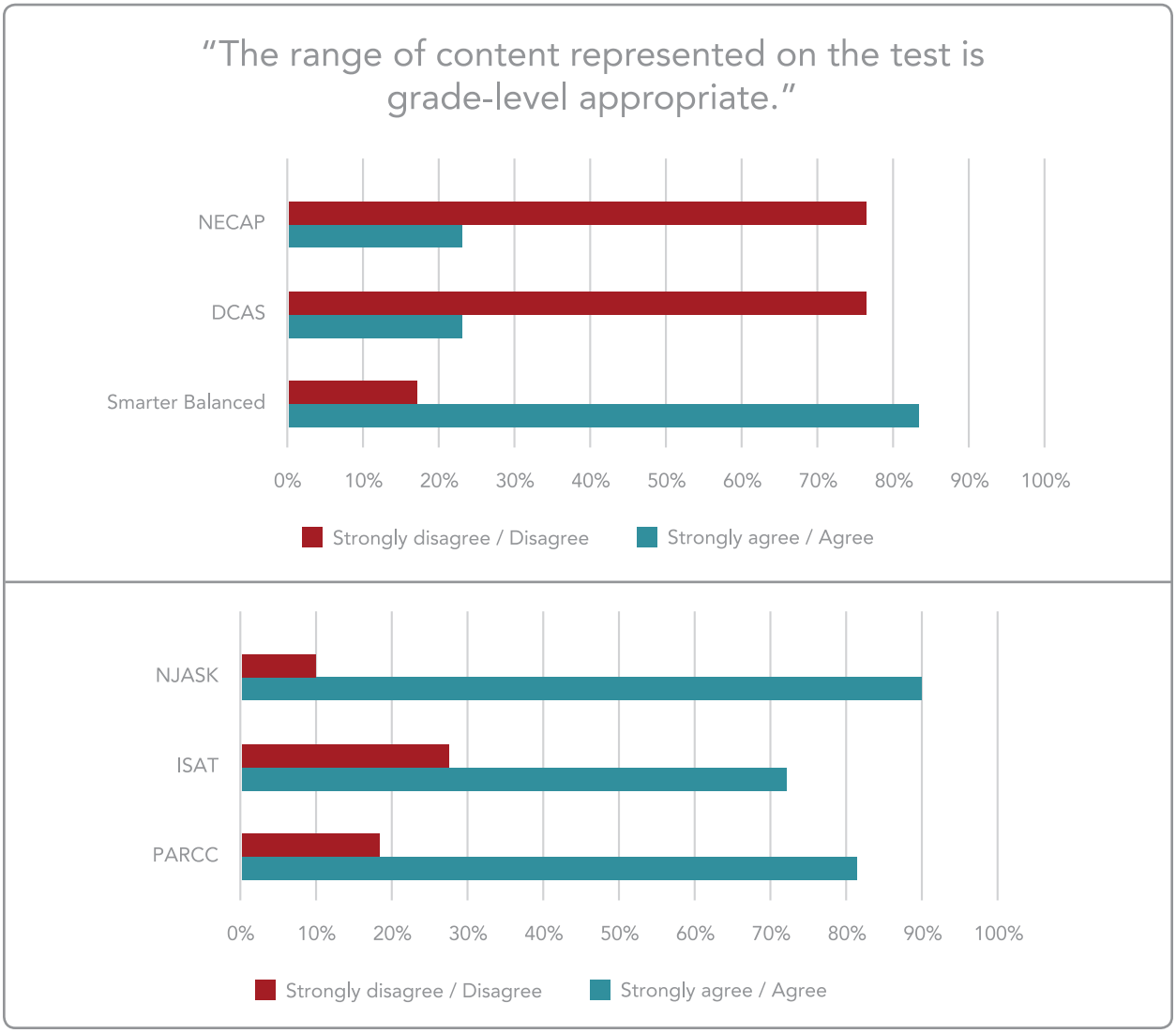


Figure 6: Percent agreement with: "The range of content represented on the test is grade-level appropriate."

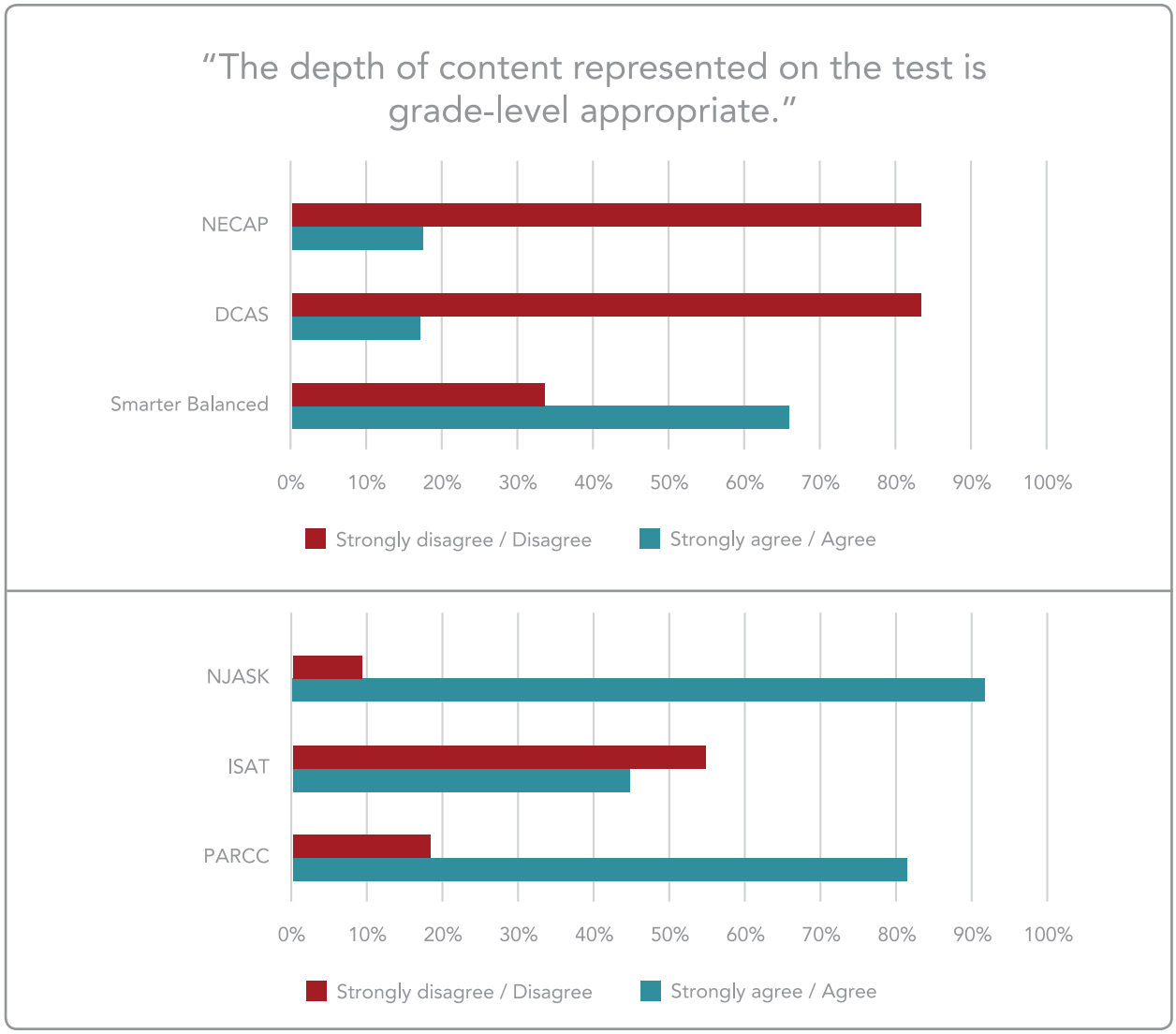


Figure 7. Percent agreement with statement: "The depth of content represented on the test is grade-level appropriate."

There was a clear consensus that the consortia tests do a better job of challenging students of middle and higher proficiency—students who might not be stretched by the former assessments. While some concerns were expressed that this improvement in the middle to upper range had come at some cost of thorough measurement at the lower end of student proficiency, panelists felt that the balance achieved in the consortia assessments was aligned with grade-level expectations and targets for students who had been well-served by the education systems in their jurisdiction. The new tests were also seen as an improvement over the prior state tests on this front.

Summary: The Transition to the New Consortia Assessments is Worth it

Our participants expressed concerns that the instruction of students currently in fifth grade had not been based on standards aligned to college and career readiness up to this point—that implementation of new curricula and teaching methods was still new and uneven. As a result, the new tests may be quite difficult for many students and initial results may be disappointing. This may be especially true in communities that have grown accustomed to high test scores year after year. However, they did not see this as a reason to retreat. To the contrary, these teachers understood that increasing expectations is the road to improved outcomes. The teachers emphasized a need for careful implementation of the new standards in the field, strong support and training for teachers using innovative techniques and novel materials, and patience from all stakeholder communities while the transition is in progress. Only then will we achieve the continuous improvement that all want.

> *"You may look good playing baseball in the A League when you are winning all the time, so you move up to AAA. And you [lose], because the caliber of player you are up against is suddenly so much higher. But that isn't a reason to drop back and play in A again—just to look good. No, you stay in AAA, your skills improve from playing at a higher standard, and soon you are winning again in the higher leagues. We all want to play in the Big Show and this is how you get there!"*

It is important to acknowledge some limitations of this study. First, this was a purposefully small study. Our eligibility criteria for participants (former state teachers of the year or finalists with direct knowledge of 5th grade ELA or math) and the rigorous, deep, and time-consuming review process meant that we only had 23 teachers participate. Second, we focused on the content of the assessments studied, not implementation. Third, our study was not able to take some of the unique features of PARCC and Smarter Balanced fully into account.

- PARCC administration in 2014-15 included two separate components: an "end of year" (EOY) test and a "performance based assessment." (Those will be merged together into a single assessment for this year.) Teachers on the PARCC panel in this study had access to both parts but only formally reviewed the EOY. As a result, some elements of the PARCC assessment, such as the extended response items that are intended to test students' knowledge and skills more deeply, were not fully taken into account in the report's findings.

- Smarter Balanced is an adaptive test, which means that the rigor and difficulty of the questions changes depending on individual student performance. Our teachers only looked at one Smarter Balanced form: one that approximates a student at the 60th percentile of performance. Students either above or below the 60th percentile would have seen a different assessment.

Overview of Study

A team convened by the National Network of State Teachers of the Year (NNSTOY) set a research direction for examining the differences between former and new state summative assessments (in this study, either the PARCC or 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, below in Figure 8.

Figure 8. Claims made about the Common Core Assessments.

A small-scale study was designed 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 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.

Methodology

The study was divided into two phases. The first phase comprised an in-depth review and alignment of four former state assessments and two new state assessments developed in consortia. Each panel examined two former assessments and one of the new consortia assessments. The assessments were grouped so that panelists from a state that gave us access to their former assessment reviewed that assessment as well as the new assessment currently being given in that state (e.g., a NJ panelist reviewed the NJ former assessment and the PARCC assessment as well as another former assessment from a state where PARCC is now given). These panels took place over two days and the study plan is described below.

The review was conducted using Norman Webb's Depth of Knowledge (DOK; Webb, 1997). 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 studying 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 items and tasks require of students. There are four DOK levels, each with increasing complexity or cognitive demand, as shown in Figure 9.

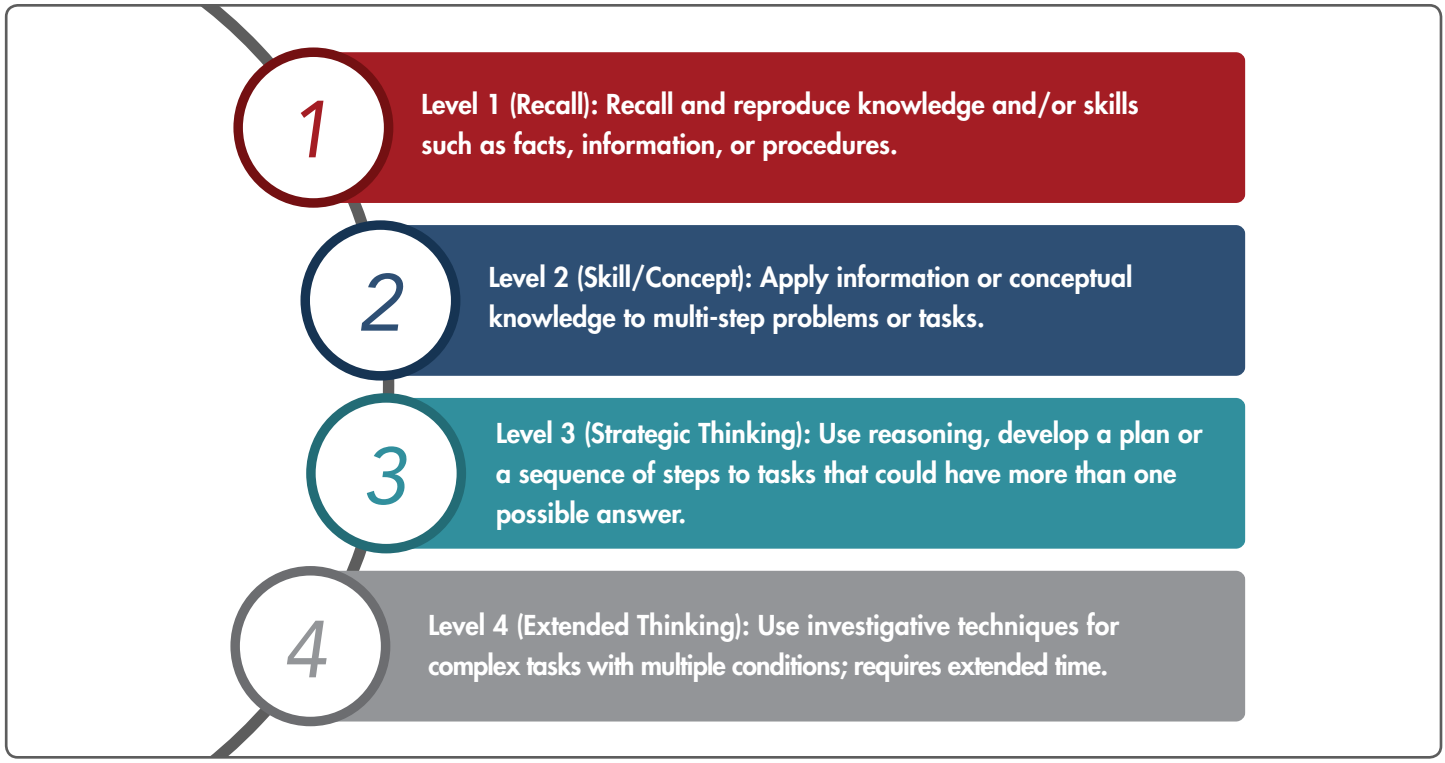


Figure 9. Webb Depth of Knowledge Levels (Webb, 2005).

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 Teacher of the Year or were finalists), this design seemed the most appropriate. The quantitative component comprised 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

Four fifth grade ELA and Math former and two new state end-of- year (EOY) summative assessments were reviewed for the study. These assessments were divided between two panels of reviewers to reduce burden and provide ample time for evaluation and discussion. The first panel reviewed New Jersey's Assessment of Skills and Knowledge (NJASK), the Illinois Standards Achievement Test (ISAT), and the PARCC consortium test. This panel will be called the PARCC panel in this report. The second panel reviewed New Hampshire's New England Common Assessment Program (NECAP), the Delaware Comprehensive Assessment System (DCAS), and the Smarter Balanced consortium test . This panel will be called the Smarter Balanced panel in this report. The former state assessments were all viewed in paper format, which is the format in which they were administered to students. The new state assessments are computer-based. Appendix A includes more details about the assessments reviewed.

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 fifth grade allowed us to call upon 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 engaging with the assessments.

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.

---

[2] The PARCC assessments are composed of two components: a performance-based assessment (PBA) and end-of-year (EOY) assessment. To ensure that participating teachers were reviewing substantially similar test forms for each assessment, the PARCC panel only formally reviewed the PARCC end-of-year (EOY) component, though optional access to the PARCC performance-based assessments (PBA) component was available. This meant that teachers did not formally review the extended response items in ELA and math that were included in PARCC in the PBA form component.

[3] Smarter Balanced is an adaptive test, but teachers only reviewed one linear form based on a student at the 60th percentile of the proficiency distribution at 5th grade.

Participants

We convened 23 outstanding educators for the study, each consisting of State Teachers of the Year and Finalists recognized for excellence in classroom practice. Participants were selected and divided into two panels . The panels were 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 fifth grade assessments as a transition point between elementary and middle school. We included teachers with familiarity of the fifth-grade content through vertical grade-level alignment;
- States –we included two or three teachers from each of the states whose assessments we examined and we sought geographic diversity. The full group of participants included teachers from AZ, CA, DC, DE, GA, IL, KY, MD, NH, NJ, NM, SD, UT, and VT;
- 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.

Teachers were assigned to one of the two panels based on which of the two consortia assessments their state is using. Eleven teachers participated on the PARCC panel, 12 on the Smarter Balanced panel. There were two or three teachers representing the state in which the prior assessments were administered on each panel. A person on each panel was included from a state that is not using either of the new assessments 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 with familiarity of the fifth grade instruction, as the focus of the evaluation was fifth grade Math and ELA assessment. 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.

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. Both panels met in Chicago, Illinois. Chicago served as a neutral and accessible location for panelists who represented multiple regions of the country. The PARCC panel met in late August and the Smarter Balanced panel met in early September, each for two days of on-site activities. 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
2 they would be reviewing. During the orientation, participants were encouraged to work through the items as if they were a typical well-prepared fifth grade student as described in the box; not necessarily the kind of student who happened to be in their individual classrooms. This provided a common lens through which to evaluate the cognitive demand associated with a particular assessment item.

Who is the fifth 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 for us that the educators demonstrated consistency in their interpretation and application of DOK levels. Publicly available 4th through 7th grade ELA, Math, and Social Studies items 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 two panels reviewed between four and six sample items, depending on how quickly they were able to come to consensus. Both panels' DOK ratings were quite similar even at the beginning of this process; perhaps because a number of the participants already had strong familiarity with Webb through their own professional experience.

After the consensus activity, the order in which the assessment was reviewed was randomly assigned to participants. The order of review was different for each group of panelists to mitigate fatigue effects on the data. For example, for the PARCC panel, one group of participants started with the assessment NJASK, another group started with the PARCC assessment, and the third group started with the ISAT. Participants were given a tutorial on how to access and navigate the two computer-based Common Core assessments.

Paper copies of the other state assessments were distributed. The state tests were provided in paper form as this was the form in which they were administered. Participants were given approximately two hours to complete their review of each assessment. Depending on their 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 the participant was 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.

---

[4] As a condition of participation in the study, two representatives from Smarter Balanced assessment consortium gave a 30- minute in-person presentation on the development and structure of the assessment to the Smarter Balanced panel. The representatives were not present for any of the other study activities.

Many of the assessments in this study have had 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.

The major takeaways from the DOK ratings assigned by the participants were that the former state math assessments and the consortia math assessments both largely comprised Level 1 and 2 items. There was a marginal increase in the cognitive complexity seen in the consortia tests, but students were generally expected to recall information and apply that information and conceptual knowledge to respond to test items and tasks on all three assessments. On the ELA assessments, while the former state assessments also largely captured Level 1 and 2 items, the consortia assessments largely captured Level 2 and 3 items. The cognitive complexity increased, as intended. Students are generally expected to apply information and conceptual knowledge, as well as apply logic and reasoning to respond to items and tasks on the consortia tests.

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 fifth grade student" and only fifth 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.

In addition, clarification was given to the items that spoke to how well the items would surface information about student performance levels. The intent of these particular survey items was to determine the extent each test contained content that would measure student ability at the low-, mid-, and high-performance levels. For example, if all of the test items were written at a high complexity level (Level 3 or Level 4), the test would surface a good deal of information about mid- to high-performing students, but very little information about low-performing students. Those low-performing students would likely not be able to respond correctly for many of the items. 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 to rest. Items with interesting or unclear responses were selected for clarification during the whole-group discussion. The discussion was audio recorded with participants' permission and transcribed 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. 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?

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. The detailed response data from each test and the individual panels is presented in Appendix F.

Question 1: Range of Important Knowledge and Skills

Some example items will illustrate the results on this focus question. Panelists were asked the extent to which they agreed with the statement: "The range of content represented on the test is grade-level appropriate." The results are shown in Figure 10.

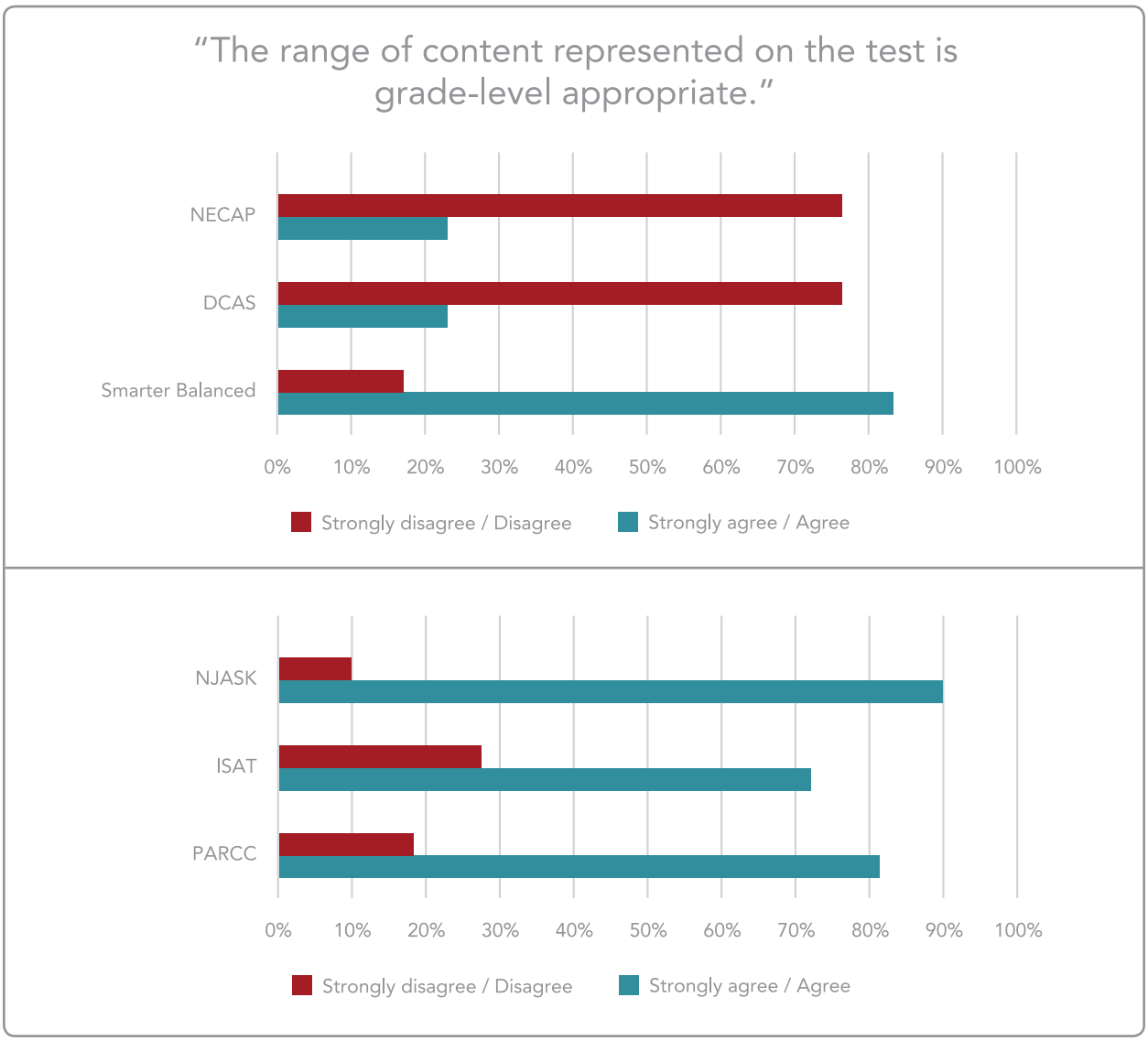


Figure 10: Percent agreement with: "The range of content represented on the test is grade-level appropriate."

Both consortia assessments were highly rated. The former state tests received more mixed evaluations. The NJASK was rated similarly to the consortium assessments, and the ISAT nearly as well. The NECAP and DCAS did not fare as well. As a group, the former state assessments were not rated as highly as the consortia assessments, but as can be seen from the data, some individual former state assessments did quite well on this item.

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 fifth grade Math/ELA classroom." for all three tests in their panel. The results are shown in Figure 8. On average, 78% of teachers strongly agreed or agreed that the consortia tests measure the most important Math and ELA knowledge and skills taught in fifth grade classrooms. They did not endorse this statement for the former assessments. On average, 57% of the teachers strongly disagreed or disagreed that the former assessments measured the most important Math and ELA fifth grade knowledge and skills taught in excellent classrooms. For this item, the consortia assessments clearly were rated higher than all four of the former state assessments.

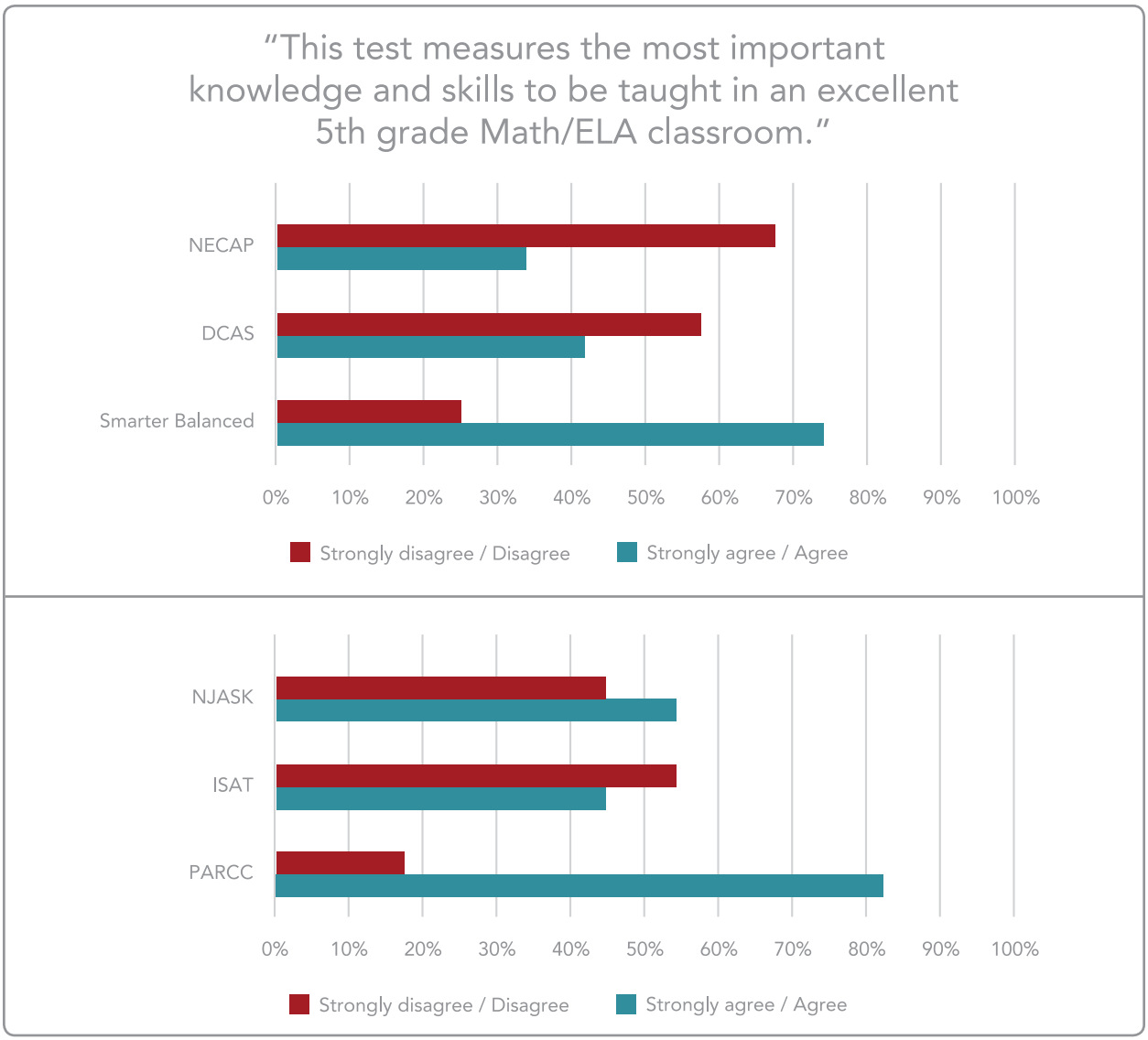


Figure 11. Percent agreement with statement: "This test measures the most important knowledge and skills to be taught in an excellent fifth grade Math/ELA classroom."

There was clear consensus that the consortia tests do a better job of measuring higher-level cognitive skills than the former assessments, also reflected in teachers' DOK ratings.

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 consortia tests as having about the right amount or enough items (the green-colored bars in Figure 12) of this type.

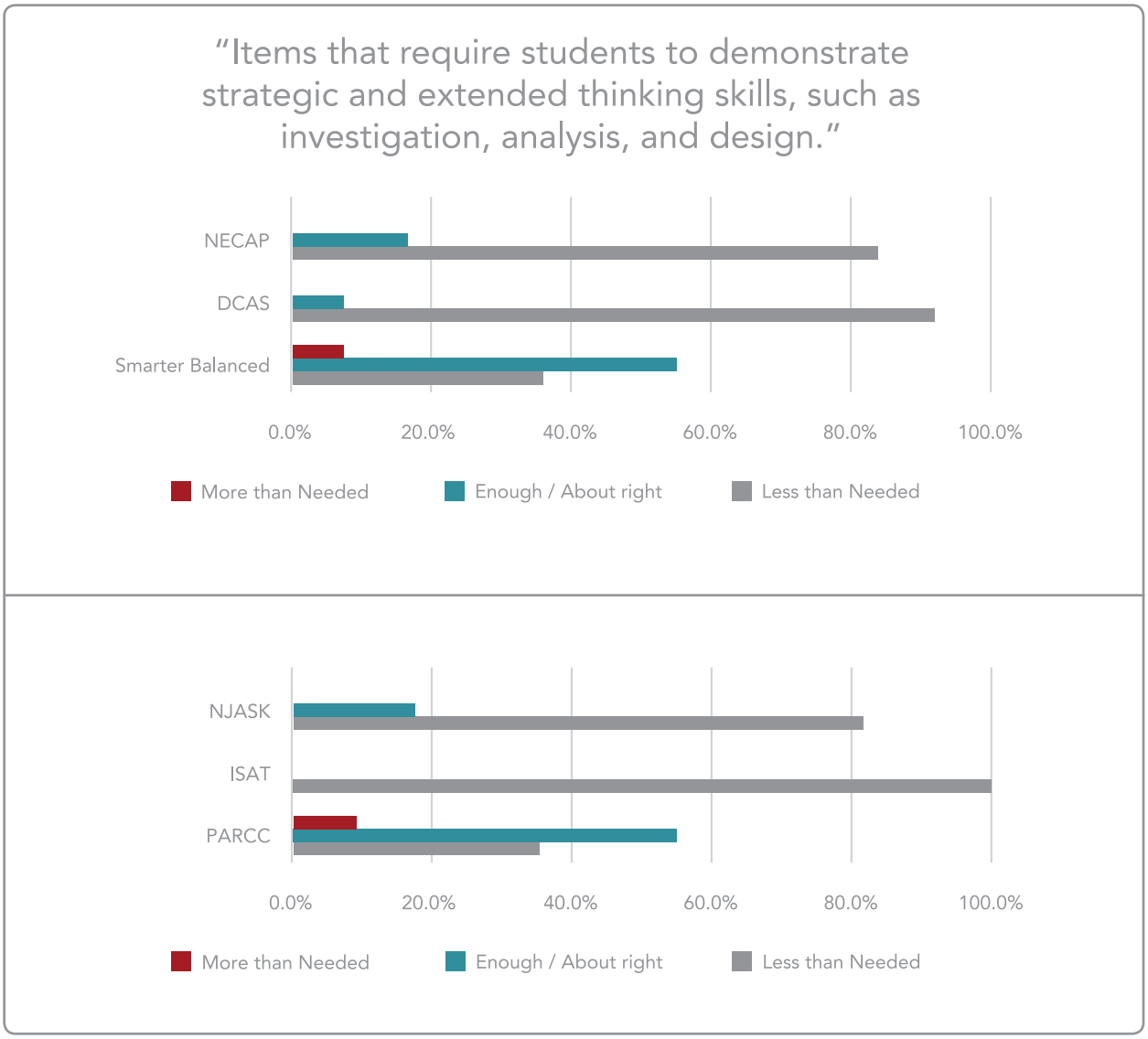


Figure 12. Percent of teachers who indicated the number of test items that "require students to demonstrate strategic and extended thinking such as investigation, analysis, and design" was "more than needed," "about right/enough," or "less than needed."

In contrast, the former assessments were perceived as having gaps in their measurement of the kinds of deep learning that take place in their classrooms. An average of 89% of the teachers, across the two panels, indicated the former assessments contained fewer items than needed (the gray-colored bars in Figure 9) that "require students to demonstrate strategic and extended thinking such as investigation, analysis, and design." Only the consortia assessments were considered by any panelist to have more items than needed of this type (the green-colored bars in Figure 9). The next concern we addressed is whether the consortia tests measure the full range of cognitive complexity that is important to teachers in a balanced way for fifth graders.

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

Teachers tended to agree or strongly agree that the consortia 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, the former assessments emphasized lower-level skills than the kinds of skills that would require strategic or extended thinking, for example. The data are shown in Figure 13.

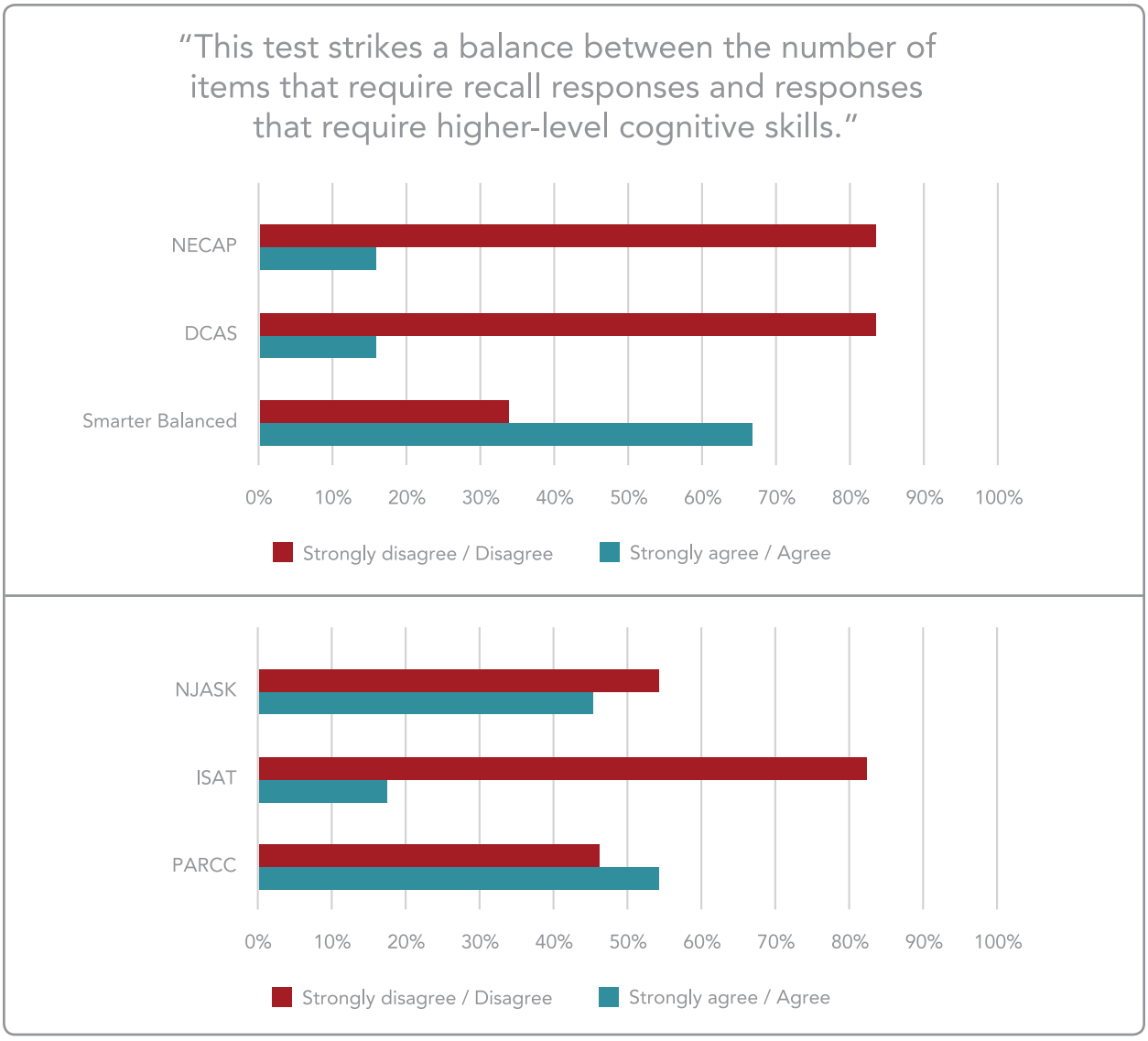


Figure 13. Percent agreement with statement: "This test strikes a balance between the number of items that require recall responses and responses that require higher-level cognitive skills."

These are strong differences between the former state and new consortia assessments when averaged together as a group; 75% strongly disagree or disagree with the statement versus 40% strongly disagree or disagree, respectively. But the former state assessments were not all evaluated similarly when the data are broken out, as can be seen above. On this item, the NJASK scores were more evenly balanced compared to the other three former state assessments, although it was not rated as highly as either of the consortium assessments.

Another set of items asked about test items that required students know and do different types of things and placed a variety of levels of cognitive demand for response. The panelists' responses are shown in Figures 14, 15, and 16.

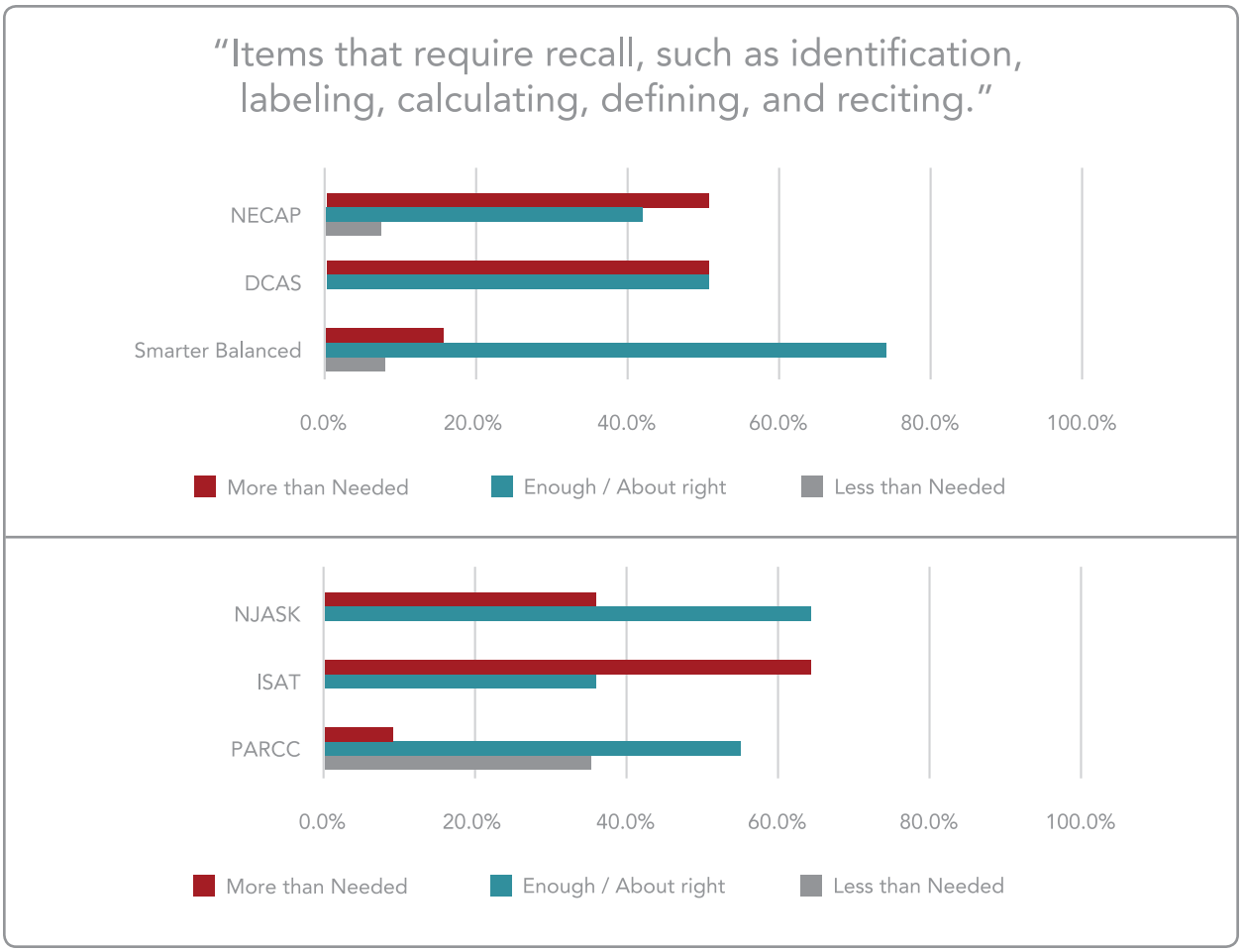


Figure 14. Percent of teachers who indicated the number of "Items that require recall, such as identification, labeling, calculating, defining, and reciting" was "more than needed," "about right/enough," or "less than needed."

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

In Figure 15, the response data for the next type of 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 assessments receiving the highest ratings in this category. Several also received several ratings indicating that there are less items than needed (the gray-colored bar) of this type as well. The evaluation of the DCAS assessment is fairly evenly split among having too many, enough, and less items than needed of this type.

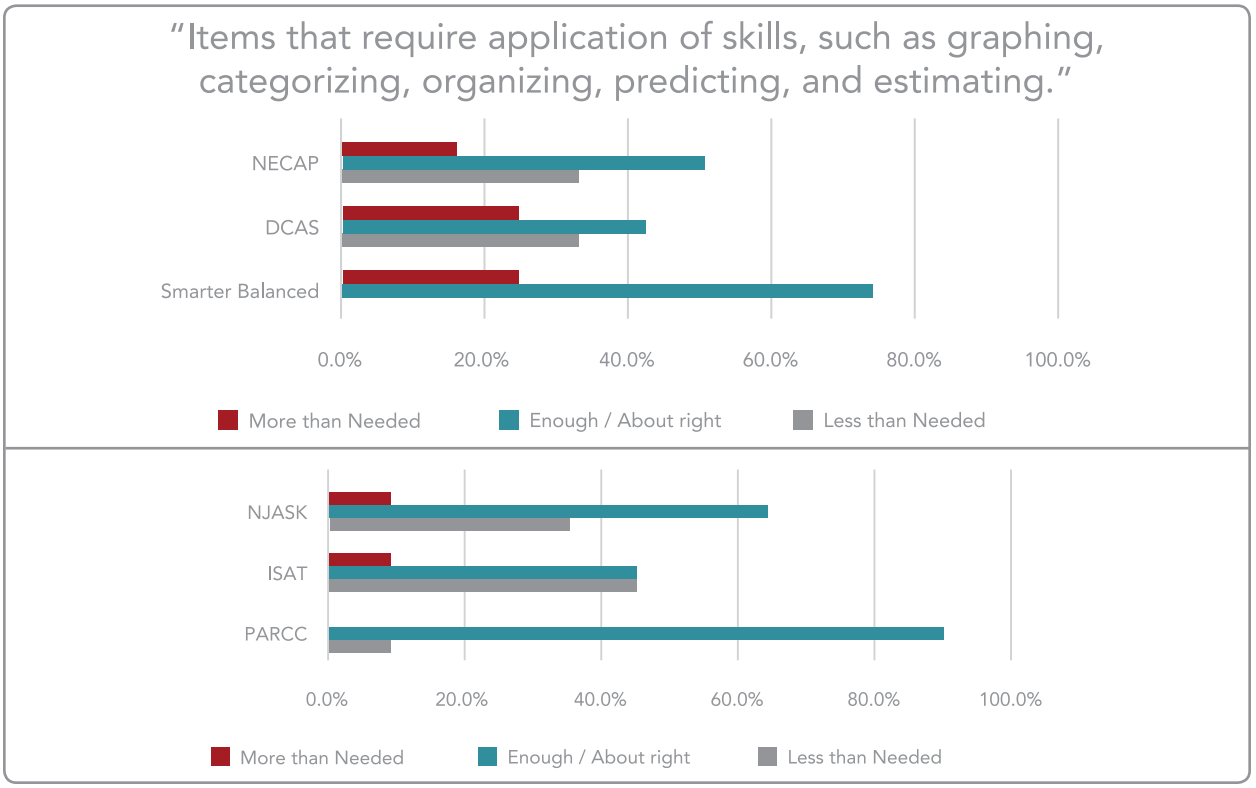


Figure 15. Percent of teachers who indicated the number of "Items that require application of skills, such as graphing, categorizing, organizing, predicting, and estimating," was "more than needed," "about right/ enough," or "less than needed."

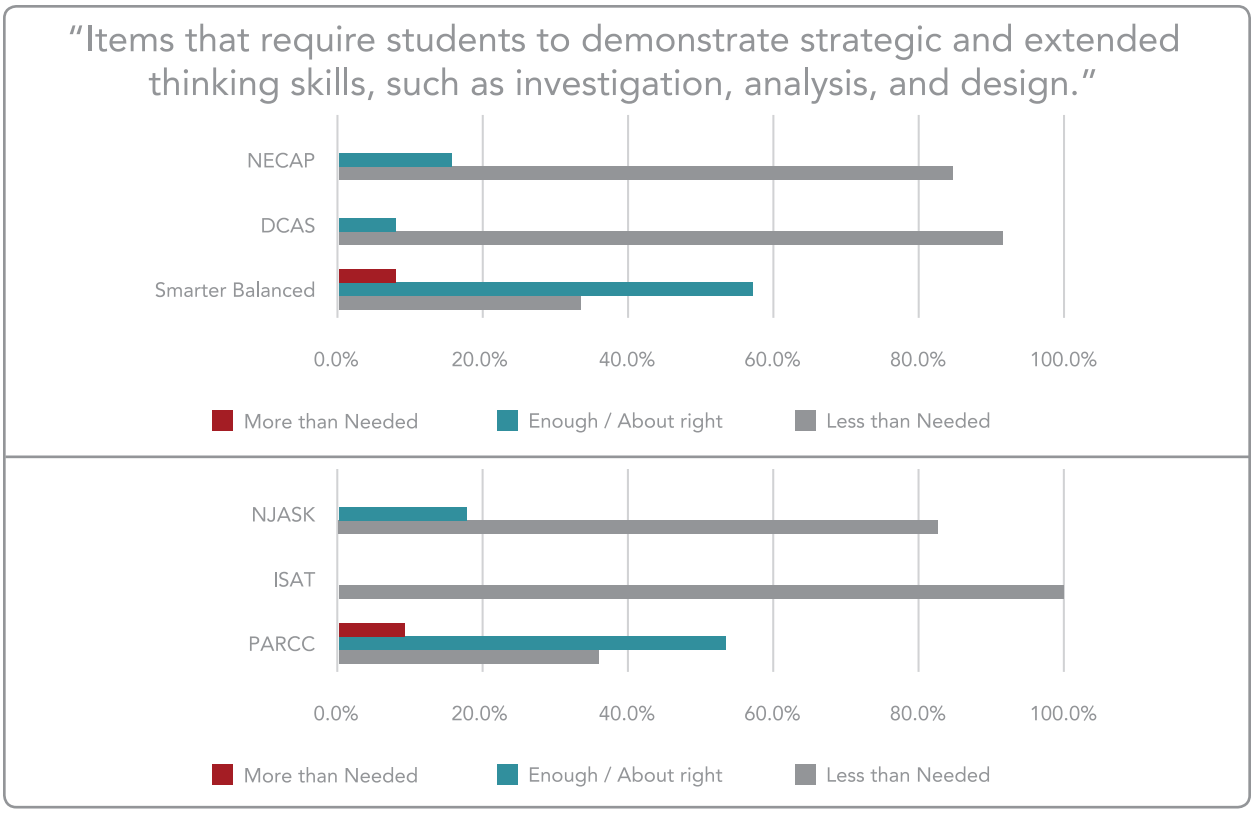


Figure 16. Percent of teachers who indicated the number of "Items that require students to demonstrate strategic and extended thinking skills, such as investigation, analysis, and design," was "more than needed," "about right/enough," or "less than needed."

In Figure 16, the participants' evaluation of the occurrence of items that are cognitively complex and demanding is presented. This description is consistent with items typically aligned into Webb DOK Level 3 or 4. All the former state assessments are rated as deficient in the number of this type of item (the gray-colored bars). The consortium assessments clearly stand out on this item, each with more than half of the panelists rating the number of this type of items on consortium assessment as "about right/enough."

While no assessment appears to have assessed of the full range cognitive complexity perfectly, the balance achieved on the consortium assessments is clearly an improvement on the former state assessments.

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

There was strong consensus that the two consortia assessments measured excellent fifth grade instruction. The views concerning the former assessments were varied, again. One panel generally perceived the former assessments they reviewed as measuring content that was aligned with excellent fifth grade instruction (i.e., NJASK and ISAT). The other panel, however, did not share this perspective. The former assessments they reviewed did not measure content that was aligned with excellent fifth grade instruction (DCAS and NECAP).

In addition, consortia tests measure the learning outcomes that participant teachers would set for student learning in fifth grade classes. As shown in Figure 17, approximately 73% and 91% of teachers across the Smarter Balanced and PARCC panels, respectively, strongly agreed or agreed with this statement in regards to the consortia tests. In comparison, between 36% and 64% of teachers across the panels strongly agreed or agreed with this statement in regards to the former assessments. With the exception of NJASK, this was not a majority.

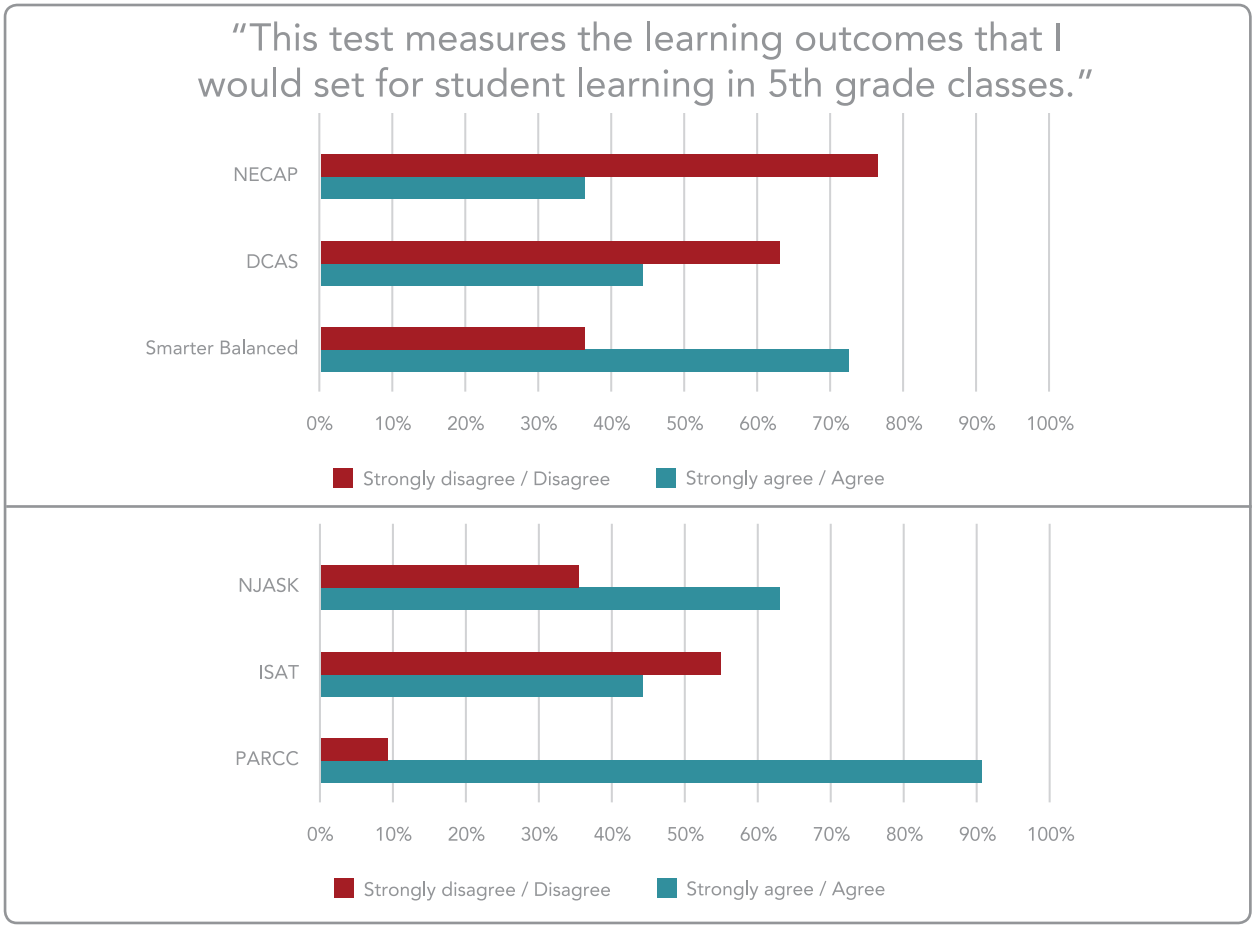


Figure 17. Percent agreement with statement: "This test measures the learning outcomes that I would set for student learning in 5th grade classes."

The data also show that 91% of the teachers strongly agreed or agreed with this statement when evaluating the consortia tests: "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 two consortia tests are rated very highly on this item as shown in Figure 18. The former state tests have less support for this statement. NJASK and NECAP scored higher, but even in those cases, more than half the panelists disagreed or strongly disagreed with this statement in reference to those assessments.

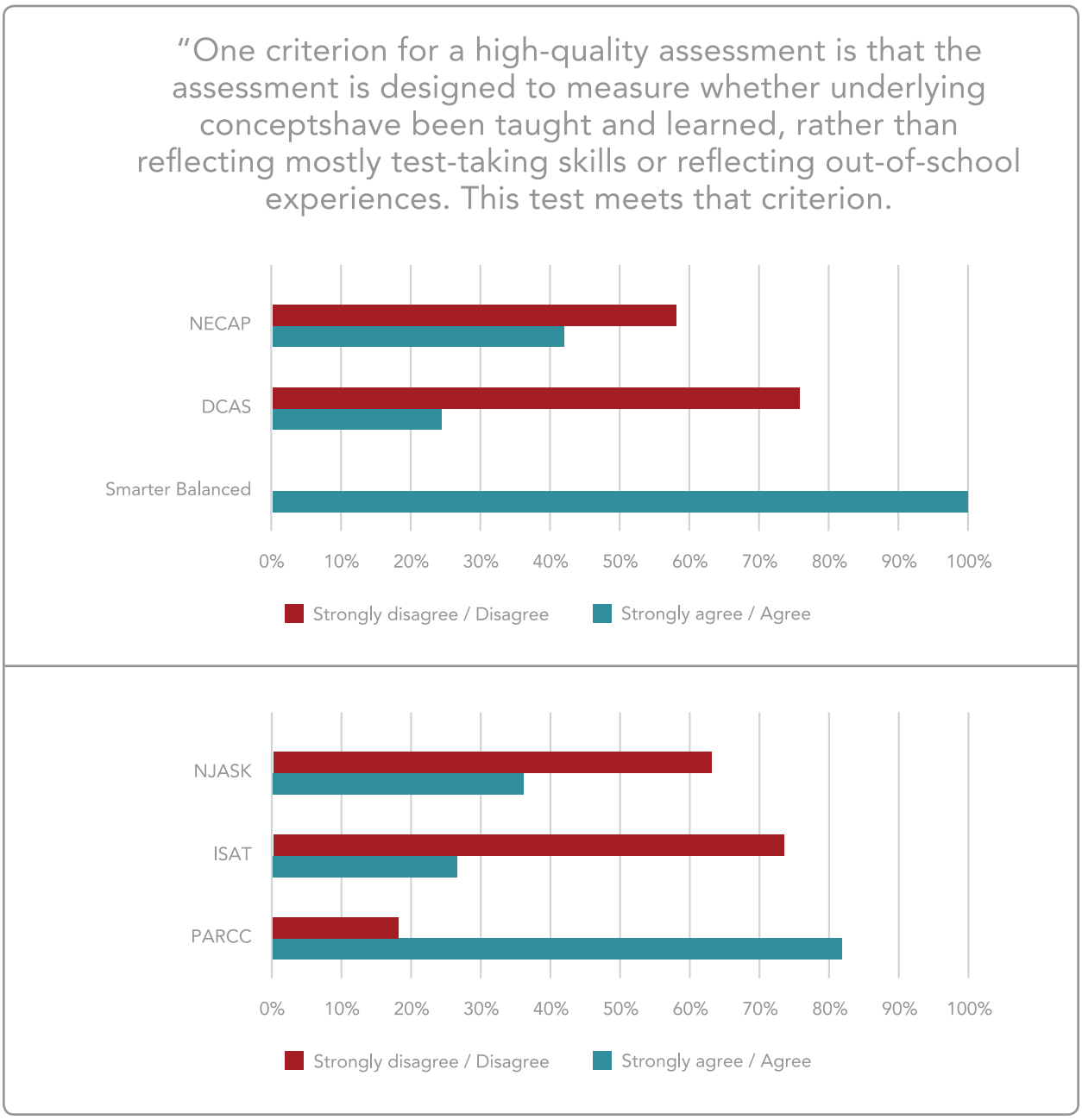


Figure 18. Percent agreement with statement: "One criterion for a high-quality assessment is that the assessment is designed to measure whether underlying concepts have been taught and learned, rather than reflecting mostly test-taking skills or reflecting out-of-school experiences. This test meets that criterion."

Teachers also agreed that the consortia assessments would complement and support their lesson planning and efforts toward high-quality instruction. An overwhelming majority (91% and 75%) strongly agreed or agreed with the following statement: "If I backwards-mapped a fifth grade lesson against items like those on [PARCC and Smarter Balanced]", it would help inform my lesson plan and guide me toward high-quality instruction." The results varied for the former assessments.

On one panel, over one-half of the teachers strongly agreed or agreed that the assessment items would inform their lesson plans and guide high-quality instruction, on average. On the other panel, teachers endorsed this statement more strongly for one assessment than the other.

Further, the majority of teachers strongly agreed or agreed that preparing students for the PARCC and Smarter Balanced assessments (100% and 75%, respectively) would require meaningful lessons and learning beyond skill and drill practice (Figure 19). This was also true for the NJASK assessment. However, teachers did not strongly endorse this statement for the other three former assessments.

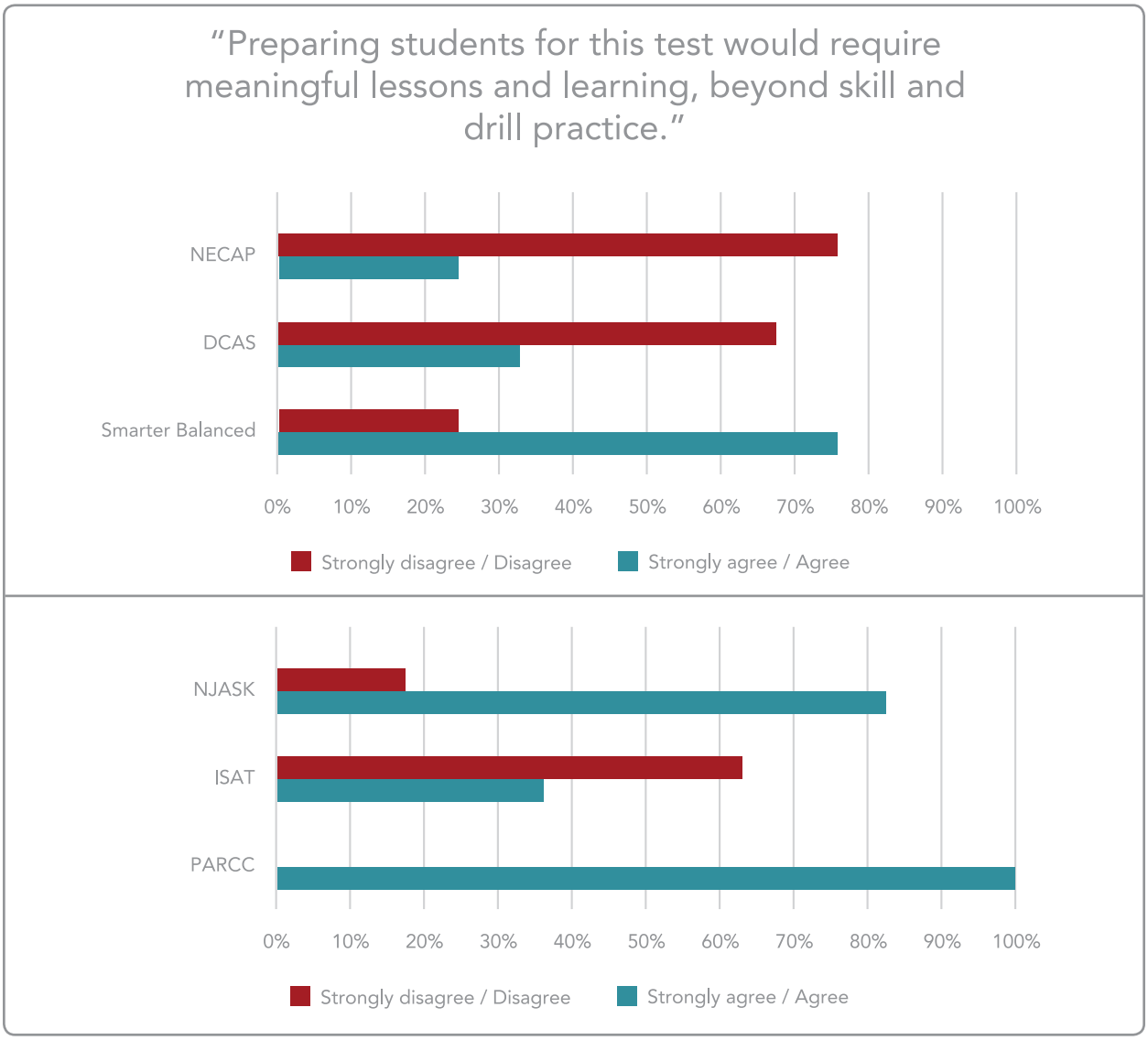


Figure 19. Percent agreement with statement: "Preparing students for this test would require meaningful lessons and learning, beyond skill and drill practice."

An important aspect of successful and effective teaching is formative assessment. The findings were mixed across the board concerning teachers' agreement with the following statement for each test: "The optimal formative assessments that I would give to fifth grade students measure concepts not addressed by this test." Disagreement was above 80% for the consortia assessments, indicating that the teachers believed that the concepts were addressed on these assessments. NJASK, NECAP, and DCAS all received mixed endorsements on this item, nearly evenly split on agreement and

disagreement. The strongest agreement was on ISAT, with more than 70% of teachers indicating that their best formative assessments measured concepts that the ISAT did not address.

Generally, the consortia assessments are better aligned with strong instructional practices used in the classroom than former assessments. However, NJASK was the standout among the former assessments, particularly with regard to the alignment between its content and learning outcomes and excellent instruction for fifth grade students.

Question 4: Information relevant to a wide range of performers

The consortia assessments generally provide information that is relevant to mid-performing and high- performing students. When asked if there were less than, enough, or more than the number of items that would surface information about fifth grade students at higher-ability levels to inform instructional strategies, the majority of teachers indicated that the consortium assessments had enough of those items. However, the same was not true for the former assessments. Across both panels, over 80% of the teachers, on average, indicated that the former assessments had less than the number of items needed to surface information about fifth grade students at higher-ability levels to information instructional strategies.

> *"For my really high-performing students, I actually really do want those high levels of questions there. Because if you're going to give me a [test targeted at] low-level ability [students] for fifth grade, that doesn't give me an overall perception of a summative assessment or where they're at in the ultimate journey. I don't want to give a [former] assessment that tells me that my students have basic skills that aren't going to get them anywhere. I want to give something like an [Consortium Test] that's going to tell me where—if I think they're highly performing, are they performing highly? I actually want the bar to be a high standard. And then it's my job to give formative assessments and other pieces to determine what I need to change for instruction to get them to that higher level.*

The former assessments tended to skew toward the low-performing students, although there were some clear exceptions in certain cases. For example, teachers were asked to indicate whether the items that would surface information about lower-ability levels to inform their instructional strategies was less than needed, enough, or more than needed. They tended to view the PARCC assessments as offering less than the number of items needed and the Smarter Balanced assessment as offering enough of the number items needed, as shown in Figure 20. A large majority of teachers indicated that for two of the four former assessments (NJASK and ISAT), the number of items was enough or about right. For the other two (NECAP and DECAS), they tended to believe the number of items that would surface information about lower-ability levels was less than what was needed, similar to PARCC.

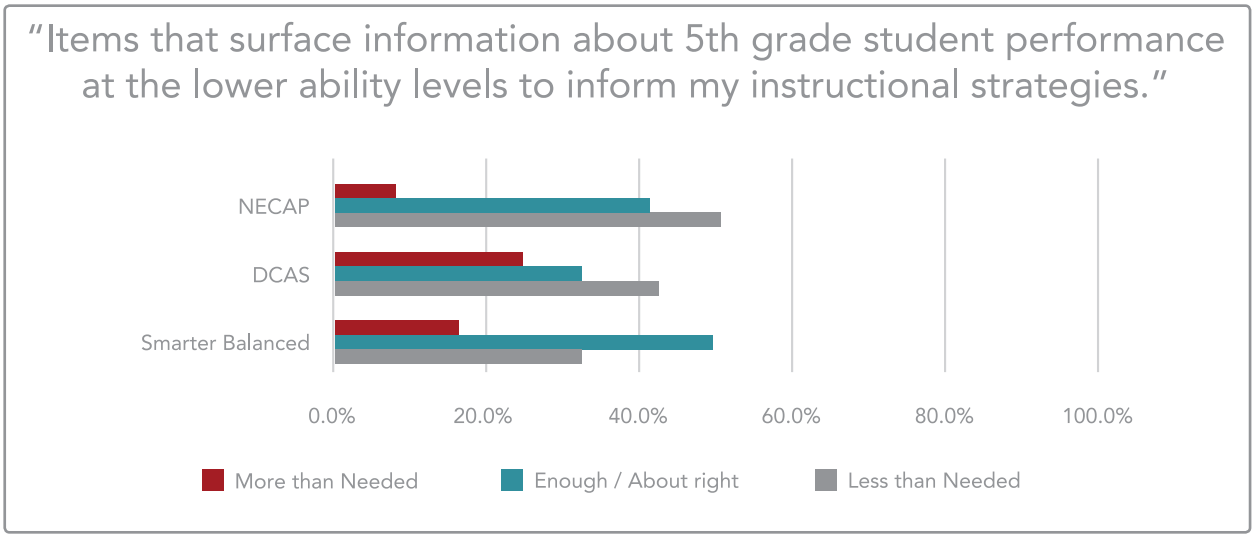


Figure 20 (A)

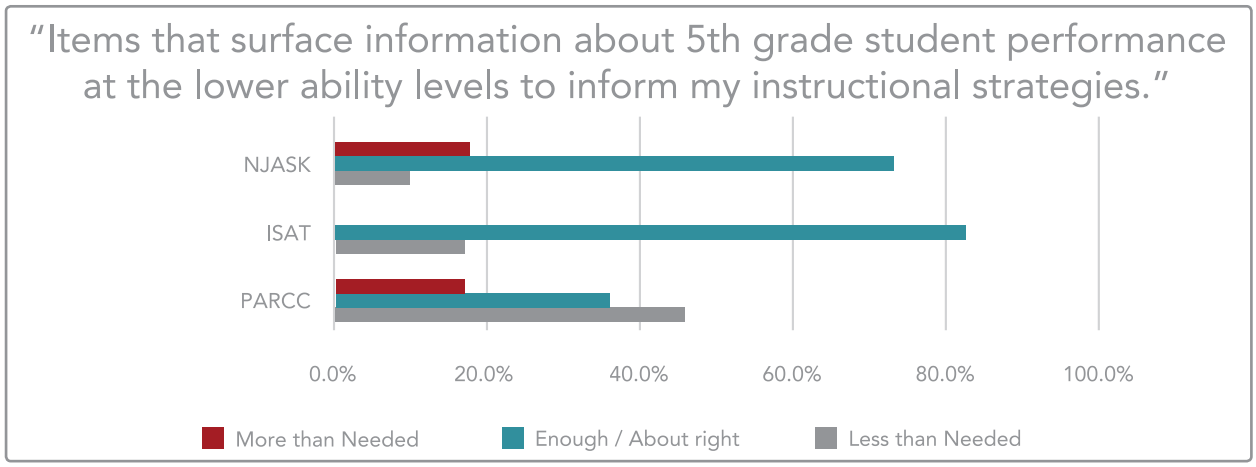


Figure 20 (B). Percent of teachers who indicated the: "Number of items that surface information about fifth grade student performance at the lower ability levels to inform my instructional strategies" was more than needed, enough/about right, or less than needed.

When asked if the number of items that would allow them to distinguish between low-level student performance and mid-level student performances was less than needed, about right, or more than needed, teachers' responses also differed quite a bit between panels. One panel indicated the number of items that require application of skill to distinguish between low-performing and mid-performing fifth grade students was about right for all three assessments (NJASK, ISAT, and PARCC), as shown in Figure 21. The other panel indicated that the number of items was less than what was needed for the former assessments and about right or enough only for the consortium assessment (Smarter Balanced).

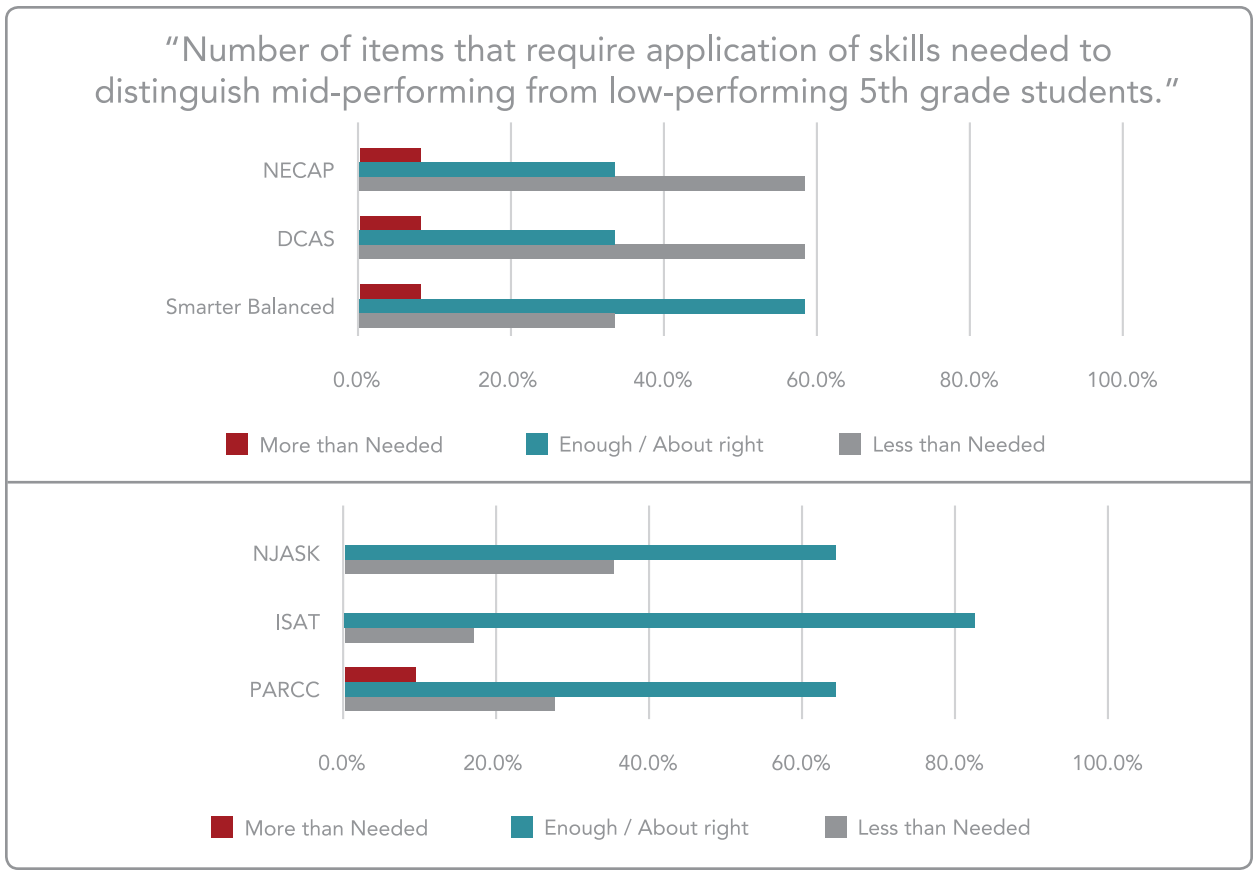


Figure 21. Average percent of teachers who indicated the: "Number of items that require application of skills needed to distinguish mid-performing from low-performing fifth grade students" was more than needed, enough/about right, or less than needed.

In keeping with the perception that former assessments skewed toward lower cognitive skills, teachers generally thought that the former assessments had fewer items that required complex thinking skills than were needed to distinguish mid-performing and high-performing students, but the new consortia assessments possessed about the right amount or enough of those items, as shown in Figure 22.

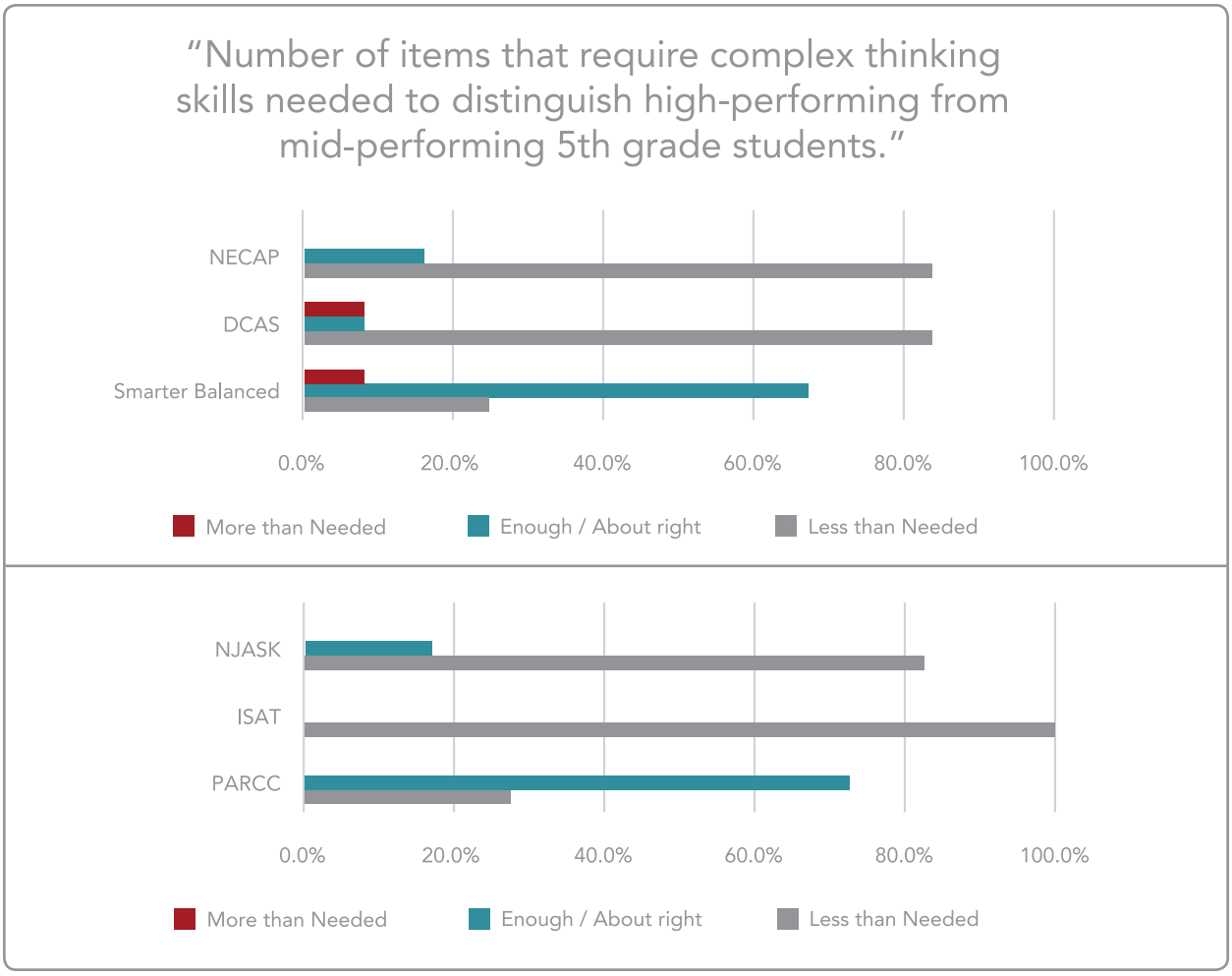


Figure 22. Average percent of teachers who indicated the: "Number of items that require complex thinking skills needed to distinguish high-performing from mid-performing fifth grade students" was more than needed, enough/about right, or less than needed.

Question 5: Grade appropriate

Finally, evidence shows that that the new consortia tests measure the learning outcomes that the teachers believe are appropriate for student learning in fifth 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 consortia assessments once they had reviewed them to get at this issue. The results are shown in Figures 23 and 24.

First, in Figure 23, we asked panelists if the assessments were more cognitively demanding than warranted for the grade level. Recall that the fifth grader to be considered for this study was one who has been well-served by the educational system in the state or jurisdiction of the panelist (because the goal of summative assessment should not be to expect less than adequate education). The results are quite striking. For three of the four former state assessments, 100% of the panelists disagreed or strongly disagreed with this statement, and for the remaining one (DCAS), the

disagreement was above 80%. Clearly, the educators did not think that the former assessments were too cognitively demanding for the grade level. But neither did the majority of the panelists think this of the consortia assessments, although the margins were not quite as wide as for the former state assessments.

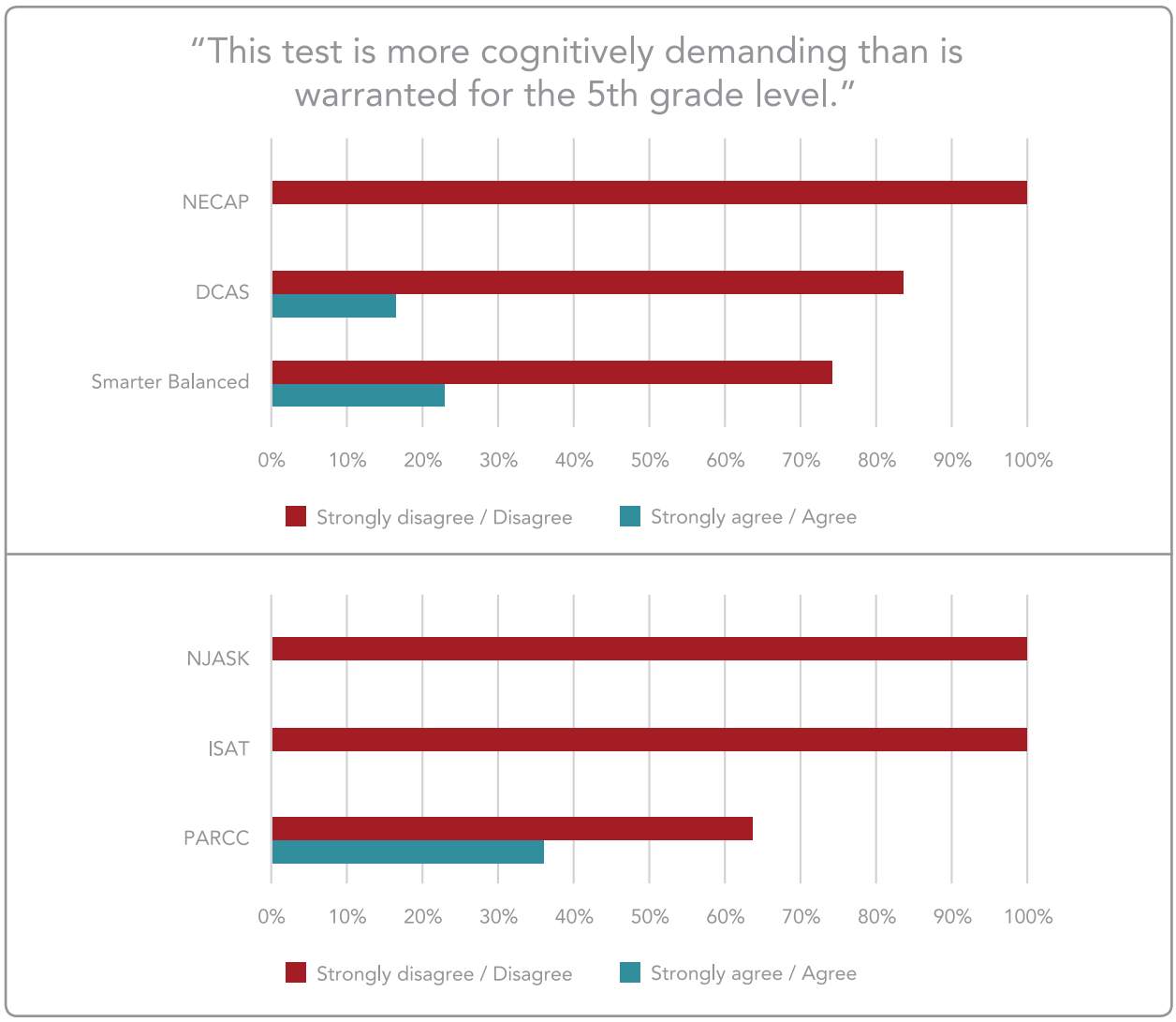


Figure 23. Percent agreement with statement: "This test is more cognitively demanding than is warranted for the fifth grade level."

Having asked if the assessments were too demanding for the grade level, we also asked the opposite question: are they not demanding enough? Part of the ongoing debate around educational standards is about setting them at the "Goldilocks" point, neither too high nor too low. The responses are shown in Figure 24, with interesting patterns. Neither consortium assessment is seen as less cognitively demanding than warranted for the grade level, which is not a surprise given the overall pattern of the data. DCAS, ISAT, and NECAP were all seen as insufficiently cognitively demanding for the grade level by the majority of the panelists. The results for the NJASK are nearly evenly split, with about half agreeing and half disagreeing that it is less cognitively demanding than warranted for the fifth grade level.

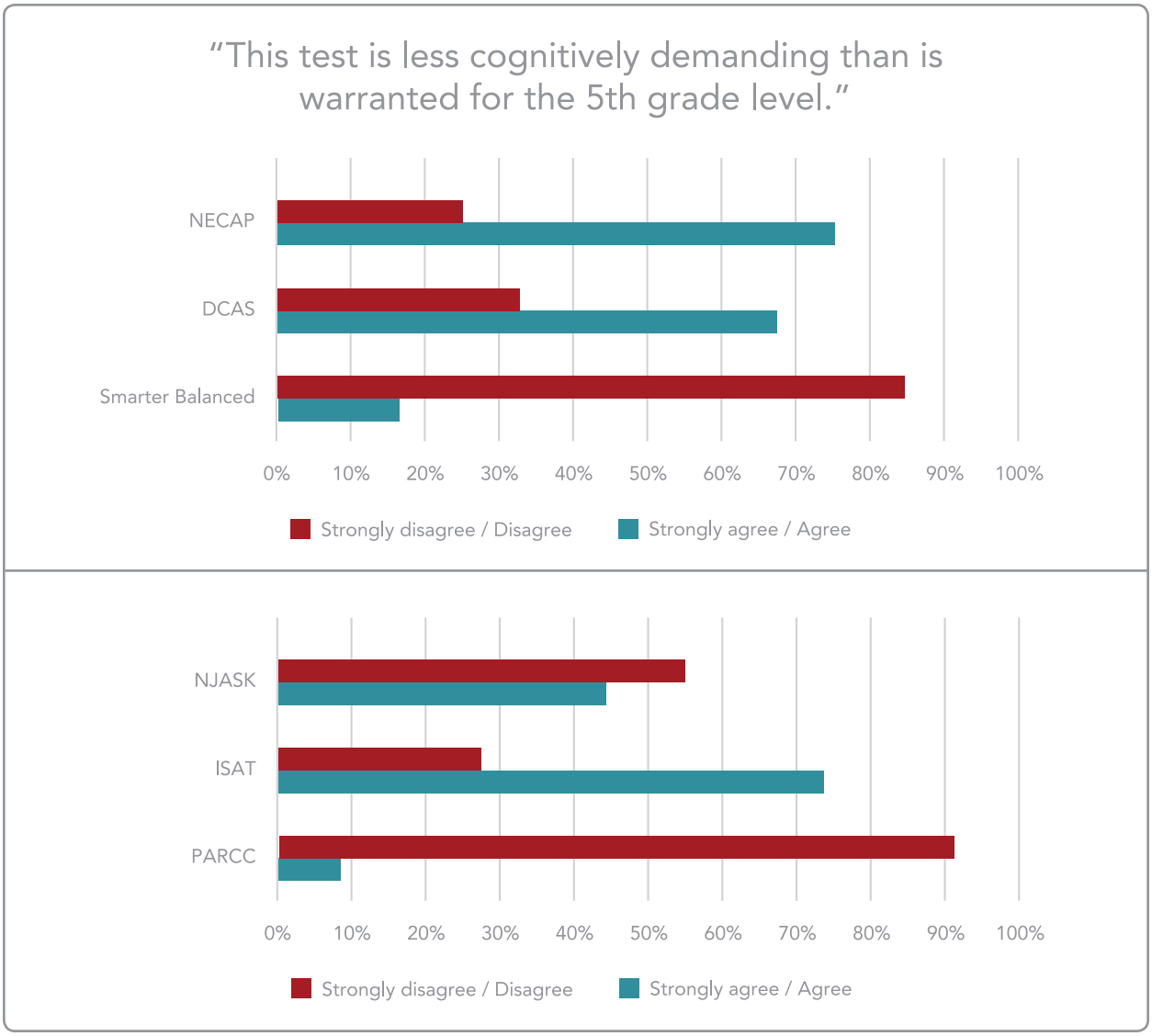


Figure 24. Percent agreement with statement: "This test is less cognitively demanding than is warranted for the fifth grade level."

> *"Even though the content may be grade-level appropriate, the depth is not there on the former tests. Smarter Balanced does a better job at structuring items to reach a deep complexity. I think the range of content on the former tests not being grade-appropriate could be based on the content standards they were written to assess."*

As shown in Figure 25, 78% of teachers across the panels strongly agreed or agreed that "this test measures the most important knowledge and skills to be taught in an excellent fifth grade math/ELA classroom" in regards to the consortia tests, on average. In comparison, 44% of teachers across the panels strongly agreed or agreed with this statement in regards to the former assessments, on average. There was strong consensus that the two consortia assessments measured the most important knowledge and skills taught in fifth grade classrooms. The former state assessments had varied evaluations on this item, with NECAP receiving the strongest disagreement with the statement at 67%.

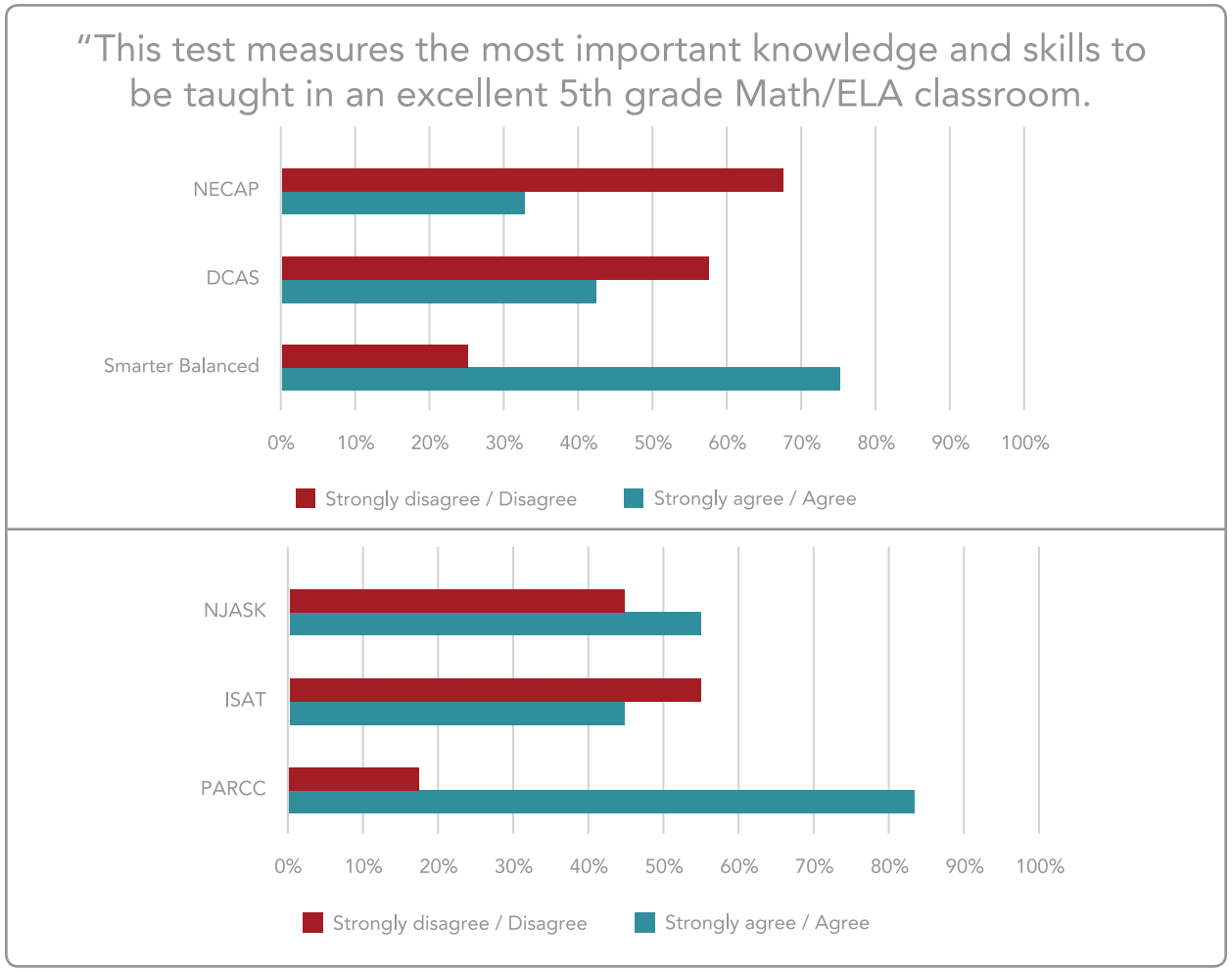


Figure 25. Percent agreement with statement: "This test measures the most important knowledge and skills to be taught in an excellent fifth grade Math/ELA classroom."

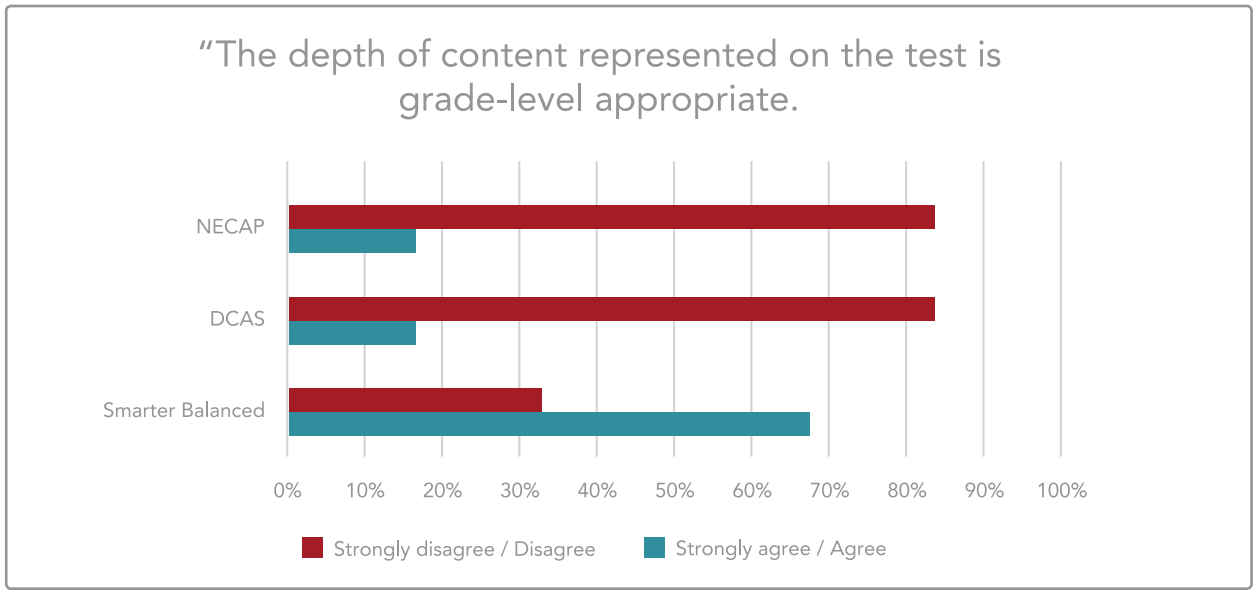


Figure 26 (A). Percent agreement with statement: "The depth of content represented on the test is grade-level appropriate."

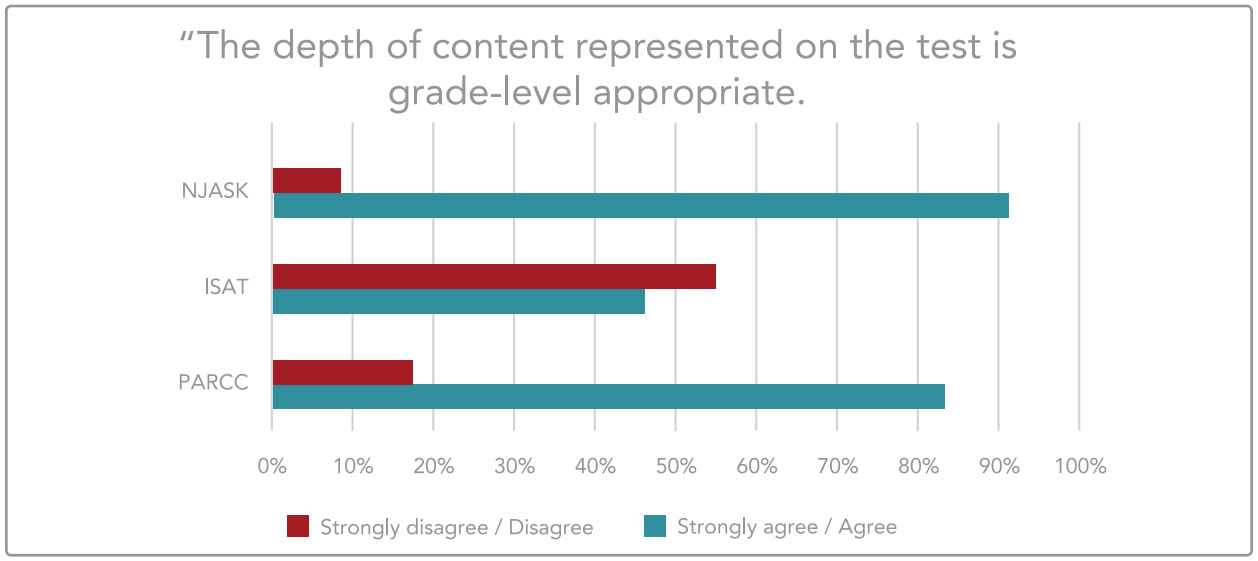


Figure 26 (B). Percent agreement with statement: "The depth of content represented on the test is grade-level appropriate."

On average, 74% of the teachers across both panels agreed or strongly agreed that the depth of content represented on the new consortia tests is grade-level appropriate. Approximately 83% of the teachers typically agreed or strongly agreed that the range[5] of content represented on the new tests is grade- level appropriate. Fewer, but over one-half (53%) of the teachers generally agreed that the former tests represented appropriate range, but not appropriate depth (42%) for fifth grade instruction. As has often been the case, the overall averages mask a great deal of individual variation in how the individual former state assessments were viewed. This is especially true of the NJASK, to which this statement was agreed upon by 91% of the panelists for both the range and depth items; somewhat so for the ISAT, with 73% agreement on range but only 45% agreement on depth.

One aspect of grade-appropriateness might be found in how engaging students find the test items. We asked teachers to evaluate the likelihood of authentic student engagement with the items for each test. The results are presented in Figure 27. Across the two panels, 61% of the teachers tended to strongly agree or agree that the consortia test items would authentically engage fifth grade students, with PARCC being rated somewhat higher than Smarter Balanced.

For the former assessments, 29% of the teachers tended to strongly agree or agree that students would be authentically engaged by the items. NECAP and DCAS were rated much lower than ISAT and NJASK. Recall that one of the criticisms of all the tests was that the content was repetitive and lacked diversity. So while the content was rather repetitive, teachers thought the items on the consortia assessments were more interesting and more likely to authentically engage students. Generally, the panelists were realistic about the moderate extent to which assessments can be expected to engage students.

Teacher 1: *"However, I have yet to come across a student who is "authentically" engaged in a testing situation aside from they know they have to take the assessment and their scores will be used to determine a number of things in the school career from their classroom level to their placement in future classes to the ratings of their teachers. I DO think the listening portions of [Consortium test] were awesome. The feedback I received from my students after our testing session last year was that those were a welcomed change and did engage them."*

Teacher 2: *"I do not know if this is possible in any given testing situation. It is not an authentic learning environment."*

Teacher 3: *"Unless the content is extremely novel and relevant, most students aren't looking to "bond" with a test. When you say "authentically engaged" I don't think that this is a goal you can attain in a test item. You can interest kids, but because this is a test and a required one, engagement is forced. I haven't yet met a child who is excited or enthused about going through the test experiences. "*

[5]See Figure 7 for data on range of content.

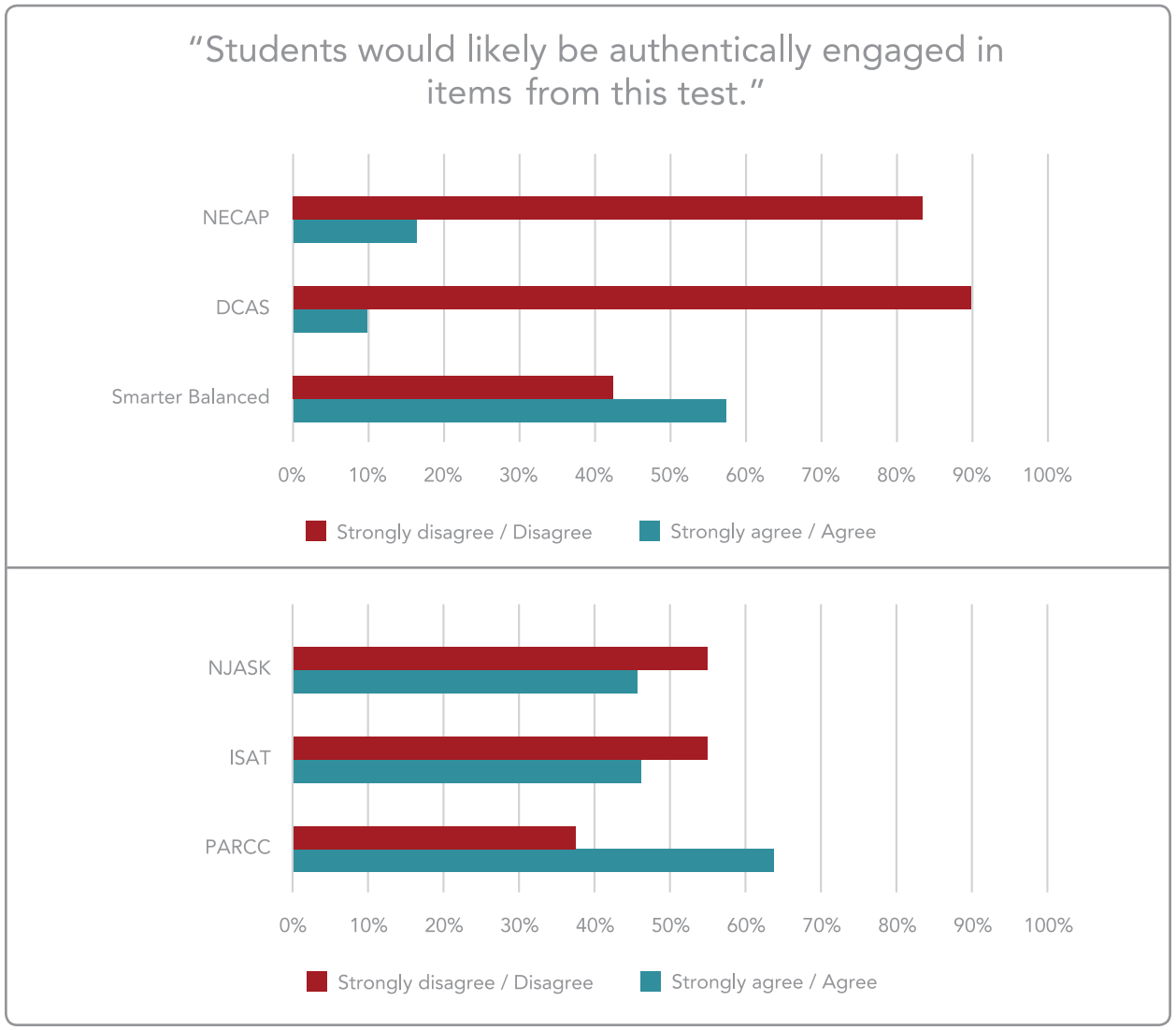


Figure 27. Average percent agreement with statement: "Students would likely be authentically engaged in items from this test."

Consistently throughout these items, the consortia assessments 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 fifth 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 change in mean scores were for the statements:

- "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." Teachers generally disagreed with this statement across the two panels. Mean scores ranged from 1.7 to 2.0 on a scale of 1 to 4. However, fewer teachers on the PARCC panel agreed and strongly agreed with this statement after they evaluated the assessments than before they evaluated the assessments.

- "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." Teachers generally agreed or strongly agreed with this statement across the two panels. Mean scores ranged from 3.2 to 3.5. More teachers on the Smarter Balanced panel agreed or strongly agreed with this statement after they evaluated the assessments than before they evaluated the assessments.

Notably, within the PARCC panel, there was no change in teachers' endorsement of this statement: "Selected-response items can be used to measure complex thinking skills." On average, more teachers disagreed than agreed with this statement before and after their evaluation of the PARCC panel assessments. A panelist from the PARCC panel noted the following:

> *"There wasn't a lot of constructed responses, not a lot of open-ended [responses on all three assessments]. There was so many different things I was looking for, I just thought there would be more. [I thought] there would be more of those selective responses where there was more than one option. Even that would have changed kind of the dynamic that was there."*

Table 1. Average Attitudes toward Tests Results for PARCC and Smarter Balanced Panels

Pre_Mean (1 to 4)		Attitudes toward Tests items	Post_Mean (1 to 4)		Pre-Post Difference	
PARCC Panel	Smarter Balanced Panel		PARCC Panel	Smarter Balanced Panel	PARCC Panel	Smarter Balanced Panel
1.9	1.8	I prefer tests that are comprised mostly of selected-response items	1.8	1.8	-0.1	0.1
2.0	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.	1.7	1.9	-0.3	0.2
2.8	3.1	I prefer tests that are comprised mostly of constructed-response or performance-based items.	2.9	2.9	0.1	-0.2
2.8	2.9	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	3.0	-0.1	0.1
3.5	3.3	I prefer tests with some selected-response and some constructed-response items.	3.4	3.4	-0.1	0.1

3.4	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.**	3.5	3.5	0.1	0.3
3.1	3.0	**Selected-response tests are simply easier to administer than constructed-response or performance-based tests.**	3.2	2.9	0.1	-0.1
2.7	2.7	**Selected-response items can be used to measure complex thinking skills.**	2.7	2.8	0.0	0.1

Note: Attitudes Toward Tests were measured on a scale of 1 to 4, where 1 = strongly disagree and 4 = strongly agree.

Recommendations for Continuous Improvement

The findings from our study suggest that the new consortia assessments have greater efficacy for reflecting high-quality teaching and learning than the former state assessments. They represent the breadth and depth of content in excellent fifth grade classrooms appreciably better than former assessments. Nonetheless, the study also reflected ways these teachers believe that these tests can be improved upon.

The most common concern from participating teachers related to whether all students – particularly low performing students – are prepared to do well on these new, more rigorous assessments at this moment in time. Participating teachers did not believe that the new assessments represented an insurmountable bar, but simply believed that students would need quality instruction and effective preparation to be able to do well on the new tests – and that can take time. Teachers did believe that the consortia assessments represent the right direction for teaching and learning, but emphasized that they do implicate a higher bar for students that not all students are currently prepared to reach.

Some of the teachers noted during the follow-up conversations that all of the tests tended to be overly repetitive and narrow in focus. That is, they asked the same kinds of questions frequently, to the point of predictability. The following excerpts from the focus-group discussion illustrate teachers' thoughts on this topic:

- *For instance on the [consortium] test, I felt like there was a huge amount of multi-digit multiplication. Over and over, more questions of multi-digit multiplication were being asked. And I understand with statistics, you have to ask two or three questions, or maybe four questions to get a valid prediction of that child's ability to multiply, but you don't have to ask ten multi-digit multiplication questions to get an idea of what they know and can understand, and I feel like there was so much more of common core math, fifth-grade common core math that was absent from those tests.*

- *When you asked that question about knowledge and skill, that's the problem. You put both of them together, knowledge and skill. None of those tests really did an excellent job of covering both of those, knowledge and skill.*

However, the survey data also suggest that the new tests do a better job of representing the depth and range of content that is appropriate for fifth grade instruction than the former assessments. This taken with the teachers' thoughts above may indicate that although the new tests represent an improvement over former state assessments, there is still opportunity to go deeper and wider in their content.

Our conversations with teachers, especially those in Math, also showed a concern about not conflating literacy with content knowledge. For example, the consortium math tests require far more reading and a deeper understanding of context than the former assessments. These elements add to the complexity of the tasks students are asked to complete. And, though teachers understand that literacy is an important foundational skill, they also encouraged the assessment consortia to ensure that each section appropriately assesses the primary knowledge and skills it was designed to assess.

- *I really liked how you said that they need to be sort of test savvy, and being able to go from the question back into the passage to find what they were looking for, to answer the question, toggle back and forth from the passage to the question. There's a very similar thing in math where they really have to be able to use that equation editor, so when they want to make a fraction, they have to be able to choose the fraction maker, and then put the cursor in the numerator to put in their numerator, and then move the cursor to the denominator, and then move the cursor out of the fraction and then choose the operation side. So it's a very involved sort of thing of being test savvy as well as content savvy. That's a nice connection between the two.*

- *I think that we have to be very careful when these questions are created that we evaluate not only the reading level that's required, but also the mathematical context that's required, and is it grade appropriate or not.*

Concluding Thoughts

Through the insight and expertise of excellent teachers, we sought data and evidence, using five key questions to evaluate three claims we wanted excellent teachers to support or refute about the new state assessments:

1. **The new tests are better suited to supporting instruction than former tests.**
2. **The new assessments reflect great teaching.**
3. **The new tests are of higher quality and worth the transition.**

We compared the new assessments to a group of former state assessments, to ground this evaluation in the concrete actuality of where states had been and where they are now. The results were clear, and included messages from our best teachers about ways to adjust our course as we progress.

The findings from our study suggest that the consortia assessments indeed are better for teaching and learning than the former assessments. They improve representation of the breadth and depth of content in excellent fifth grade classrooms over former assessments. While the new tests were seen as challenging, they were seen as appropriate for the grade level. 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. Consortium tests do not assess outside the range of what fifth grade students are expected to know and do. In fact, they represent the shift toward better alignment between classroom instruction and standardized testing.

Participants suggested that there may need to be a wider range of items that help to differentiate academic accomplishment between low- and mid-performing students. Nevertheless, the new tests represent 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, district, states and jurisdictions, and in the nation as a whole.

Yes, excellent teachers do want and prefer these assessments. Several times during the panel discussions, the idea that "we aren't there yet, but we're on the right trajectory" was raised. This is promising and hopeful, from policy, teaching, and student achievement perspectives. Teachers' evaluation of the consortia tests, which were largely positive, helps to validate the investments that have been made in developing stronger standards for content and learning and, concomitantly, stronger assessments. Teachers acknowledged the challenges of transitioning to a new state assessment. At the same time, they want to pursue the promise the consortia tests embody. Despite the negative press and the misinformation shrouding the tests, it's important to keep in mind that many teachers really do believe they are of higher quality than the former state assessments. One teacher put it this way:

> *"I think the prejudice that I came into it with was different, it was more positive, because I live in a state where we opted out of PARCC, so we started in PARCC and then we decided that we weren't going to do PARCC, and that we were going to create our own test. Well my particular state is notorious for making tests a little less rigorous than what the standard is for everyone else in the nation, so I'm like yes, I'm going to get to see this great test and look at it, and just think about how to better give my students the instruction they need, other than what we've created in our state."*

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%20DCAS%20Summary%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%20PDF/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

NJASK

The NJASK was administered annually until New Jersey's adoption of the PARCC as its statewide assessment. The NJASK was designed to measure student achievement based on New Jersey's Core Curriculum Standards. The Grade 5 ELA test used in the study was a retired (non-operational) form administered in 2012 (NJDOE, 2013). The form comprised 3 ELA reading passages that included 10 selected-response items and 1 essay task for each passage. The writing section was excluded from the ELA review. The Grade 5 Math test, also from the retired 2012 form, comprised 33 selected-response items, 8 short-response items, and 3 extended-response items.

ISAT

Similar to New Jersey, the ISAT was administered annually until the adoption of PARCC as the statewide assessment. The ISAT was designed to measure student achievement based on the Illinois Learning Standards (ISBE, 1997). The Grade 5 ELA and Math items reviewed in the study were from the 2013 sample book. The sample book contains items that are representative of items used on operational test forms. ELA comprised 1 short reading passage with 4 accompanying selected-response items, 1 long reading passage with 10 accompanying selected-response items, and 1 essay task (no sample essay responses were included). The writing section was excluded from the ELA review. The Math test comprised 45 selected-response items, 2 short-response items and 1 extended-response item.

NECAP

The NECAP was administered in New Hampshire until the state's adoption of the Smarter Balanced assessment as its statewide test for grades 3 through 8. The NECAP was designed to assess student outcomes against achievement targets set for each grade level ("Guide to Using the 2013 NECAP Reports," 2014). The Grade 5 ELA test comprised a reading and writing section. The writing section was omitted from the study to maintain consistency between the two former state assessments included in the Smarter Balanced panel[6]. The ELA test, therefore, only contained the reading component. The ELA and Math items reviewed were among the 2013 released items – these are items that were previously administered, but are no longer considered operational. The items included on the released form were representative of the operational form. The ELA test contained 2 short reading passages with 4 accompanying selected-response items and 1 short-response item, each, and 2 additional selected- response items that focused on grammar and word use. The Math test contained 15 selected-response items and a reference sheet and tool kit at the back of the form for students to use.

DCAS

The DCAS was administered in Delaware as a computer-adaptive test from 2010 until the Smarter Balanced assessment was adopted as the statewide measure of achievement in the Spring of 2015 for Grades 3 through 8 and 11. A 2012 form of the Grade five ELA and Math items were used for the study (DDOE, 2012). The ELA section included 1 short reading passage and 8 long reading passages. Each passage had a range of 5-6 selected-response items; there were 50 total items. A writing section was not included. The Math section included 50 selected-response items.

[6] The DCAS made available for the study did not contain a writing section.

PARCC

The PARCC End-of-Year (EOY) assessment was used for the study. Participants were allowed to view the performance-based assessment (PBA), but did not evaluate the PBA for the study. The ELA section of the EOY assessment reviewed for this study included 3 reading passages. Two of the passages had 8 selected-response items, the other had 13 selected-response items. Two of the reading passages also had 1 drag-and-drop item each. The Math EOY assessment used for the study included 2 sections. One section required students to utilize a calculator (provided in the test interface). There were 11 selected-response items that included drag-and-drop item types and 9 short-response items that required students to calculate and enter the answer, and 2 graph items. The second section included 10 selected-response items and 12 short-response items, both similar item formats (i.e., drag-and-drop).

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 fifth grade. 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

Attitudes Toward Tests

The Attitudes Toward Tests survey was designed by the research team to capture teacher's 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 comprising 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 reviewer 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[7].				
2	Tests that are largely selected-response are more appropriate for the knowledge and skills embedded in my learning outcomes than constructed-response[8] or performance-based[9] 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.				

[7] Selected-response items – Items for which the test taker must select a response from a set of options (e.g., true/false, multiple-choice, drag and drop, matching).

[8] Constructed-response items – Items for which the test taker must develop or create an original response (e.g., fill-in-the-blank, paragraph, and essay).

[9] Performance-based items/assessments – Items for which the test taker must develop an original response in the context or conditions in which the knowledge and skills are actually applied (e.g. act in a play, dance, play a musical instrument, complete a laboratory experiment).

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. 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?

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 fifth grade students			

5	Cognitive demand for mid-performing fifth grade students			
6	Cognitive demand for high-performing fifth grade students			
7	Items that require fifth grade students to demonstrate basic knowledge of concepts.			
8	Items that surface information about fifth grade student performance at the lower ability levels that would be useful to inform my instructional strategies.			
9	Items that low-performing fifth grade students would be expected to get right.			
10	Items that low-performing fifth grade students would be expected to get wrong.			
11	Items that surface information about fifth grade student performance at the middle ability levels that would be useful to inform my instructional strategies.			
12	Items that mid-performing fifth grade students would be expected to get right.			
13	Items the mid-performing fifth grade students would be expected to get wrong.			
14	Items that surface information about fifth grade student performance at the high ability levels that would be useful to inform my instructional strategies.			
15	Items that high-performing fifth grade students would be expected to get right.			
16	Items that high-performing fifth 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 fifth grade students.			

18	Number of items that require complex thinking skills needed to distinguish high-performing from mid-performing fifth grade ..students.			
19	The number of items that are above fifth grade-level.			
20	The number of items that are below fifth 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.			
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 fifth grade level.			
9	This test is less cognitively demanding than is warranted for the fifth grade level.			
10	Items on this test are consistent with what excellent fifth 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[10]. This test meets that criterion.			
13	This test measures an appropriately broad sampling of the ELA/ Math knowledge and skills in instruction an excellent fifth grade classroom.			
14	Excellent fifth 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 fifth grade Math/ELA classroom.			
16	This test measures the learning outcomes that I would set for student learning in fifth grade classes.			

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

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 fifth grade classes than the test does.			
20	The distribution of content on the test is representative of excellent instruction at the fifth 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[8]. This test meets that criterion.			
24	If I backwards-mapped a fifth 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 fifth grade classroom.			
26	The optimal formative assessments that I would give to fifth 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.			

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

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 section, 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

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

Figure C1: Gender

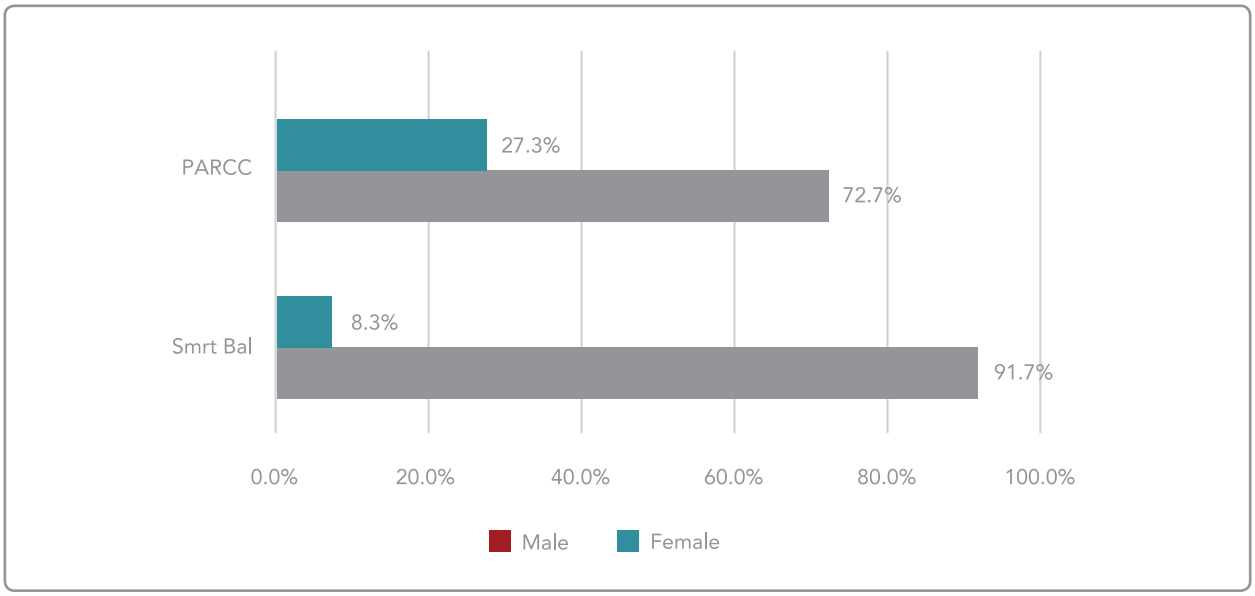


Figure C2: Race/Ethnicity

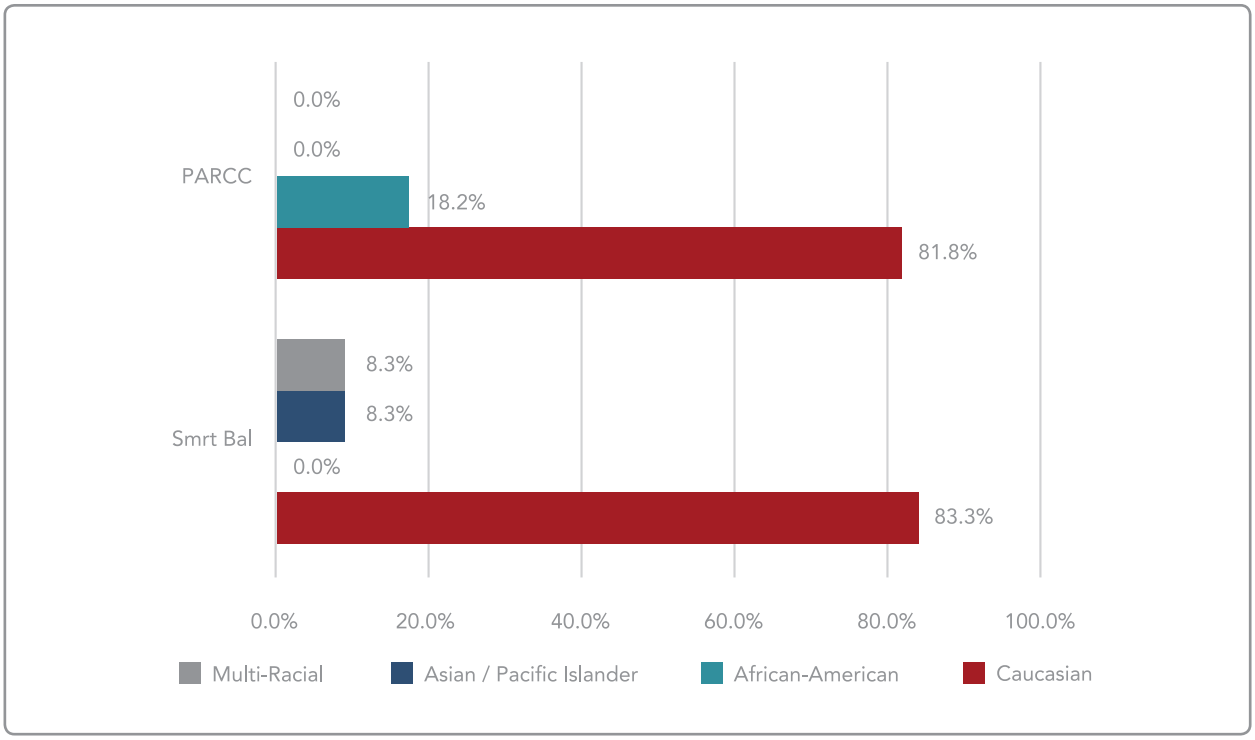


Figure C3: Years of Teaching Experience—PARCC panel (left) and Smarter Balanced panel (right)

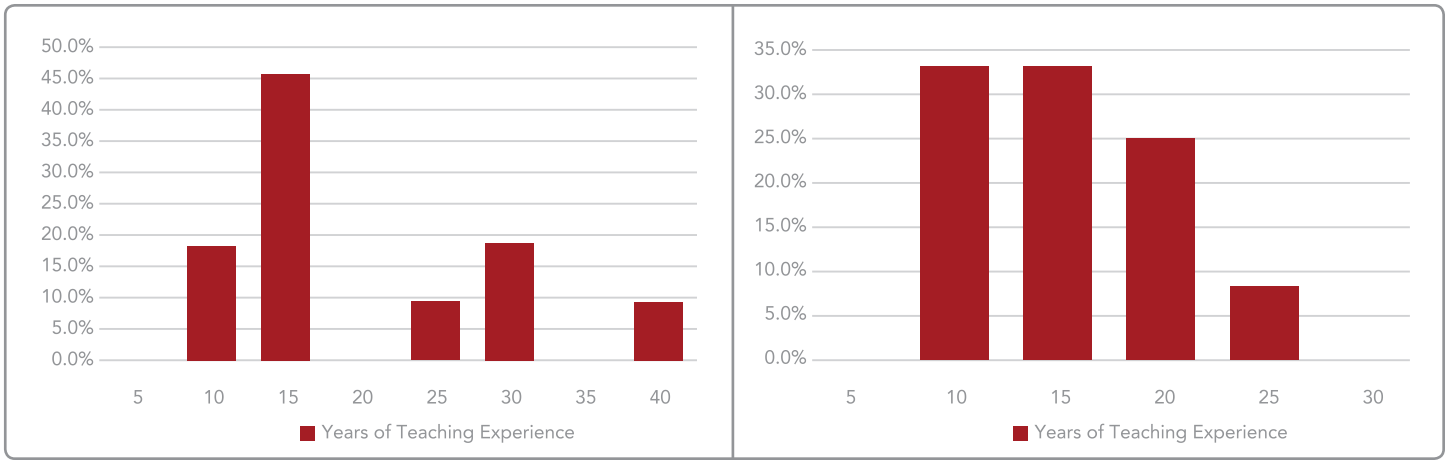
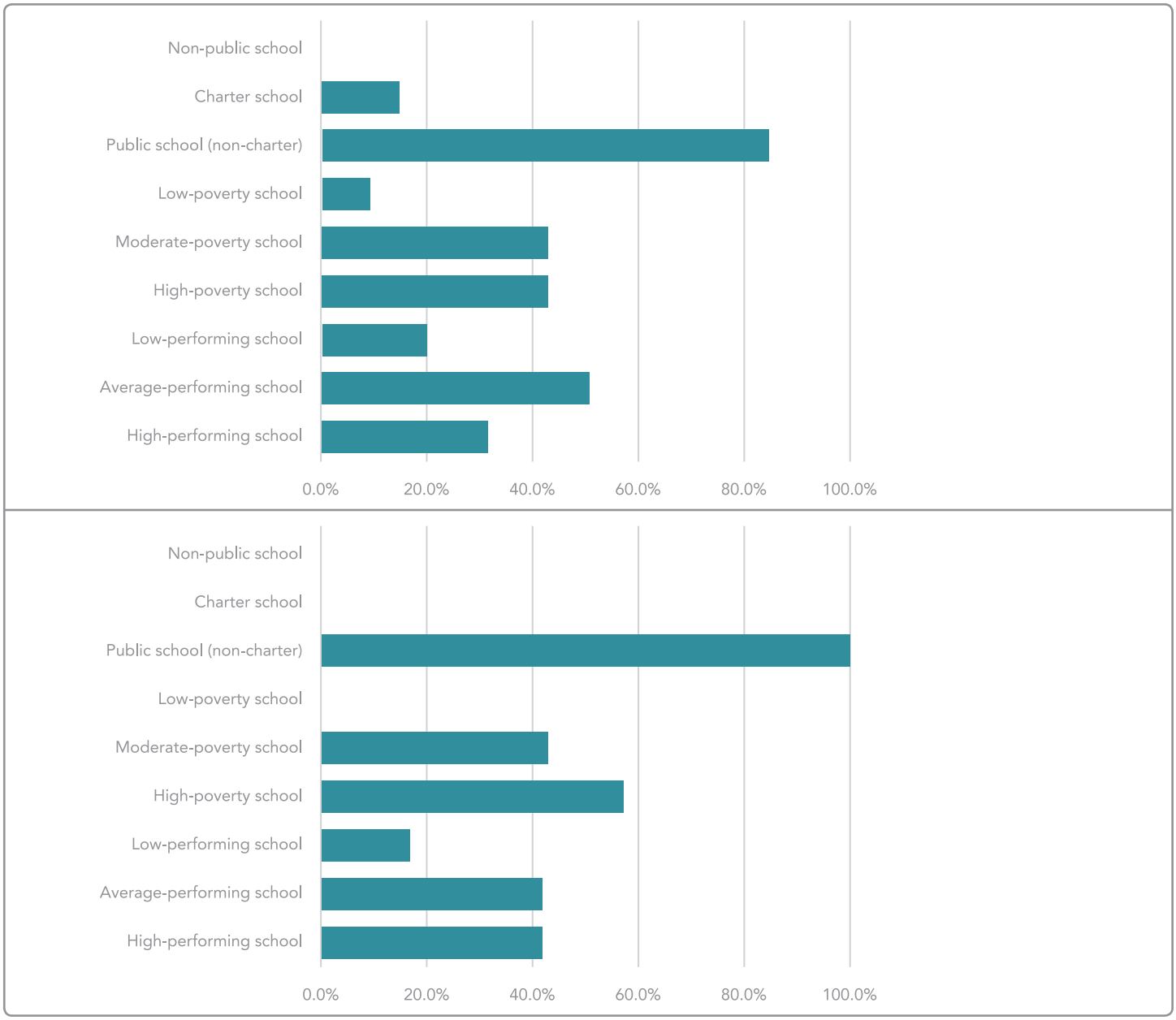


Figure C4: Teaching contexts—PARCC panel (top) and Smarter Balanced panel (bottom)



Appendix D: Guiding Questions for Panel Discussions

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 varied by panel, depending on the response patterns in the survey data. The PARCC panel's survey data generated the following prompts:

1. A number of you disagreed with the statement that these tests, all three, measures an appropriately broad sampling of the ELA/Math knowledge and skills in instruction in an excellent fifth grade classroom. What knowledge and skills were missing, under-represented, or over- represented in each of these tests?

2. The results of the survey indicate the PARCC test has enough items that will surface information about fifth grade high and mid-performing students that would inform your instruction, but there aren't enough items that will surface information about fifth grade low-performing students. If you were asked to redesign this test, how would you fix this problem?

    a. If they say "add items," then ask them to consider other alternatives that wouldn't require additional testing time.

The Smarter Balanced panel's survey data generated the following prompts:

1. How did you all interpret the statement "students are required to integrate knowledge from different domains?"

    a. A number of you thought the DE and NECAP assessments did not require students to integrate knowledge from different "domains" (concepts). However, you did agree that Smarter Balanced does require integration of knowledge from different domains. Can you draw some examples from the Smarter Balanced and talk about specific ways in which it requires students to integrate knowledge from different concepts?

2. The results of the survey indicate the Smarter Balanced test has enough items that will surface information about fifth grade high-performing students that would inform your instruction, but there aren't enough items that will surface information about fifth grade low- and mid- performing students. If you were asked to redesign this test, how would you fix this problem?

    a. If they say "add items," then ask them to consider other alternatives that wouldn't require additional testing time.

3. A number of you disagreed that the distribution of content on the NECAP was representative of excellent fifth grade instruction. Can one or more of you who responded in this way talk the ways in which the distribution of content missed the mark?

4. Several of you thought the Smarter Balanced did well in providing information for mid- and high- performers. You also thought the Smarter Balanced did not provide enough items to discriminate between mid- and high-performers. These are seemingly conflicting statements. Are they? How could the test be more discriminating between these groups?

Appendix E: Survey of Assessment Quality Items

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 PARCC 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; PARCC, Illinois, and New Jersey assessments

"Amount" items	PARCC				ISAT				NJASK			
	Mean Score (1 to 3)	Less than Needed	Enough/ About right	More than Needed	Less than Needed	Enough/ About right	More than Needed	Mean Score (1 to 3)	Less than Needed	Enough/ About right	More than Needed	Mean Score (1 to 3)
Items that require recall, such as identification, labeling, calculating, defining, and reciting.	1.7	36.40%	54.50%	9.10%	0.00%	36.40%	63.60%	2.6	0.00%	63.60%	36.40%	2.4
Items that require application of skills, such as graphing, categorizing, organizing, predicting, and estimating.	1.9	9.10%	90.90%	0.00%	45.50%	45.50%	9.10%	1.6	36.40%	54.50%	9.10%	1.7
Items that require students to demonstrate strategic and extended thinking skills, such as investigation, analysis, and design.	1.7	36.40%	54.50%	9.10%	100.00%	0.00%	0.00%	1	81.80%	18.20%	0.00%	1.2
Cognitive demand for low-performing 5th grade students	2.6	0.00%	36.40%	63.60%	9.10%	81.80%	9.10%	2	0.00%	72.70%	27.30%	2.3
Cognitive demand for mid-performing 5th grade students	2.4	0.00%	63.60%	36.40%	54.50%	45.50%	0.00%	1.5	27.30%	72.70%	0.00%	1.7
Cognitive demand for high-performing 5th grade students	1.9	9.10%	90.90%	0.00%	100.00%	0.00%	0.00%	1	72.70%	27.30%	0.00%	1.3
Items that require 5th grade students to demonstrate basic knowledge of concepts.	1.7	36.40%	54.50%	9.10%	0.00%	54.50%	45.50%	2.5	0.00%	45.50%	54.50%	2.5
Items that surface information about 5th grade student performance at the lower ability levels to inform my instructional strategies.	1.7	45.50%	36.40%	18.20%	18.20%	81.80%	0.00%	1.8	9.10%	72.70%	18.20%	2.1
Items that low-performing 5th grade students would be expected to get right.	1.4	63.60%	36.40%	0.00%	27.30%	45.50%	27.30%	2	9.10%	72.70%	18.20%	2.1

Items that low-performing 5th grade students would be expected to get wrong.	2.5	0.00%	54.50%	45.50%	9.10%	81.80%	9.10%	2	9.10%	90.90%	0.00%	1.9
Items that surface information about 5th grade student performance at the middle ability levels to inform my instructional strategies.	2	9.10%	81.80%	9.10%	18.20%	81.80%	0.00%	1.8	27.30%	63.60%	9.10%	1.8
Items that mid-performing 5th grade students would be expected to get right.	2	9.10%	81.80%	9.10%	9.10%	54.50%	36.40%	2.3	9.10%	72.70%	18.20%	2.1
Items the mid-performing 5th grade students would be expected to get wrong.	2.2	0.00%	81.80%	18.20%	45.50%	54.50%	0.00%	1.5	18.20%	72.70%	9.10%	1.9
Items that surface information about 5th grade student performance at the high ability levels to inform my instructional strategies.	2	0.00%	100.00%	0.00%	90.90%	9.10%	0.00%	1.1	81.80%	18.20%	0.00%	1.2
Items that high-performing 5th grade students would be expected to get right.	2	9.10%	81.80%	9.10%	0.00%	0.00%	100.00%	3	0.00%	27.30%	72.70%	2.7
Items that high-performing 5th grade students would be expected to get wrong.	1.9	18.20%	72.70%	9.10%	100.00%	0.00%	0.00%	1	72.70%	27.30%	0.00%	1.3
Number of items that require application of skills needed to distinguish mid-performing from low-performing 5th grade students.	1.8	27.30%	63.60%	9.10%	18.20%	81.80%	0.00%	1.8	36.40%	63.60%	0.00%	1.6
Number of items that require complex thinking skills needed to distinguish high-performing from mid-performing 5th grade students.	1.7	27.30%	72.70%	0.00%	100.00%	0.00%	0.00%	1	81.80%	18.20%	0.00%	1.2
The number of items that are above 5th grade-level.	2.3	9.10%	54.50%	36.40%	54.50%	45.50%	0.00%	1.5	45.50%	54.50%	0.00%	1.5
The number of items that are below 5th grade-level.	1.6	36.40%	63.60%	0.00%	0.00%	54.50%	45.50%	2.5	18.20%	54.50%	27.30%	2.1
Items that are likely to authentically engage student interest.	1.8	18.20%	81.80%	0.00%	54.50%	45.50%	0.00%	1.5	54.50%	45.50%	0.00%	1.5

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 PARCC 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; PARCC, Illinois, and New Jersey assessments

"Agree" Items	PARCC					ISAT					NJASK				
	Mean Score (1 to 4)	Strongly agree	Agree	Disagree	Strongly Disagree	Strongly agree	Agree	Disagree	Strongly Disagree	Mean Score (1 to 4)	Strongly agree	Agree	Disagree	Strongly Disagree	Mean Score (1 to 4)
Students are required to integrate a variety of knowledge and skills from a single domain.	3.3	27.30%	72.70%	0.00%	0.00%	0.00%	63.60%	18.20%	18.20%	2.5	18.20%	63.60%	9.10%	9.10%	2.9
Students are required to transfer knowledge from different domains.	3.5	54.50%	45.50%	0.00%	0.00%	0.00%	36.40%	45.50%	18.20%	2.2	27.30%	36.40%	27.30%	9.10%	2.8
Students are required to integrate a variety of knowledge and skills from different domains.	3.5	54.50%	45.50%	0.00%	0.00%	0.00%	36.40%	54.50%	9.10%	2.3	18.20%	45.50%	27.30%	9.10%	2.7
This test provides sufficient opportunity to evaluate students' ability to communicate in writing.	2.4	18.20%	18.20%	45.50%	18.20%	9.10%	18.20%	27.30%	45.50%	1.9	18.20%	36.40%	18.20%	27.30%	2.5
This test provides sufficient opportunity to evaluate students' ability to show their reasoning when solving a problem or arguing a case.	3	36.40%	36.40%	18.20%	9.10%	9.10%	18.20%	27.30%	45.50%	1.9	18.20%	36.40%	18.20%	27.30%	2.5
This test strikes a balance between the number of items that require recall responses and responses that require higher-level cognitive skills.	2.5	9.10%	45.50%	36.40%	9.10%	0.00%	18.20%	36.40%	45.50%	1.7	18.20%	27.30%	27.30%	27.30%	2.4
Students are required to demonstrate complex thinking skills, such as experimentation, analysis, and synthesis.	2.8	36.40%	36.40%	0.00%	27.30%	0.00%	18.20%	27.30%	54.50%	1.6	0.00%	54.50%	9.10%	36.40%	2.2
This test is more cognitively demanding than is warranted for the 5th grade level.	2.5	27.30%	9.10%	54.50%	9.10%	0.00%	0.00%	45.50%	54.50%	1.5	0.00%	0.00%	63.60%	36.40%	1.6
This test is less cognitively demanding than is warranted for the 5th grade level.	1.6	0.00%	9.10%	45.50%	45.50%	36.40%	36.40%	27.30%	0.00%	3.1	9.10%	36.40%	54.50%	0.00%	2.5
Items on this test are consistent with what excellent 5th grade Math/ELA teachers ask their students to know and do.	3.3	45.50%	45.50%	0.00%	9.10%	0.00%	54.50%	27.30%	18.20%	2.4	9.10%	72.70%	9.10%	9.10%	2.8

Preparing students for this test would require meaningful lessons and learning, beyond skill and drill practice.	3.5	54.50%	45.50%	0.00%	0.00%	9.10%	27.30%	45.50%	18.20%	2.3	18.20%	63.60%	9.10%	9.10%	2.9
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.	3.3	36.40%	54.50%	9.10%	0.00%	0.00%	18.20%	63.60%	18.20%	2	9.10%	45.50%	36.40%	9.10%	2.5
This test measures an appropriately broad sampling of the ELA/Math knowledge and skills in instruction an excellent 5th grade classroom.	3	27.30%	45.50%	27.30%	0.00%	0.00%	27.30%	63.60%	9.10%	2.2	9.10%	63.60%	27.30%	0.00%	2.8
Excellent 5th grade instruction generally aligns with the content measured on this test.	3.3	27.30%	72.70%	0.00%	0.00%	0.00%	63.60%	36.40%	0.00%	2.6	18.20%	72.70%	9.10%	0.00%	3.1
This test measures the most important knowledge and skills to be taught in an excellent 5th grade Math/ELA classroom.	3.1	27.30%	54.50%	18.20%	0.00%	0.00%	45.50%	45.50%	9.10%	2.4	18.20%	36.40%	45.50%	0.00%	2.7
This test measures the learning outcomes that I would set for student learning in 5th grade classes.	3.1	18.20%	72.70%	9.10%	0.00%	0.00%	45.50%	45.50%	9.10%	2.4	18.20%	45.50%	36.40%	0.00%	2.8
Certain item types are emphasized more heavily on the test than is warranted for the grade level.	2.5	18.20%	9.10%	72.70%	0.00%	27.30%	27.30%	36.40%	9.10%	2.7	18.20%	18.20%	63.60%	0.00%	2.5
Certain content areas are emphasized more heavily on the test than is warranted for the grade level.	2.5	18.20%	9.10%	72.70%	0.00%	0.00%	9.10%	90.90%	0.00%	2.1	9.10%	9.10%	81.80%	0.00%	2.3
I would give more emphasis to certain content areas in 5th grade classes than the test does.	2.4	9.10%	18.20%	72.70%	0.00%	27.30%	36.40%	36.40%	0.00%	2.9	9.10%	45.50%	45.50%	0.00%	2.6
The distribution of content on the test is representative of excellent instruction at the 5th grade level.	2.9	27.30%	45.50%	18.20%	9.10%	0.00%	27.30%	72.70%	0.00%	2.3	9.10%	45.50%	45.50%	0.00%	2.6

The depth of content represented on the test is grade-level appropriate.	2.9	9.10%	72.70%	18.20%	0.00%	0.00%	45.50%	18.20%	36.40%	2.1	9.10%	81.80%	0.00%	9.10%	2.9
The range of content represented on the test is grade-level appropriate.	2.8	0.00%	81.80%	18.20%	0.00%	0.00%	72.70%	27.30%	0.00%	2.7	9.10%	81.80%	9.10%	0.00%	3
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.	3.1	27.30%	54.50%	18.20%	0.00%	9.10%	18.20%	45.50%	27.30%	2.1	0.00%	36.40%	54.50%	9.10%	2.3
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.	3.3	36.40%	54.50%	9.10%	0.00%	0.00%	45.50%	36.40%	18.20%	2.3	18.20%	45.50%	27.30%	9.10%	2.7
I would like to use formative assessments built using items from this test in a 5th grade classroom.	3.3	27.30%	72.70%	0.00%	0.00%	9.10%	45.50%	45.50%	0.00%	2.6	18.20%	54.50%	27.30%	0.00%	2.9
The optimal formative assessments that I would give to 5th grade students measure concepts not addressed by this test.	2	0.00%	18.20%	63.60%	18.20%	18.20%	54.50%	18.20%	9.10%	2.8	18.20%	36.40%	36.40%	9.10%	2.6
If used for formative assessment, items on this test would help me make decisions about instruction.	3.4	36.40%	63.60%	0.00%	0.00%	0.00%	72.70%	27.30%	0.00%	2.7	27.30%	63.60%	9.10%	0.00%	3.2
Student results from this test would give me valuable information about how students are learning.	3.2	18.20%	81.80%	0.00%	0.00%	0.00%	72.70%	27.30%	0.00%	2.7	18.20%	54.50%	27.30%	0.00%	2.9
The item types on this test are aligned with the skills they appear to be designed to measure.	3.3	27.30%	72.70%	0.00%	0.00%	9.10%	54.50%	27.30%	9.10%	2.6	18.20%	63.60%	0.00%	18.20%	2.8
This test provides a satisfactory balance between selected-response items and constructed response/performance-based items.	2.6	9.10%	63.60%	9.10%	18.20%	0.00%	0.00%	54.50%	45.50%	1.5	0.00%	36.40%	45.50%	18.20%	2.2

Low-performing students would find it easy to get most of the items on this test correct.	1.4	0.00%	0.00%	36.40%	63.60%	0.00%	54.50%	45.50%	0.00%	2.5	9.10%	18.20%	54.50%	18.20%	2.2
Mid-performing students would find it easy to get most of the items on this test correct.	2	0.00%	27.30%	45.50%	27.30%	18.20%	81.80%	0.00%	0.00%	3.2	36.40%	18.20%	45.50%	0.00%	2.9
High-performing students would find it easy to get most of the items on this test correct.	2.6	9.10%	45.50%	45.50%	0.00%	72.70%	27.30%	0.00%	0.00%	3.7	45.50%	45.50%	9.10%	0.00%	3.4
Low-performing students would generally perform well on this test.	1.5	0.00%	9.10%	36.40%	54.50%	0.00%	63.60%	36.40%	0.00%	2.6	0.00%	27.30%	63.60%	9.10%	2.2
Mid-performing students would generally perform well on this test.	2.5	9.10%	36.40%	54.50%	0.00%	45.50%	54.50%	0.00%	0.00%	3.5	36.40%	54.50%	9.10%	0.00%	3.3
High-performing students would generally perform well on this test.	3.1	9.10%	90.90%	0.00%	0.00%	72.70%	27.30%	0.00%	0.00%	3.7	63.60%	36.40%	0.00%	0.00%	3.6
Students would likely be authentically engaged in items from this test.	2.7	9.10%	54.50%	36.40%	0.00%	9.10%	36.40%	45.50%	9.10%	2.5	0.00%	45.50%	54.50%	0.00%	2.5

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 E3 for the Smarter Balanced 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 E3 for each assessment as well.

Table E3: "Amount" Items; Smarter Balanced, Delaware, and New Hampshire assessments

"Amount" items	Smarter Balanced				DCAS				NECAP			
	Mean Score (1 to	Less than Needed	Enough/ About right	More than Needed	Less than Needed	Enough/ About right	More than Needed	Mean Score (1 to	Less than Needed	Enough/ About right	More than Needed	Mean Score (1 to 3)
Items that require recall, such as identification, labeling, calculating, defining, and reciting.	2.1	8.3%	75.0%	16.7%	0.0%	50.0%	50.0%	2.5	8.3%	41.7%	50.0%	2.4
Items that require application of skills, such as graphing, categorizing, organizing, predicting, and estimating.	2.3	0.0%	75.0%	25.0%	33.3%	41.7%	25.0%	1.9	33.3%	50.0%	16.7%	1.8
Items that require students to demonstrate strategic and extended thinking skills, such as investigation, analysis, and design.	1.8	33.3%	58.3%	8.3%	91.7%	8.3%	0.0%	1.1	83.3%	16.7%	0.0%	1.2
Cognitive demand for low-performing 5th grade students	2.7	0.0%	33.3%	66.7%	16.7%	75.0%	8.3%	1.9	33.3%	41.7%	25.0%	1.9
Cognitive demand for mid-performing 5th grade students	2.3	8.3%	50.0%	41.7%	66.7%	16.7%	16.7%	1.5	66.7%	16.7%	16.7%	1.5
Cognitive demand for high-performing 5th grade students	2.0	16.7%	66.7%	16.7%	83.3%	16.7%	0.0%	1.2	83.3%	16.7%	0.0%	1.2
Items that require 5th grade students to demonstrate basic knowledge of concepts.	2.1	16.7%	58.3%	25.0%	16.7%	41.7%	41.7%	2.3	8.3%	33.3%	58.3%	2.5
Items that surface information about 5th grade student performance at the lower ability levels to inform my instructional strategies.	1.8	33.3%	50.0%	16.7%	41.7%	33.3%	25.0%	1.8	50.0%	41.7%	8.3%	1.6
Items that low-performing 5th grade students would be expected to get right.	1.7	50.0%	33.3%	16.7%	8.3%	50.0%	41.7%	2.3	0.0%	58.3%	41.7%	2.4
Items that low-performing 5th grade students would be expected to get wrong.	2.4	16.7%	25.0%	58.3%	41.7%	50.0%	8.3%	1.7	50.0%	50.0%	0.0%	1.5
Items that surface information about 5th grade student performance at the middle ability levels to inform my instructional strategies.	1.9	16.7%	75.0%	8.3%	58.3%	16.7%	25.0%	1.7	66.7%	25.0%	8.3%	1.4
Items that mid-performing 5th grade students would be expected to get right.	1.9	25.0%	58.3%	16.7%	16.7%	25.0%	58.3%	2.4	0.0%	50.0%	50.0%	2.5
Items the mid-performing 5th grade students would be expected to get wrong.	2.1	16.7%	58.3%	25.0%	58.3%	33.3%	8.3%	1.5	50.0%	50.0%	0.0%	1.5

Items that surface information about 5th grade student performance at the high ability levels to inform my instructional strategies.	1.6	41.7%	58.3%	0.0%	83.3%	16.7%	0.0%	1.2	83.3%	16.7%	0.0%	1.2
Items that high-performing 5th grade students would be expected to get right.	2.3	0.0%	66.7%	33.3%	33.3%	8.3%	58.3%	2.3	16.7%	25.0%	58.3%	2.4
Items that high-performing 5th grade students would be expected to get wrong.	1.9	25.0%	58.3%	16.7%	83.3%	16.7%	0.0%	1.2	58.3%	33.3%	8.3%	1.5
Number of items that require application of skills needed to distinguish mid-performing from low-performing 5th grade students.	1.8	33.3%	58.3%	8.3%	58.3%	33.3%	8.3%	1.5	58.3%	33.3%	8.3%	1.5
Number of items that require complex thinking skills needed to distinguish high-performing from mid-performing 5th grade students.	1.8	25.0%	66.7%	8.3%	83.3%	8.3%	8.3%	1.3	83.3%	16.7%	0.0%	1.2
The number of items that are above 5th grade-level.	2.0	9.1%	81.8%	9.1%	81.8%	18.2%	0.0%	1.2	63.6%	36.4%	0.0%	1.4
The number of items that are below 5th grade-level.	1.6	58.3%	25.0%	16.7%	8.3%	58.3%	33.3%	2.3	25.0%	50.0%	25.0%	2.0
Items that are likely to authentically engage student interest.	1.5	50.0%	50.0%	0.0%	83.3%	8.3%	8.3%	1.3	91.7%	8.3%	0.0%	1.1

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 E4 for the Smarter Balanced 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 F4 for each assessment as well.

Table E4: "Agree" Items; Smarter Balanced, Delaware, and New Hampshire assessments

"Agree" Items	Smarter Balanced					DCAS					NECAP				
	Mean Score (1 to	Strongly agree	Agree	Disagree	Strongly Disagree	Strongly agree	Agree	Disagree	Strongly Disagree	Mean Score (1 to 4)	Strongly agree	Agree	Disagree	Strongly Disagree	Mean Score (1 to 4)
Students are required to integrate a variety of knowledge and skills from a single domain.	3.3	33.3%	58.3%	8.3%	0.0%	0.0%	58.3%	25.0%	16.7%	2.4	8.3%	33.3%	50.0%	8.3%	2.4
Students are required to transfer knowledge from different domains.	3.3	33.3%	66.7%	0.0%	0.0%	8.3%	16.7%	50.0%	25.0%	2.1	0.0%	25.0%	66.7%	8.3%	2.2
Students are required to integrate a variety of knowledge and skills from different domains.	3.3	25.0%	75.0%	0.0%	0.0%	8.3%	16.7%	50.0%	25.0%	2.1	0.0%	16.7%	58.3%	25.0%	1.9
This test provides sufficient opportunity to evaluate students' ability to communicate in writing.	3.1	33.3%	41.7%	25.0%	0.0%	0.0%	0.0%	0.0%	100.0%	1.0	8.3%	41.7%	16.7%	33.3%	2.3
This test provides sufficient opportunity to evaluate students' ability to show their reasoning when solving a problem or arguing a case.	3.2	41.7%	33.3%	25.0%	0.0%	0.0%	0.0%	8.3%	91.7%	1.1	8.3%	33.3%	16.7%	41.7%	2.1
This test strikes a balance between the number of items that require recall responses and responses that require higher-level cognitive skills.	2.7	16.7%	50.0%	16.7%	16.7%	8.3%	8.3%	16.7%	66.7%	1.6	0.0%	16.7%	25.0%	58.3%	1.6
Students are required to demonstrate complex thinking skills, such as experimentation, analysis, and synthesis.	3.0	41.7%	33.3%	8.3%	16.7%	8.3%	0.0%	33.3%	58.3%	1.6	0.0%	25.0%	25.0%	50.0%	1.8
This test is more cognitively demanding than is warranted for the 5th grade level.	2.3	8.3%	16.7%	75.0%	0.0%	0.0%	16.7%	8.3%	75.0%	1.4	0.0%	0.0%	33.3%	66.7%	1.3
This test is less cognitively demanding than is warranted for the 5th grade level.	1.8	0.0%	16.7%	41.7%	41.7%	41.7%	25.0%	16.7%	16.7%	2.9	41.7%	33.3%	25.0%	0.0%	3.2
Items on this test are consistent with what excellent 5th grade Math/ELA teachers ask their students to know and do.	2.8	25.0%	25.0%	50.0%	0.0%	8.3%	33.3%	8.3%	50.0%	2.0	8.3%	41.7%	25.0%	25.0%	2.3
Preparing students for this test would require meaningful lessons and learning, beyond skill and drill practice.	3.3	58.3%	16.7%	25.0%	0.0%	8.3%	25.0%	25.0%	41.7%	2.0	0.0%	25.0%	50.0%	25.0%	2.0
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.	3.3	41.7%	50.0%	8.3%	0.0%	8.3%	16.7%	41.7%	33.3%	2.0	0.0%	16.7%	58.3%	25.0%	1.9
This test measures an appropriately broad sampling of the ELA/Math knowledge and skills in instruction an excellent 5th grade classroom.	2.8	16.7%	50.0%	33.3%	0.0%	0.0%	16.7%	41.7%	41.7%	1.8	0.0%	16.7%	58.3%	25.0%	1.9

Excellent 5th grade instruction generally aligns with the content measured on this test.	3.1	25.0%	58.3%	16.7%	0.0%		0.0%	41.7%	33.3%	25.0%	2.2		8.3%	25.0%	50.0%	16.7%	2.3
This test measures the most important knowledge and skills to be taught in an excellent 5th grade Math/ELA classroom.	3.0	25.0%	50.0%	25.0%	0.0%		0.0%	41.7%	33.3%	25.0%	2.2		0.0%	33.3%	50.0%	16.7%	2.2
This test measures the learning outcomes that I would set for student learning in 5th grade classes.	2.9	25.0%	41.7%	33.3%	0.0%		0.0%	41.7%	25.0%	33.3%	2.1		8.3%	25.0%	50.0%	16.7%	2.3
Certain item types are emphasized more heavily on the test than is warranted for the grade level.	2.3	0.0%	33.3%	66.7%	0.0%		58.3%	25.0%	16.7%	0.0%	3.4		33.3%	41.7%	25.0%	0.0%	3.1
Certain content areas are emphasized more heavily on the test than is warranted for the grade level.	2.3	0.0%	33.3%	66.7%	0.0%		16.7%	41.7%	41.7%	0.0%	2.8		8.3%	33.3%	50.0%	8.3%	2.4
I would give more emphasis to certain content areas in 5th grade classes than the test does.	2.8	25.0%	33.3%	41.7%	0.0%		25.0%	33.3%	41.7%	0.0%	2.8		33.3%	33.3%	33.3%	0.0%	3.0
The distribution of content on the test is representative of excellent instruction at the 5th grade level.	2.8	0.0%	75.0%	25.0%	0.0%		0.0%	33.3%	41.7%	25.0%	2.1		0.0%	33.3%	41.7%	25.0%	2.1
The depth of content represented on the test is grade-level appropriate.	2.7	0.0%	66.7%	33.3%	0.0%		8.3%	8.3%	41.7%	41.7%	1.8		0.0%	16.7%	58.3%	25.0%	1.9
The range of content represented on the test is grade-level appropriate.	2.8	0.0%	83.3%	16.7%	0.0%		0.0%	25.0%	50.0%	25.0%	2.0		0.0%	25.0%	50.0%	25.0%	2.0
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.	3.3	33.3%	66.7%	0.0%	0.0%		8.3%	16.7%	41.7%	33.3%	2.0		0.0%	41.7%	33.3%	25.0%	2.2
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.	3.3	58.3%	16.7%	16.7%	8.3%		16.7%	25.0%	16.7%	41.7%	2.2		8.3%	16.7%	50.0%	25.0%	2.1
I would like to use formative assessments built using items from this test in a 5th grade classroom.	3.3	41.7%	50.0%	0.0%	8.3%		8.3%	41.7%	33.3%	16.7%	2.4		8.3%	33.3%	33.3%	25.0%	2.3
The optimal formative assessments that I would give to 5th grade students measure concepts not addressed by this test.	2.2	8.3%	8.3%	75.0%	8.3%		25.0%	25.0%	41.7%	8.3%	2.7		25.0%	25.0%	41.7%	8.3%	2.7
If used for formative assessment, items on this test would help me make decisions about instruction.	3.3	33.3%	58.3%	8.3%	0.0%		8.3%	50.0%	33.3%	8.3%	2.6		0.0%	58.3%	33.3%	8.3%	2.5

Student results from this test would give me valuable information about how students are learning.	3.0	33.3%	33.3%	33.3%	0.0%	0.0%	41.7%	33.3%	25.0%	2.2	8.3%	16.7%	66.7%	8.3%	2.3
The item types on this test are aligned with the skills they appear to be designed to measure.	3.0	25.0%	58.3%	8.3%	8.3%	0.0%	75.0%	0.0%	25.0%	2.5	0.0%	58.3%	33.3%	8.3%	2.5
This test provides a satisfactory balance between selected-response items and constructed response/performance-based items.	3.0	25.0%	50.0%	25.0%	0.0%	0.0%	0.0%	0.0%	100.0%	1.0	8.3%	50.0%	25.0%	16.7%	2.5
Low-performing students would find it easy to get most of the items on this test correct.	1.4	8.3%	0.0%	16.7%	75.0%	0.0%	41.7%	41.7%	16.7%	2.3	0.0%	25.0%	58.3%	16.7%	2.1
Mid-performing students would find it easy to get most of the items on this test correct.	2.3	0.0%	25.0%	75.0%	0.0%	25.0%	50.0%	16.7%	8.3%	2.9	16.7%	66.7%	16.7%	0.0%	3.0
High-performing students would find it easy to get most of the items on this test correct.	2.8	8.3%	58.3%	33.3%	0.0%	66.7%	33.3%	0.0%	0.0%	3.7	66.7%	33.3%	0.0%	0.0%	3.7
Low-performing students would generally perform well on this test.	1.4	0.0%	0.0%	41.7%	58.3%	16.7%	41.7%	25.0%	16.7%	2.6	16.7%	16.7%	58.3%	8.3%	2.4
Mid-performing students would generally perform well on this test.	2.5	0.0%	50.0%	50.0%	0.0%	41.7%	50.0%	0.0%	8.3%	3.3	25.0%	58.3%	16.7%	0.0%	3.1
High-performing students would generally perform well on this test.	3.3	33.3%	58.3%	8.3%	0.0%	91.7%	8.3%	0.0%	0.0%	3.9	91.7%	8.3%	0.0%	0.0%	3.9
Students would likely be authentically engaged in items from this test.	2.6	8.3%	50.0%	33.3%	8.3%	0.0%	8.3%	50.0%	41.7%	1.7	0.0%	16.7%	50.0%	33.3%	1.8