#3 IN THE SERIES

Demystifying Statewide Standardized Assessments

Developing High-Quality Assessments and Items

By Michelle Croft, Hailly T.N. Korman, and Titilayo Tinubu Ali

APRIL 2023

ACCESS THE FULL SERIES

Overview

While K-12 students take many kinds of assessments (also called tests) for different purposes, the statewide standardized assessments used as part of annual federal accountability for schools often receive the greatest public scrutiny.

Statewide assessments measure what a student knows and can do. They are based on what schools are expected to teach for that grade level or content area within each state — and they can play a valuable role in improving education. Through comparable and consistent data, they allow decision-makers and educators to better understand how the education system is serving students, particularly historically marginalized students, including students of color, English learners, and students with disabilities. Test scores help educators identify students' strengths and areas of needed support to guide changes in instruction, inform large-scale instructional decisions (for example, identifying gaps in a district's curriculum), and measure how much a student has learned in a full academic year. These scores can also provide information about the effectiveness of instructional programs and other student supports to help state and district leaders and policymakers direct resources to schools and student populations.

Due to the complexity of assessment development and uncertainty around how scores are reported and can be used, statewide assessments can seem mysterious to some policymakers, educators, parents, students, and the general public. We developed six briefs to provide an overview of the test development process — from the initial stages of assessment design to the final process of scoring and reporting results — to help readers improve their understanding of how statewide assessments are developed and used. These briefs are organized by six topics:

1. What Statewide Assessments Are Designed to Measure
2. Ensuring That Assessments Accurately Measure Academic Standards
3. **Developing High-Quality Assessments and Items**
4. Ensuring Comparability Across Administrations
5. Making Assessments Accessible for Students With Disabilities and English Learners
6. Reporting Assessment Scores

WHAT ARE SUMMATIVE ASSESSMENTS?

Different tests have different purposes. State assessments are one type of summative assessment because they are administered to measure what a student has learned relative to what students should have been taught over the course of a school year. State summative assessments also can inform large-scale instructional decisions, such as identifying gaps in a district's curriculum, and measuring how much a student has learned in a full academic year.

While these briefs focus on statewide summative assessments, we also acknowledge the importance of other types of assessments that may be included in a state's assessment system, including formative and interim assessments.[1]

***Note**: Throughout the six briefs, we use the terms "test" and "assessment" interchangeably.*

Developing High-Quality Assessments and Items

Test developers have a thorough process for designing tests and items (commonly referred to as the test's questions). The process includes rigorous educator reviews through the development process to help improve the tests' and items' quality.

Key Takeaways

- Effective assessment development starts with a plan and blueprint.
- Educators who are experienced experts in their fields write the test items based on guidelines and instructions provided by test developers.
- Every test item undergoes an extensive review process by diverse groups of internal and external experts, including educators.
- Test development involves a comprehensive bias and sensitivity review process to ensure every item is reviewed and evaluated, resulting in a fair and appropriate assessment of student learning.

High-quality assessment development and item writing are essential components to produce fair, unbiased tests that provide useful information. Test items are what a student sees and responds to on the test and are often written by current or former teachers.

How do test developers create a high-quality assessment?

To develop a high-quality assessment that is fair, unbiased, and useful, test developers start by creating a plan for testing.[2] The plan includes information about:

- The assessment's purpose.
- How the scores may be used (e.g., to inform instruction or as part of an accountability system).
- The types of items on the assessment (e.g., multiple-choice items or written responses).
- The types of accommodations offered to English learners and students with disabilities.[3]
- The administration mode (e.g., paper and pencil or computer-based).
- (If on computer), whether students will see items sequentially (fixed form) or based on their answers to previous questions (adaptive).

It's important to have a diversity of voices in the planning stage to create an assessment that better meets the needs of students. For example, having discussions about the inclusion of students with disabilities early in the design process helps to build a more accessible assessment by avoiding items that may present accessibility challenges.[4]

The test developer, in partnership with the state, creates a test blueprint based on the plan. The blueprint identifies what will be assessed, the types of items that will be used, and the number or percentage of items for each content area on the test.[5]

Many decisions are made in creating the plan and blueprint that will influence other test development decisions and activities, presenting trade-offs that impact the utility of the scores and student experiences during testing. For statewide assessments, the blueprint generally is not changed over time to allow for comparisons of scores across years.[6]

What steps do test developers take when deciding how to structure a test?

1. Decide what content will be tested.

The blueprint specifies the types of content to be tested based on the state's academic content standards.[7] Test developers do this for two reasons:

- To make sure students are being tested on material that they should have been taught.
- To ensure that the test covers the depth and breadth of the standards being measured.[8]

2. Decide what types of items should be used.

The two most common item types for standardized tests are **selected-response** (e.g., multiple-choice) items and **constructed-response** (e.g., written-response) items. There are trade-offs associated with the choice of item type.[9]

The most common selected-response items are multiple-choice or true-false items. An advantage of selected-response items is that they generally take less time to respond to, so the test can cover more content within a shorter amount of time. They are also efficient in the time and cost to score. Disadvantages of selected-response items include limitations in representing the variety of ways students may show their knowledge. In addition, selected-response items are also open to guessing and do not allow for students to demonstrate partial knowledge when scoring only allows for correct or incorrect responses.

Common constructed-response items include short answer items or essays. The advantages of constructed-response items are that they allow students to demonstrate knowledge in different ways and tend to be more complex. Disadvantages of constructed-response items include that they take longer to respond to and take longer to score. Moreover, as these items tend to be more complex and take more time to respond to during testing, students are more likely to skip these items altogether.[10]

3. Decide how many items should be on the test and how difficult they should be.

Identifying the number of items a student may see is a balancing act.[11] If there are too few items, the test covers less content and there may not be enough items to allow for the reporting of subscores (the scores that describe smaller units of similar content, such as geometry within a math test).[12] This can reduce the perceived value of a test, since the test would not provide educators data on any of the specific skills omitted. If there are too many items, testing time is increased, which can be overwhelming for students and may consume additional instructional time.

Similarly, the test developer makes decisions about the range of difficulty of the items. For fixed-form assessments (i.e., where all students see the same questions), this ensures that there is a mix of easier items and more difficult items. For adaptive assessments (i.e., where students see different questions based on their earlier responses), this ensures that there are sufficient items for students of different performance levels.

How are items written to be fair and unbiased?

Once the test developer designs the testing plan and blueprint, item writers develop items that match the blueprint. Effective item writing is important to ensure that tests are fair and unbiased. Item writers are subject-matter experts and are typically current or former classroom teachers.[13] This experience ensures item writers are familiar with state content standards and expectations for students at different grade levels.

The item writers create the directions, the prompt, and the answer options and select any stimuli the student will see and respond to.[14] For instance, for a reading assessment, a student may be asked to read a passage and respond to questions about it. The item writer would identify a passage, then write a set of items to accompany the passage. Similarly, for science assessments, a stimulus may be graphics or data that accompany the item.

Test developers provide item writers with tools to focus the items on the content the test is trying to measure and reduce the potential for bias (i.e., by systematically over- or under-estimating performance) for students bringing different background knowledge to apply based on a range of different identity attributes (e.g., race, ethnicity, gender, disability status, geographic location, language proficiency, and/or socioeconomic status).[15] These ranges of identities are another reason why having a diversity of perspectives during the test development process is so important.

One tool is an **item writing guide**, which provides general instructions to the item writers. The item writing guide may include information about content, formatting, and style as well as advice on things like avoiding trivial content and keeping vocabulary similar.[16] The goal of the item writing guide is to increase the quality of the items by reducing irrelevant or distracting information so that students focus on the content that is intended to be assessed.[17]

Another set of tools are **item specifications** or **task models**. These tools are used to define important characteristics of an item and what the particular item should be assessing.[18] The item specifications or task model may provide information to the item writer on what types of skills the particular item should focus on. It may also provide boundaries for the type of content to *not* include in the item.[19]

How does the test developer ensure items are high quality?

Each item goes through a rigorous review process to ensure quality with a focus on producing fair and unbiased questions.[20] At each stage in the review process, test development staff may edit the items to fix any deficiencies. If an item is flagged and cannot be appropriately edited, it may be removed altogether.

Internal Reviews

The first stage is an internal review process.[21] In this stage other test development staff, who are also subject-matter experts in the content area, review items for accuracy and clarity. Editors also review items for clarity, consistency in the answer options, and compliance with the testing program's guidelines (e.g., capitalization and punctuation rules). These reviews help to ensure that items are accurate and comply with the item writing guidelines and item specifications.

External Reviews

After internal reviews, items go through content and bias reviews.[22] The reviews are conducted by panels of experts, typically current classroom teachers or district instructional staff, who did not write the items. The panels are selected to be diverse, both demographically (e.g., race/ethnicity, geography) and in terms of work experience (e.g., teaching different student populations).[23] Facilitators train the panelists using fairness guidebooks that are developed in partnership with staff from the state's department of education. Panelists then review items to ensure that the content is accurate and to look for potential bias in the items.

One source of bias that reviewers are vigilant about is sensitive topics. For instance, in Florida, bias reviewers are specifically instructed to flag items about hurricanes or wildfires, as they could be "offensive or distracting" to students.[24] Panelists also look closely at how different racial/ethnic groups are portrayed. The Florida panelists are asked to review items to see if the portrayal of any group is "demeaning, offensive, condescending, or insensitive," if stereotypes are used, or if any group is included too much or too little.[25]

Field Testing

Once the items have been reviewed by experts both internal and external to the testing organization, the items are "field tested" with students.[26] Field testing items means that students complete the items during a typical administration of the test, but the results from that item are not used to calculate the students' test scores. Often, items are embedded into a test so that students generally do not know which items are being used to calculate their score. If items are revised, they are often field tested again with students.

Item Statistics

The field test provides the test developer with data to determine if the items should be revised or excluded from the item pool. Field testing items also provides developers with data to inform how the items could fit within the test blueprint. These statistics are regularly computed each time the item is administered.

One statistic that is calculated is the item's **difficulty**. Difficulty represents how easy or hard an item is for students to answer correctly.[27] The test developer will review the item difficulty to ensure that the items are not too easy or too difficult. For instance, if every student answers an item incorrectly, the item may be too difficult and should be revised or removed. The test developer will also review the difficulty levels to make sure items fit within the target difficulty ranges for the blueprint.

A second statistic is the item's **discrimination**.[28] In this context, discrimination means how well the item differentiates among students based on how well students know the material. In testing, discrimination is a valuable property of a test item. It is measured on a 0 to 1 scale. An item with high discrimination (i.e., close to 1) is one where high-performing students answer the item correctly and low-performing students answer the item incorrectly. This means that items with higher discrimination values do well at identifying students who know the content compared to students who do not know the content. If an item has a low discrimination value (i.e., close to 0), it cannot identify students who know or do not know the content. The item might be too easy, or a previous item might inadvertently provide the answer, enabling nearly all students to answer it correctly. When an item has a low discrimination value, the item is checked to make sure that the answer key is accurate and the content of the item is correct.

A third statistic is **differential item functioning** (DIF). DIF occurs when students from different groups (e.g., gender, race/ethnicity, disability status) have similar levels of ability on what is being tested but have different responses to a particular item.[29] This means that students of different groups expected to answer the item similarly, based on their responses to other items, offer different answers instead. DIF analyses are a further check to make sure that a test is fair for all students.

If the analyses flag an item for potential DIF, the test developer, and in many cases educators, review the item to see why it may be functioning differently for certain groups of students.[30] If the item is reviewed and there is not a plausible explanation for the difference in responses, the test developer may continue to use the item. If the review identifies an issue with the item, the item may be revised or dropped from the item pool entirely.

It should be noted that writing items is not a one-time process. Items may be reused for multiple test administrations until they are eventually retired. New items are continually developed to refresh the item pool. Each item goes through this review process, beginning with the test developer's internal reviews through analyzing the data after students take the test. The process provides a safeguard to ensure that the item is fair and unbiased.✦

Endnotes

[1] Formative assessments are "a planned, ongoing process" that provide evidence of student learning to improve student outcomes. Formative assessments help educators design activities and instructional material to better align with student needs during learning. Interim assessments are another form of assessment that are administered periodically during the year and — depending on the assessment — can serve a formative function (i.e., for learning) or a summative function (i.e., to measure how much a student has learned). Formative Assessment for Students and Teachers (FAST) State Collaborative on Assessment and Student Standards, *Revising the Definition of Formative Assessment*, Council of Chief State School Officers, 2018, https://ccsso.org/sites/default/files/2018-06/Revising%20the%20Definition%20of%20Formative%20Assessment.pdf.

[2] Steven M. Downing, "Twelve Steps for Effective Test Development," in Steven M. Downing and Thomas M. Haladyna, eds., *Handbook of Test Development* (Mahwah, NJ: Lawrence Erlbaum Associates, 2006), 3–26. Robert J. Mislevy and Michelle M. Riconscente, "Evidence-Centered Assessment Design," in *Handbook of Test Development*, 61–90.

[3] For more information, see the "Making Assessments Accessible for Students With Disabilities and English Learners" brief.

[4] Trisha Klein, "Making Sense of Test-taking by Touch," Smarter Balanced, January 22, 2020, https://smarterbalanced.org/making-sense-of-test-taking-by-touch/; Martha L. Thurlow, Sandra H. Warren, and Magda Chia, *Guidebook to Including Students With Disabilities and English Learners in Assessments*, National Center on Educational Outcomes, June 2020, https://files.eric.ed.gov/fulltext/ED609705.pdf; American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, eds. *Standards for Educational and Psychological Testing* (Lanham, MD: American Educational Research Association, 2014), https://www.testingstandards.net/open-access-files.html.

[5] American Educational Research Association et al., *Standards for Educational and Psychological Testing*, 79.

[6] For more information, see the "Ensuring Comparability Across Administrations" brief.

[7] For more information, see the "What Statewide Assessments Are Designed to Measure" brief.

[8] Test developers seek to create a test blueprint that adequately covers the range of the standards that students are expected to meet to avoid assessing only some of the grade-level standards or assessing them using a narrow range of cognitive complexity (e.g., including only those items that require the recalling of facts, not the application of knowledge). See Downing, "Twelve Steps for Effective Test Development" in *Handbook of Test Development* for more information.

[9] Steven M. Downing, "Selected-Response Item Formats in Test Development," in *Handbook of Test Development*, 287–302.

[10] Pamela M. Jakewerth and Frances B. Stancavage, "An Investigation of Why Students Do Not Respond to Questions," working paper, NCES, 2003, https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=200312.

[11] Depending on the type of assessment, students may see different numbers of items. In an adaptive test, students see different items depending on prior responses. The number of items a student sees can vary depending on the test developer's decision rules.

[12] William Monaghan, "The Facts About Subscores," *R&D Connections*, July 2006, https://www.ets.org/Media/Research/pdf/RD_Connections4.pdf.

[13] Steven M. Downing, "Twelve Steps for Effective Test Development," in *Handbook of Test Development*.

[14] Thomas M. Haladyna and Michael C. Rodriguez, *Developing and Validating Test Items* (New York: Routledge, 2013).

[15] "Test Bias," *APA Dictionary of Psychology* (Washington, DC: American Psychological Association, 2023), https://dictionary.apa.org/test-bias.

[16] Thomas M. Haladyna, Steven M. Downing, & Michael C. Rodriguez, "A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment," *Applied Measurement in Education* 15, no. 3 (2002): 309–334.

[17] Research is emerging in culturally relevant assessment that questions the assumption of content neutrality. Jennifer Randall, "'Color-Neutral' Is Not a Thing: Redefining Construct Definition and Representation Through a Justice-Oriented Critical Antiracist Lens," *Educational Measurement: Issues and Practice* 40, no. 4 (2021): 82–90, https://onlinelibrary.wiley.com/doi/abs/10.1111/emip.12429. Some contend that the degree of cultural relevance depends on the type of assessment, such that state summative assessments should remain content neutral while classroom assessments should be designed with cultural relevance. Carla Evans, "Are There Limits to Making Large-Scale Standardized Testing Culturally Responsive?," Center for Assessment, November 3, 2021, https://www.nciea.org/blog/culturally-sensitive-relevant-responsive-and-sustaining-assessment/.

[18] Mislevy and Riconscente, "Evidence-Centered Assessment Design," in *Handbook of Test Development.*

[19] For an example of Next Generation Science Standards item specification, see CCSSO Science Assessment Item Collaborative, *Item Specifications Guidelines for the Next Generation Science Standards* (Washington, DC: CCSSO, 2015), https://csaa.wested.org/wp-content/uploads/2019/12/SAIC_Item_Specifications_Guidelines.pdf.

[20] Haladyna and Rodriguez, *Developing and Validating Test Items*; American Educational Research Association et al., *Standards for Educational and Psychological Testing.*

[21] Downing, "Twelve Steps for Effective Test Development," in *Handbook of Test Development.*

[22] Haladyna and Rodriguez, *Developing and Validating Test Items.*

[23] Ibid.

[24] *Bias and Sensitivity Review: District Developed Assessments*, Florida Department of Education, March 2012, https://www.fldoe.org/core/fileparse.php/5423/urlt/bsrdda.pdf.

[25] Ibid.

[26] American Educational Research Association et al., *Standards for Educational and Psychological Testing.*

[27] "Understanding Item Analyses," Office of Educational Assessment, University of Washington, https://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis/; See also American Educational Research Association et al., *Standards for Educational and Psychological Testing.*

[28] "Understanding Item Analyses," Office of Educational Assessment, University of Washington.

[29] American Educational Research Association et al., *Standards for Educational and Psychological Testing*, 82.

[30] For examples of biased items, see W. James Popham, *Assessment Bias: How to Banish It*, second edition (Boston, MA: Pearson, 2012), https://iarss.org/wp-content/uploads/2016/05/Popham_Bias_BK04.pdf.

About the Authors

MICHELLE CROFT

Michelle Croft is a senior analyst at Bellwether in the Policy and Evaluation practice area. She can be reached at **michelle.croft@bellwether.org**.

HAILLY T.N. KORMAN

Hailly T.N. Korman is a senior associate partner at Bellwether in the Policy and Evaluation practice area. She can be reached at **hailly.korman@bellwether.org**.

TITILAYO TINUBU ALI

Titilayo Tinubu Ali is a partner at Bellwether in the Policy and Evaluation practice area. She can be reached at **titilayo.ali@bellwether.org**.

About Bellwether

Bellwether is a national nonprofit that exists to transform education to ensure systemically marginalized young people achieve outcomes that lead to fulfilling lives and flourishing communities. Founded in 2010, we work hand in hand with education leaders and organizations to accelerate their impact, inform and influence policy and program design, and share what we learn along the way. For more, visit **bellwether.org**.

ACKNOWLEDGMENTS

We would like to thank the many individuals who gave their time and shared their knowledge with us to inform our work, including the Consortium on Assessment Policy and Advocacy (CAPA) for their support of this project.

We would also like to thank our Bellwether colleagues Andrew J. Rotherham, Harold Hinds, and Alexis Richardson. Thank you to Alyssa Schwenk, Abby Marco, Andy Jacob, Zoe Campbell, Julie Nguyen, and Amber Walker for shepherding and disseminating this work, and to Super Copy Editors.

The contributions of these individuals and entities significantly enhanced our work; however, any errors in fact or analysis remain the responsibility of the authors.

Bellwether

FORWARD THINKING. FORWARD MOVING.