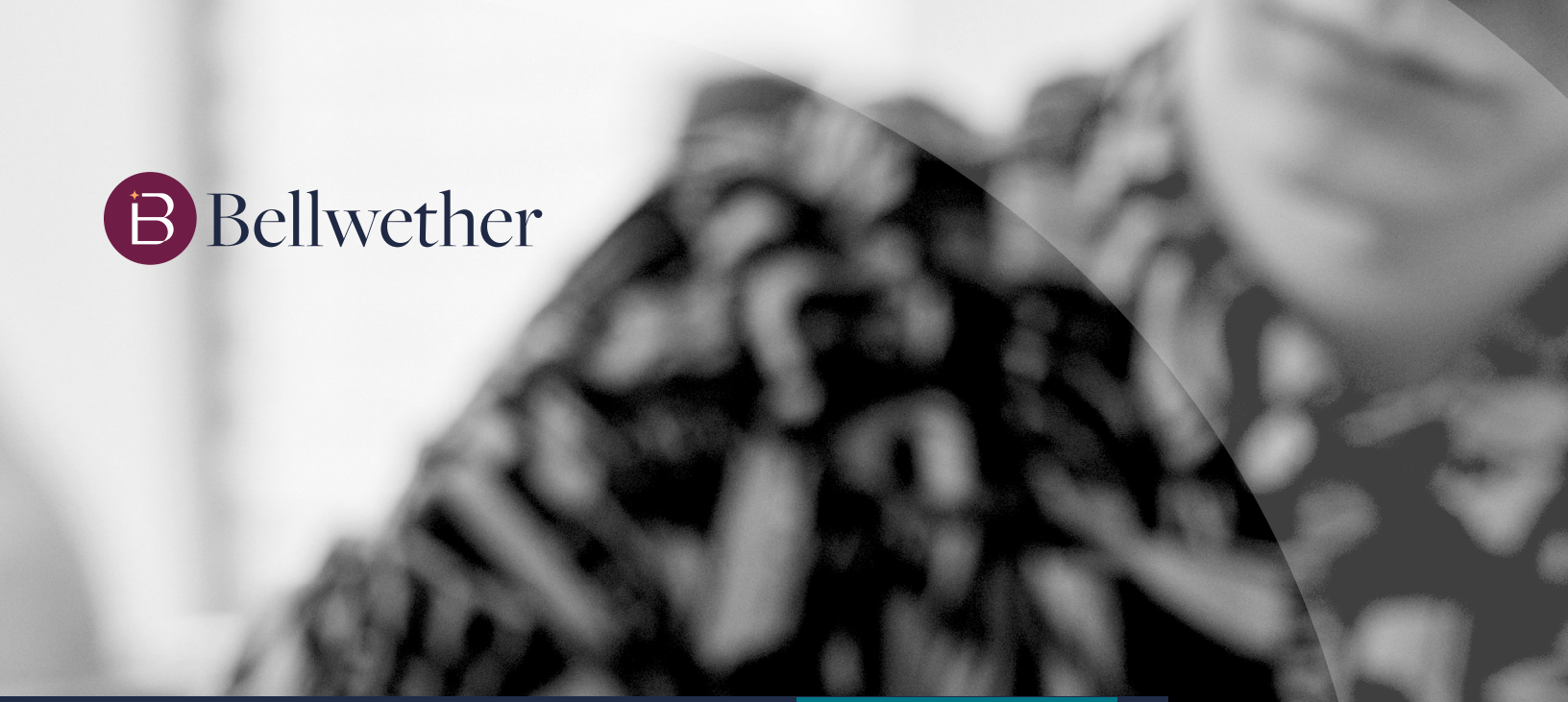

Bellwether

#2 IN THE SERIES

Demystifying Statewide Standardized Assessments

Ensuring That Assessments Accurately Measure Academic Standards

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

APRIL 2023

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.*

Ensuring That Assessments Accurately Measure Academic Standards

Test developers collect validity evidence to provide support for the interpretations and uses of the test scores. Validity evidence comes from multiple sources including what's being tested, how students respond to the questions, the internal structure of the test, how the scores relate to other types of data, and the reliability of the scores. The collection of validity evidence is an ongoing process.

Key Takeaways

- Assessments must demonstrate through multiple types of evidence that the conclusions drawn from the results are appropriate.
- The collection of validity evidence is an ongoing process because content standards, teaching practices, and students change over time.
- The U.S. Department of Education has an important role in reviewing statewide tests to ensure that students are taking high-quality assessments and that tests are meeting technical requirements.

To use statewide assessment scores for accountability and to inform learning, there must be evidence to support the appropriateness of the inferences (or conclusions) made based on those test scores. For example, an educator may infer that a student needs remediation in math based on the student's math score, or a parent may infer that their student is reading on grade level based on the student's reading score. All of these inferences need evidence to support them, and **validity** is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests."[2] Test developers must build a case that any inference based on a test score is appropriate. The different types of evidence test developers collect to measure the validity of the test's scores (often called "validity evidence") depends on how test scores will be used and interpreted.

The reason that validity evidence focuses on how assessment scores are interpreted and not the assessment itself is because test scores could be used in many different ways, and the test developer (or user) must provide evidence to support each particular use.[3] For example, if a district develops a policy that students must receive a certain score on the next grade level's end-of-year test to be accelerated (i.e., skip a grade), the district must collect evidence to support the use of the test score for acceleration purposes; it cannot solely rely on the test developer's evidence for students taking the test at the end of the year in which they were taught the content. Even though a test score may appear credible for a *related* use, even a small difference in application may require additional evidence to support that other use.

How do test developers identify and collect validity evidence to establish a validity argument?

Test developers start collecting evidence at the very beginning of the assessment design process, documenting the assessment's purpose, what will be assessed, what the scores are supposed to convey, how the items will be written and presented to students,[4] and how the scores will be reported.[5] All of these factors play a role in developing evidence of validity.[6]

Assessment's Content

Test developers collect evidence of the assessment's content by examining the relationship between the content of the test and what the test is intended to measure.[7] Test developers have internal processes during assessment and item development (e.g., instructions to item writers and item reviewers) to monitor that the test is indeed measuring what is intended.[8]

Test developers also conduct alignment studies typically involving a diverse set of educators within the state.[9] Educators compare the test items to the state's standards to ensure that the items match the content and complexity for that grade level and that they adequately cover the required standards.

Students' Response Process

Another form of validity evidence is students' response process, meaning how the students come to their answer when responding to a test item.[10] Validity evidence of the response process looks at the relationship between what the test is trying to measure (also known as the construct) and the task the student is doing. For example, if the test is designed to measure math reasoning but students can follow a standard algorithm to answer the question, the test may not be providing information about students' reasoning ability.

One strategy to achieve this is a special research study called a "cognitive lab," in which participating students think aloud while completing items from the assessment to demonstrate how they are approaching the problem.[11] The cognitive labs allow test developers to see if the test item is measuring what was intended and can help confirm (or refute) assumptions on how students will respond.

Internal Structure

Validity evidence about the assessment's internal structure helps build the case for how items on the test relate to one another and what the test is designed to measure.[12] Test developers use statistical analyses to confirm that items that test similar content have similar outcomes.

Relationship to Other Variables

Test developers examine the relationship between different assessments with similar — or different — content to assess whether results are correlated as expected.[13] For instance, we would expect to see a high correlation between two math assessments. We would expect to see a lower correlation between a math assessment and a reading assessment.

Similarly, test developers look at the relationship between the assessment scores and other types of information about the student's academic performance. For example, test developers examine how closely student GPA and test scores are correlated because we would expect that students performing well on the test would have higher GPAs than lower-performing students. Likewise, test developers might also look at the relationship between teacher ratings of student performance and assessment scores. Students rated highly on academic performance by their teacher would be expected to have higher test scores compared to students rated as lower performing.

Reliability

Reliability means assessments yield dependable results by consistently measuring knowledge and skills.[14] To estimate reliability, test developers may have students retake the test. Developers can also examine the consistency of the items within the test.

COLLECTION OF VALIDITY EVIDENCE

Test developers examine all pieces of evidence to see if they are consistent with the proposed interpretation of the score. Optimally, all types of validity evidence should point in the same direction and provide evidence that any inferences made are appropriate.

The collection of validity evidence is an ongoing process and is never considered done. As test items, content standards, teaching practices, and students change over time, test developers continue to collect validity evidence and refine their validity argument.

What oversight ensures that test developers are collecting validity evidence?

Test developers follow the Standards for Educational and Psychological Measurement, which requires collecting validity evidence.[15] For statewide assessments, there are several external reviews of technical quality that provide oversight to promote the quality of the tests.

Under the Every Student Succeeds Act, states must submit technical documentation about their statewide assessments used for federal accountability to the U.S. Department of Education (ED) for peer review.[16] These peer panels are experts selected by ED to review the evidence submitted by the states to support the tests. Peer decisions are reviewed by ED, which issues a letter to the state either identifying any areas where additional evidence is needed or test quality needs to be improved, or concluding that the test meets all federal requirements. This peer review process is designed to provide a safeguard for technical quality and to improve the overall quality of tests over time. To increase transparency about testing, the ED letters and peer notes are publicly available.[17]

As an additional part of this peer review process, states are required to convene and consult a Technical Advisory Committee (TAC) to review the state's summative assessment system.[18] The TAC is comprised of experts in educational measurement. The TAC reviews technical materials and provides advice to the state and the state's test vendor.✦

To use statewide assessment scores for accountability and to inform learning, there must be evidence to support the appropriateness of the inferences (or conclusions) made based on those test scores.

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 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.

3 Ibid.

4 For more information, see the "Developing High-Quality Assessments and Items" brief.

5 For more information, see the "Reporting Assessment Scores" brief.

6 Joan L. Herman and Robert Linn, "Evidence-Centered-Design: A Summary," National Center for Research on Evaluation, Standards, & Student Testing, 2015, https://csaa.wested.org/wp-content/uploads/2019/11/ECDsummary.pdf.

7 American Educational Research Association et al., *Standards for Educational and Psychological Testing*.

8 For more information, see the "Developing High-Quality Assessments and Items" brief.

9 Ellen Forte, *Evaluating Alignment in Large-Scale Standards-Based Assessment Systems*, CCSSO, February 2017, https://edcount.com/wp-content/uploads/2019/05/ccsso_tilsa_forte_evaluating_alignment_2017.pdf.

10 American Educational Research Association et al., *Standards for Educational and Psychological Testing*.

11 Sasha Zucker, Christy Sassman, and Betsy J. Case, *Cognitive Labs*, Pearson, February 2004, http://images.pearsonassessments.com/images/tmrs/tmrs_rg/CognitiveLabs.pdf.

12 American Educational Research Association et al., *Standards for Educational and Psychological Testing*.

13 Ibid.

14 Ibid.

15 Ibid.

16 *A State's Guide to the U.S. Department of Education's Assessment Peer Review Process*, U.S. Department of Education, September 24, 2018, https://www2.ed.gov/admins/lead/account/saa/assessmentpeerreview.pdf.

17 "Decision Letters on State Assessment Systems Under Title I of the ESEA," U.S. Department of Education, https://oese.ed.gov/offices/office-of-formula-grants/school-support-and-accountability/decision-letters-on-state-final-assessment-system/.

18 *A State's Guide to the U.S. Department of Education's Assessment Peer Review Process*.

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.