Bellwether

#4 IN THE SERIES

Demystifying Statewide Standardized Assessments

Ensuring Comparability Across Administrations

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 Comparability Across Administrations

State test scores are valuable because unlike measures such as student grades, they can be directly compared to other students' scores and across schools or districts. To achieve comparability, the tests are standardized and go through a statistical process called equating.

Key Takeaways

- Comparability is the ability to compare test scores across time.
- Standardization helps ensure that test scores are comparable.
- Test scores are equated to adjust for differences in difficulty between two test administrations.

An essential feature of statewide testing is **comparability**: the ability to compare scores across test forms (i.e., versions of the same test) and/or across time.[2] For instance, for test security reasons, there may be multiple test forms for a single administration. Similarly for computer-adaptive tests, test items vary in difficulty based on the student's responses so no two students will likely see the same items.[3] Comparability allows the scores between two forms to be used interchangeably. It also allows for the comparison of scores from year to year, such as comparing one group of students who took a test in 2022 to other groups of students who take the same test in 2023.

Comparability provides a consistent and common metric that other measures like teacher feedback or grades cannot, allowing policymakers, school leaders, and families to compare scores across districts and schools.

What is standardization and how is it related to comparability?

Standardization is the process of ensuring that all students have the same (or very similar) testing experience.[4] Some components of standardization may include:

- The amount of time students have to take each section of the test.
- The directions that are read to students prior to testing.
- If (and when) reminders are given to students about how much time is remaining in the test.
- The types of accommodations that are allowed (e.g., breaks in between sections, testing in a quiet environment).
- The security of the testing environment (e.g., barriers between screens so students cannot see other student responses) or the length of the testing window (i.e., the number of days or weeks during which a student can test).
- If computer-based, the types of devices (e.g., iPad, Chromebook) that the test can be taken on.

Without standardization, we are less sure that the score a student receives is due to their knowledge and skills or to other factors. For instance, if there are not standardized directions, students may be approaching the assessment process differently. Educators, families, and policymakers would not know whether differences in the scores represent real differences in knowledge and skills or different understandings about the assessment process.

The extent of standardization varies depending on the assessment and its uses. Assessments used for routine activities, such as interim assessments or classroom assessments used to inform instruction, may require less standardization. Statewide assessments require higher standardization requirements; since test scores play such a large role in accountability systems and impact resource allocation, stricter requirements can reduce concerns about unrelated factors influencing those scores.

What is equating and how does it relate to comparability?

The process of **equating** helps address concerns about differences in the assessment items themselves. Although the test blueprint has targeted ranges of difficulty for items,[5] no two test forms will be exactly the same in terms of difficulty. Because of these differences in difficulty, test developers use a statistical method called equating to adjust scale scores, taking the differences across forms into account.[6] There are different methods of equating, but the overall goal is to have scores that are interchangeable even when there are differences across test forms.

For equating to work, several conditions need to be met.[7] They can be stated simply as follows:

- The two test forms should be built to the same specifications, so that the test forms cover essentially the same content at similar levels of difficulty.
- Translating scores on Form 1 into scores on Form 2 and back again should leave you where you started (i.e., the transformation between the two sets of scores should be mathematically symmetric).
- Ideally, the test taker should not prefer one test form over another.
- The relationship between scores on the two test forms should be the same for every subpopulation (e.g., gender, race, socioeconomic status, testing year). ✦

Comparability provides a consistent and common metric that other measures like teacher feedback or grades cannot, allowing policymakers, school leaders, and families to compare scores across districts and schools.

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

4 Thomas M. Haladyna and Michael C. Rodriguez, *Developing and Validating Test Items* (New York: Routledge, 2013); American Educational Research Association et al., *Standards for Educational and Psychological Testing*.

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

6 Michael J. Kolen and Robert L. Brennan, *Test Equating, Scaling, and Linking: Methods and Practices* (New York, NY: Springer, 2014).

7 Ibid.

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.