Bellwether

#6 IN THE SERIES

Demystifying Statewide Standardized Assessments

Reporting Assessment Scores

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.*

Reporting Assessment Scores

Assessments report multiple types of scores that have different meanings and offer complementary information. For statewide assessments used for accountability, states must provide accessible score reports to families and publicly report scores for schools and districts.

Key Takeaways

- Assessment scores should be used alongside additional information educators and families have about what students know and can do.
- Test developers provide test scores in the form of individual student reports and aggregated reports to inform parents and educators of student performance.
- Assessment scores reported to the federal government must be accessible for families, including families whose primary language is not English and parents with disabilities.
- To meet federal requirements, states must publicly report disaggregated data by student subgroup (e.g., student race/ethnicity, students with disabilities, English learners).

Assessment scores are a snapshot of what a student knows and can do at one moment in time. While scores from statewide assessments can help educators make certain instructional decisions and highlight a student's or a group of students' strengths and areas of needed support,[2] they have limitations. Absent additional evidence, test scores should only be used in the ways the test developer intended when designing and developing the test.[3] Equally important, assessment scores should not be used in isolation, particularly to make important decisions about a student. Assessment scores should always be interpreted within the context of other information about the student such as other standardized test scores, classroom tests and student work, teacher observations, and parent observations. Considering additional information about a student can provide a more complete picture of a student's strengths and areas where they may need support.

What are some of the types of assessment scores?

There are multiple types of scores that are commonly reported on statewide assessments that can provide different information about a student, including raw score, scale score, subscore, percentile rank, grade equivalent, growth scores, and/or achievement levels.

Raw Score

The **raw score** is most often the total number of items answered correctly.[4] The raw score provides a general indicator of student performance, but because of differences in difficulty across assessment items, the meaning of the raw score is limited. For instance, answering three difficult items correctly out of six items may be more challenging than answering four easy items out of six items. Raw scores can be used to compare students for an assessment that uses a single test form. When assessments have multiple test forms, or change from one year to the next, raw scores are not a comparative tool.

Scale Score

After equating,[5] raw scores are usually adjusted so they can be reported on the same scale.[6] **Scale scores** allow for accurate comparisons across test forms and administration years within a grade or course and content area.[7]

A carefully designed score scale can provide families, educators, and policymakers with useful information about student performance (e.g., understanding if the student's score is generally high, low, or about average) and can facilitate meaningful comparisons across test forms and testing years. The score scale continues to gain meaning through accumulated information on the performance of groups of interest.

Subscore

The scale score represents knowledge and skills across the entire content area. Educators and parents may be interested in more specific information about a student's particular knowledge, skills, and abilities, which are called **subscores**.[8] The metric for reporting subscores varies across tests. Common metrics include raw score, scale score, perfect correct, or performance category (e.g., "proficient").[9]

Percentile Rank

A **percentile rank** indicates the percentage of a group of similar test takers that scored below a given score.[10] A score at the 50th percentile means that the student scored higher than 50% of the other test takers. A score at the 10th percentile means the student scored higher than 10% of the other test takers.

The percentile rank is helpful because it provides information about a student's relative standing compared to other students. However, it doesn't provide any information about the types of content within the test that the student performed well on or struggled with. Another disadvantage of the percentile rank is that it is often confused with the percent correct (i.e., the percentage of questions answered correctly or incorrectly). Finally, comparisons of percentile ranks are of limited value as most students will score toward the average.[11]

Grade Equivalent

A **grade equivalent** "indicates the grade in the norm group for which a certain raw score was the median performance."[12] These can be useful because they can help compare the student's performance to students in a particular grade level. The grade equivalent usually includes two numbers: the grade level and the month during the school year (with the year consisting of ten months). For example, a grade equivalent of 5.3 means the student performance is similar to the performance of a typical student taking that test in the third month of fifth grade.

Grade equivalent scores help to show some comparative information to other grade levels, but they do not show when a student is ready for advanced material. For instance, if a third-grader received a grade equivalent of 5.3 on a standardized test, it does not mean they are ready for fifth-grade material. Instead, it means that based on the third-grade content, the third-grader scored about the same as a typical fifth-grader being assessed on similar content.

Growth Scores

Growth scores describe the academic performance of a student or group of students over two or more time points.[13] There are many different ways to calculate growth, each with their own advantages and disadvantages.

Growth scores are different than the types of scores previously discussed, which all measure a single time point. Growth scores try to capture changes in student performance.

Achievement Levels

Achievement levels are used to help with score interpretation. Achievement levels divide the score scale into descriptions based on what students know and can do. For example, the National Assessment of Educational Progress (NAEP) has three achievement levels: basic, proficient, and advanced.[14] Like NAEP, states are required to set at least three achievement levels for the state summative test with accompanying definitions of what each achievement level means.[15]

Achievement levels help to provide a general indicator of student proficiency. A student who is proficient is on track to meet grade-level expectations.

HOW ARE ACHIEVEMENT LEVELS DEFINED?

Test scores are often reported as achievement levels. To set those achievement levels, the test developer engages in a process called standard setting, which identifies points along the score scale that match the achievement level (e.g., the NAEP score that indicates a student is proficient instead of basic). These points are referred to as performance standards or cut scores.

There are multiple methods for setting cut scores that involve slightly different types of data or expert judgments,[16] but the process for setting cut scores is fairly similar across these different types of methods. First, the state sponsoring the test works with the test developer to select the panelists. As is the case with content and bias reviews,[17] panelists, who are typically educators, must have content expertise and should be demographically diverse (e.g., geography, race/ethnicity, gender, work experience). The state and test developer then develop descriptions of each performance category. The achievement levels discussed earlier are used to help in developing those categories.

The next step is training the panelists, both on the standard-setting process more generally and on the specific method that they will be using to set standards. The panelists then implement the method.[18] Each panelist makes judgments about the items, and a facilitator compiles the individual panelist ratings into a summary. Panelists review the summary and the impact of setting the cut score, using actual student data, to see how many students would be classified into each of the levels. The panelists discuss the results and may complete a second set of ratings. As before, the facilitator compiles the panelist ratings and uses the aggregated ratings to set the performance standards.[19]

For statewide assessments, the state — typically through the state board of education — must formally adopt the performance levels.

How are scores reported?

To enhance the utility of assessment scores, score reports are provided to families and educators.

There are two primary types of score reports. The first are **individual student score reports**, which are provided to educators and families as soon as practicable after the test administration. For statewide assessments, the U.S. Department of Education (ED) requires states to make certain guarantees to increase the utility and accessibility for families.[20] First, the content of the score report should be understandable. The score reports are reviewed as part of the ED peer review process along with other materials states may provide to help families interpret their child's scores.[21] Second, the reports must also be accessible in language and format so that families who do not speak English or who have disabilities have access to their child's score reports.[22] States and school districts are required to provide written or oral translations for parents or guardians with limited English proficiency. They are also required to provide reports in an alternate format upon request when a parent or guardian has a disability.

The second are **aggregate-level reports**, which report scores at the classroom, school, district, or state level.[23] The reporting at the different administrative levels is used by educators and policymakers, either for instructional purposes (e.g., to help identify areas of strength or areas for improvement) or for accountability purposes, to better allocate resources to schools or districts. Aggregate-level reports are useful because they help surface trends in student performance to better highlight areas of need. For example, classroom-level reports may highlight areas to further evaluate curriculum quality or to provide professional development. State-level reports may show differences in student performance across districts and schools.

What scores are publicly reported?

As noted within the comparability brief (the fourth in this series), states must publicly report the scores for schools and districts and break out the scores for different student populations such as by race/ethnicity, socioeconomic status, disability status, English proficiency, gender, and migrant status[24] — allowing parents and community members to see how schools are educating all students.✦

Assessment scores should always be interpreted within the context of other information about the student such as other standardized test scores, classroom tests and student work, teacher observations, and parent observations ... [to] provide a more complete picture of a student's strengths and areas where they may need support.

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 For more information about how the types of uses of the scores depend on the type of assessment, see the "What Statewide Assessments Are Designed to Measure" brief.

3 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.

4 Some tests use "formula scores," where the raw score is the number of items answered correctly, minus a fraction of the number of items answered incorrectly (often called a "guessing penalty"). Joseph Ryan and Frank Brockmann, *A Practitioner's Introduction to Equating With Primers on Classical Test Theory and Item Response Theory*, CCSSO, 2018, https://ccsso.org/sites/default/files/2018-06/A%20Practitioners%20Introduction%20to%20Equating%20revised%20edition.pdf; Robert B. Frary, "Formula Scoring of Multiple-Choice Tests (Correction for Guessing)," *Educational Measurement Issues and Practice* 7, no. 2 (1998): 33–38, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-3992.1988.tb00434.x.

5 See the "Ensuring Comparability Across Administrations" brief for more information.

6 A scale associates numbers with performance, so that higher scores indicate increased achievement (e.g., 100-200). Michael J. Kolen, "Scaling and Norming," in Robert L. Brennan, ed., *Educational Measurement*, fourth edition (Westport, CT: American Council on Education/Praeger, 2006), 155–186.

7 Kolen, "Scaling and Norming," in *Educational Measurement*.

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

9 Victoria Tanaka, "Promoting Effective Practices for Subscore Reporting and Use: A Framework for Reporting Technically-Sound and Useful Subscores on State Assessments," Center for Assessment, September 11, 2019, https://www.nciea.org/blog/promoting-effective-practices-for-subscore-reporting-and-use/.

10 Craig A. Mertler, "Module 6: Norm-Referenced Test Scores and Their Interpretations," in *Interpreting Standardized Test Scores: Strategies for Data-Driven Instructional Decision Making* (London: SAGE Publications, 2007).

11 Ibid.

12 Ibid.

13 Katherine E. Castellano and Andrew D. Ho, *A Practitioner's Guide to Growth Models*, CCSSO, February 2013, https://scholar.harvard.edu/files/andrewho/files/a_pracitioners_guide_to_growth_models.pdf.

14 "Scale Scores and NAEP Achievement Levels," National Center for Education Statistics, https://nces.ed.gov/nationsreportcard/guides/scores_achv.aspx.

15 20 U.S.C. § 6311 (b)(1)(A), https://www.law.cornell.edu/uscode/text/20/6311.

16 Ron K. Hambleton and Mary J. Pitoniak, "Setting Performance Standards," in *Educational Measurement*, 433–470.

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

18 For example, in the "Bookmark" method, test items are ordered by level of difficulty from easiest to hardest, and panelists are asked to indicate the last item a "borderline" student (a student who is "just good enough" to fit the performance category) would be able to answer correctly.

19 To support their findings, panelists are asked to evaluate the process, typically by completing a survey about their satisfaction with the process (e.g., the training, the facilitation, the final performance standards). These survey results are included in a technical report that documents the standard setting process and the performance levels. The panelists' recommended cut scores may not be used — or given less weight — if the evaluation indicates there were substantial problems with the standard setting process.

20 *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.

21 For more information about peer review, see the "Ensuring That Assessments Accurately Measure Academic Standards" brief.

22 20 U.S.C. § 6311 (b)(3)(C)(xi), https://www.law.cornell.edu/uscode/text/20/6311; *A State's Guide to the U.S. Department of Education's Assessment Peer Review Process*.

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

24 20 U.S.C. § 6311(b)(2)(B), https://www.law.cornell.edu/uscode/text/20/6311.

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.