

R & D Connections

No. 16 • September 2011

Why Do Standardized Testing Programs Report Scaled Scores? Why Not Just Report the Raw or Percent-Correct Scores?

By Xuan Tan and Rochelle Michel

Key concepts

To understand why testing programs report scaled scores, it is useful to understand these ideas:

- **Raw score** — Total number of raw points a test taker receives based on the number of questions answered correctly; typically, for example, 20 correct answers means a raw score of 20
- **Scaled score** — Scores that have been mathematically transformed from one set of numbers (i.e., the raw score) to another set of numbers in order to make them comparable in some way — for example, across different editions, or “forms,” of the same test

continued on p. 2

A common question asked by policymakers, teachers, parents, and some test takers is: Why is it not appropriate to use raw or percent-correct scores for comparing test takers across different test editions? Why do standardized testing programs go through complicated statistical processes to come up with scaled scores?

Standardized tests are widely used in K–12, higher education, and in some professions for such purposes as accountability, admissions, certification, and licensure. Standardized tests provide a common basis for evaluating and comparing test-takers’ abilities in a specific content area. They are administered and scored in a predetermined, set manner that is consistent for all test takers (e.g., test questions, time allowed for each administration, scoring procedures).

However, in order for standardized testing programs to have consistency in score interpretation when there are different editions of the test, programs often transform test scores (summed raw score points assigned to different questions) into a set of values different from the raw score points obtained directly from a test. Further, testing programs often report these transformed test scores, which are called scaled scores, rather than reporting percent-correct scores derived from the raw score points. This standardization allows scores reported from a test to have consistent meaning for all test takers.

This article highlights the reasons why percent-correct scores are generally not used as the primary reported scores and provides additional details on what a scaled score is, how scaled scores are obtained, and the reasons for providing scaled scores, as well as their usefulness in interpreting test scores.

The Raw and Percent-Correct Score

A *raw score* is the total number of score points a test taker obtains by answering questions correctly on a test. A *percent-correct* score represents the percentage of questions a test taker answered correctly on a test. For example, if a test taker answered 20 out of 50 questions on a test correctly, then his or her percent-correct score would be 40%. The raw score, in this example, is 20. The percent-correct score can be considered

Editor’s note: Xuan Tan is a Psychometrician in the ETS Research & Development division’s Center for Statistical Analysis. Rochelle Michel, also a doctoral-level psychometrician by training, is a Product Manager in ETS’s Higher Education division.

Key concepts

continued from p. 1

- Equating — Process used to place all forms of the same test on the same scale and make scores comparable across forms
- Anchor items — Set of questions that is common to different forms of the same test in order to facilitate the statistical comparison of group ability and form difficulty that takes place during equating
- Percent-correct score — Percentage of questions a test taker answered correctly on a test; often used in classroom tests

an adjusted raw score to account for differences in the lengths of different tests. The percent-correct score is easy to calculate and understand, and is often used in classroom tests for score reporting.

Why are percent-correct scores not used as the primary reported scores?

In the case of many standardized testing programs, it is necessary to develop multiple editions of a test as a way of dealing with the issue of content *exposure*. Test questions from standardized tests are usually secure, but when the same test is repeatedly administered to a large number of test takers, the questions can become exposed to the public and jeopardize the testing process. Test takers can remember the test questions and share them with future test takers through different media. To address this concern, test takers taking a test at different times may receive different editions of the test. Also, test takers within the same administration may be administered different editions of the test to address security concerns. In some cases, multiple editions of a test are developed in response to state laws requesting the disclosure of test questions after each administration.

Standardized testing programs often develop different editions of the same test that contain different sets of test questions conforming to predefined content and statistical specifications. These different editions are commonly called *forms*.

Although strict adherence to common test specifications or blueprints allows test developers to create multiple forms that are remarkably *similar* in difficulty, they are rarely, if ever, exactly *equal* in difficulty (Holland & Dorans, 2006). This makes it hard to use the percent-correct score for fair comparisons of test takers' performances on different forms of the same test. For example, getting 50% correct on a hard form may mean the test taker has more knowledge and skill than another test taker getting 60% correct on a relatively easier form. For the same reason, the raw scores cannot be used to compare test takers' performances on different forms. When two test takers get the same raw score on two different forms, the test taker who took the more difficult form has demonstrated a higher level of performance than the test taker who took the relatively easier form.

Most standardized testing programs require scores that can be compared across different forms. In order for different stakeholders (states, schools, etc.) to make consistent and fair decisions based on assessment results, the scores reported from standardized tests need to be *comparable* — that is, scores must carry the same meaning regardless of which form was administered. Simply put, scores on different forms of a test should indicate the same level of performance no matter which form the test taker received. Most standardized testing programs do not use percent-correct scores as the primary scale for reporting assessment results because such scores are not comparable across forms. The raw scores are not comparable across forms either. However, they are often reported to the test takers along with the scaled scores as a direct indicator of how many points a test taker obtained from the set of questions on a test form.

“For an easier form, a test taker needs to answer slightly more questions correctly to get a particular scaled score. For a more difficult form, a test taker can get the same scaled score answering slightly fewer questions correctly.”

Table 1. Scaled Scores for Form A and Form B

Raw Score	Scaled Score	
	Form A	Form B
100	200	200
99	200	199
98	199	197
97	197	195
96	195	194
95	194	192
Etc.	Etc.	Etc.

Table 1 shows an example of scaled scores associated with different raw scores for two different forms. In this hypothetical example, Form A is the more difficult form. To achieve the same scaled score of 195, a test taker needs to answer 96 out of the 100 questions correctly on Form A, but needs to answer 97 questions correctly on Form B.

In some cases, however, percent-correct scores are used as auxiliary scores as a way of providing score users with additional information to assist in understanding their performance. In such cases, the professional guidelines used at ETS¹ and in the testing industry² as a whole typically call on testing programs to state the limitations of the percent-correct scores and provide guidelines as to the appropriate use of the test scores.

The Scaled Score

To achieve comparability, standardized testing programs report *scaled scores*. The reported scaled scores are obtained by statistically adjusting and converting raw scores onto a common scale to account for differences in difficulty across different forms. For an easier form, a test taker needs to answer slightly more questions correctly to get a particular scaled score. For a more difficult form, a test taker can get the same scaled score answering slightly fewer questions correctly. Table 1 shows an example of scaled scores associated with different raw scores for two different forms, A and B.

As illustrated in Table 1, Form A is the more difficult form. To achieve the same scaled score of 195, a test taker needs to answer 96 out of the 100 questions correctly on Form A, but needs to answer 97 questions correctly on Form B.

In order to obtain comparable scaled scores across different forms of a test, testing programs use processes known as *scaling* and *equating*. Scaling, sometimes referred to as “setting the scale,” is the process by which raw scores are transformed, either linearly

¹ ETS Standards for Quality and Fairness (ETS, 2002)

² Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 1999)

“Under most circumstances, having test takers take two forms of the same test at the same time is not practical due to issues such as increased testing time and test-taker fatigue.”

or nonlinearly, to a scale with a range of numbers that are usually different from the possible range of raw scores. The transformed scores, which are called scaled scores, are reported to the test score users. During the scaling process, the first form administered (or one form, if more than one is administered at the same administration) is considered the base form and the initial scale is set using this base form. For example, a scale of 100 to 200 could be selected as the scaled score range for a test with a possible range of raw scores from 0 to 100. Scores on all subsequent forms are placed on the same scale (100 to 200) as the base form through another process known as equating.

Equating is the process by which raw scores on a new form are adjusted to account for the differences in form difficulty from a base or *reference* form. In order to quantify and adjust for differences in form difficulty, it would be desirable to have the same group of test takers take the two forms (the new form and the reference form) at the same time. The difference in average performance on the two forms is a direct indication of the difference in form difficulty. Then, scores on the new form can be statistically adjusted to make average performances on both forms equal.

However, under most circumstances, having test takers take two forms of the same test at the same time is not practical due to issues such as increased testing time and test-taker fatigue. Another option is to have two different groups of test takers take the two forms at the same administration or at two different administrations. However, because these two groups of test takers could have different average abilities, the difference in average performance on the two forms could indicate the existence of both group ability differences and form difficulty differences.

Anchor Items

In order to isolate and quantify the difference in form difficulty, a common practice for standardized testing programs is to embed a common set of test questions, called an anchor, in both the new form and the reference form. Since both groups of test takers answer the same set of anchor questions, the difference in average performance on the anchor questions provides an indication of group ability differences. When the group ability difference is quantified and removed from the difference in average performance on the two forms, what is left in the average performance difference is an indication of the difference in form difficulty. With the difference in form difficulty identified and quantified, scores on the new form can then be statistically adjusted to remove the impact of the form difficulty difference.

Figure 1 illustrates the scaling and equating process to obtain the scaled scores for two forms. Form A is the first form and the base form on which the initial scaling is done. Form B is the second form that is equated to Form A. In this illustration, Form A is the harder form. Two steps are involved in obtaining the raw-to-scale score relationship of Form B. First, the raw scores on Form B are equated to raw scores on Form A and, as a result, higher scores on Form B correspond to lower scores on Form A. Second, the raw-to-scale relationship for Form A is applied to the equated scores on Form B. Once the scaling and equating processes are completed, scaled scores from different test forms are considered interchangeable, which means the scaled scores indicate the same levels of performance across forms of the test.

Application

Scaled scores can provide useful information about the test taker. For example, they may, if designed well:

- Facilitate meaningful comparisons of scores from test takers who took different editions of the test at different times
- Help score users to form meaningful inferences about what test takers know and can do, but discourage them from making misinterpretations and inappropriate inferences
- Allow for measurement precision while avoiding overemphasis on minor differences between points along the scale

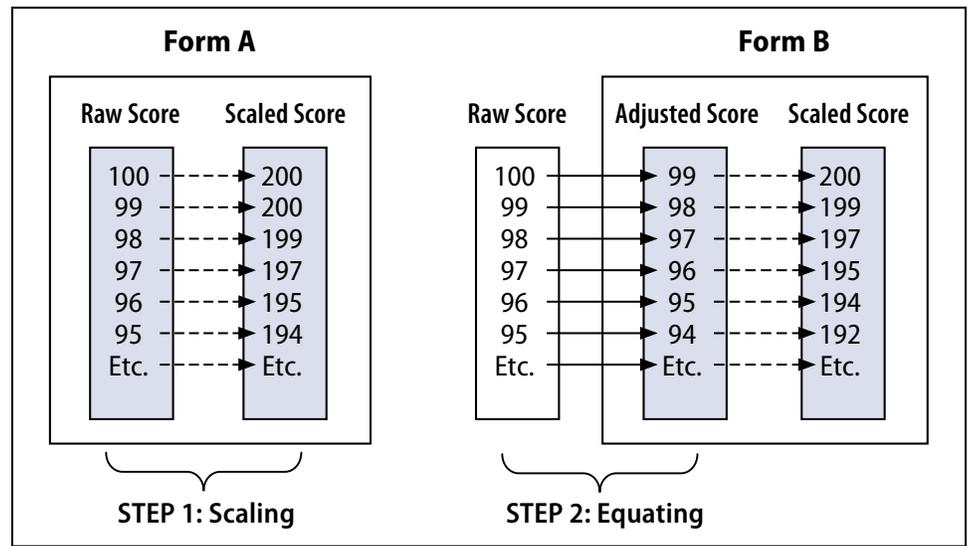


Figure 1. The scaling and equating process for obtaining scaled scores. Forms A and B are from the hypothetical example introduced in Table 1, found on page 3.

As shown in Figure 1, after the scaling and equating processes, scaled scores obtained from Form A and Form B are equivalent and interchangeable. If two test takers taking Forms A and B respectively got the same scaled score of 194 (corresponding to raw scores of 95 on Form A and 96 on Form B), we know these two test takers exhibited the same level of performance on this test. One might ask: Why not report the adjusted scores for Form B instead of the scaled scores? This is because the adjusted scores would be on the same scale as the raw scores and could be easily misinterpreted as the raw scores. Thus, the scaled scores are used and are commonly set to a range of values different from the raw score values.

Regardless of the scaling procedure used, a norm or reference group is often used, and data collected on the norm group is often reported to accompany the scale for score interpretation purposes. This initial sample of test takers is usually selected to be representative and reflect the demographics of the intended testing population (ETS, 2010). When the group is representative of the population of test takers, this allows for better interpretation of the scaled scores. The reference group's performance can serve as a benchmark against which a new group of test takers can compare their subsequent performances.

Usefulness of the Scaled Score

The utility of the scaled score comes from allowing for meaningful score interpretations and, at the same time, minimizing misinterpretations and inappropriate inferences. Test-score users frequently want additional information to assist in the interpretation of scaled scores. Providing information related to content, norm or reference groups, and precision of scores helps with the meaningful interpretation of these scores (Kolen & Brennan, 2004; Petersen, Kolen, & Hoover, 1989).

Information about the precision of the scaled scores assists test users with making appropriate interpretations based on the reported scaled scores. The number of distinct scores on the scale should allow for measurement precision but avoid encouraging too much emphasis on differences between points along the score scale (e.g., Kolen & Brennan, 2004).

For example, the scaled score can be reported in various increments (1-point increments, 5-point increments, 10-point increments, etc.). Usually, we want each additional correct answer to make a difference in the test takers' scaled score, but not such a large difference that people exaggerate its importance. The selection of a score scale, with appropriate increments, aids in the usefulness of the reported scaled scores to the test-score users.

Thus, although percent-correct scores are easy to calculate and easy to understand, they are often misinterpreted, especially in circumstances where more than one edition of a test exists. Alternatively, scaled scores should be the primary scores provided to test-score users so that reported scores are more appropriately used and correctly interpreted.

References

- American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Educational Testing Service. (2010). *Glossary of standardized testing terms*. Retrieved from http://www.ets.org/understanding_testing/glossary/
- Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.
- Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: Praeger.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). New York: Springer.
- Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan.

R&D Connections is published by

ETS Research & Development
Educational Testing Service
Rosedale Road, 19-T
Princeton, NJ 08541-0001
email: RDWeb@ets.org

Editor: Jeff Johnson
Copy Editor: Eileen Kerrigan
Layout Design: Tom Hutchinson

Visit ETS Research & Development
on the web at
www.ets.org/research

Follow ETS Research on Twitter
([@ETSresearch](https://twitter.com/ETSresearch))

Copyright © 2011 by Educational Testing Service.
All rights reserved. ETS, the ETS logo and LISTENING.
LEARNING. LEADING. are registered trademarks of
Educational Testing Service (ETS). 17284