DOCUMENT RESUME

ED 480 039 CG 032 612

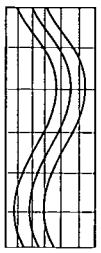
AUTHOR	Harris, Deborah J.
TITLE	Reporting and Interpreting Test Results.
PUB DATE	2003-08-00
NOTE	14p.; In: Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators; see CG 032 608.
PUB TYPE	Information Analyses (070)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Decision Making; *Educational Assessment; *Educational Testing; Evaluation Methods; *Scores; *Test Interpretation

ABSTRACT

Tests and assessments are generally administered to gather data to aid in decision making, with at an individual student level or at an aggregated level. In order to incorporate assessment data in informed decision making, test users need to understand the test results. This chapter highlights the types of test scores and test score interpretations and, specifically, the information needed to interpret test results. (Contains 10 references.) (GCP)

Reporting and Interpreting Test Results

By
Deborah J. Harris



Chapter 4
Reporting and Interpreting Test Results

Deborah J. Harris

Tests and assessments are generally administered to gather data to aid in decision making, either at an individual student level ("What math class should Kyra be placed in next year?" "Is Heru showing improvement in science this year?" "Should Jae apply for early admission to State University?") or at an aggregated level ("Has our school shown enough improvement since we adopted the new curriculum to warrant continuing it?" "What percentage of our students is meeting the new state standards, and how do we increase it?"). In order to incorporate assessment data in informed decision making, test users need to understand the test results.

Types of Test Scores

Test results are typically reported as scores, both scores for individuals and scores aggregated over individuals to obtain group averages. Just as there are many types of assessments, numerous types of scores can be reported. For most tests, the raw score is the fundamental score. Ironically, the raw score is seldom the score on which decisions are based; for many tests, it may not even be reported.

Raw scores are generally derived by counting the number of points a student obtained on the test administered. For a multiple-choice achievement test, this might be either the number of questions answered correctly or the number answered correctly adjusted for guessing. Raw scores can be useful when all students are administered the same test—as in a situation where a teacher administers a classroom test to determine whether to go on to the next science unit or spend more time on the current one—but they are generally inadequate when students take different forms of a test. Test developers try to build multiple test forms to be equivalent, but they are unlikely to be able to make the forms exactly equal in difficulty; thus, using raw scores would advantage those students receiving the easier form. (The statistical process of

equating is used to adjust for these differences when scale or derived scores are reported; see Angoff, 1971; Petersen, Kolen, & Hoover, 1989).

Although raw scores generally do not appear on score reports, sometimes percentage correct scores do. For example, the report might show the number of items answered correctly in a particular content category or skill area divided by the total number of questions in that area, to give an idea of whether the student mastered that content or skill, or needs more instruction in it. Scores typically reported include normative scores, such as percentile ranks, stanines, and normal curve equivalents. *Percentile ranks* provide an indication of how an individual's score compares to other scores by reporting the percentage of examinees in some well-defined group who earned the same or a lower score. *Stanines* are integer scores ranging from 1 to 9, with a mean of 5 and a standard deviation of 1; they are a legacy from the punch-card days, when it was desirable to have a single-digit standard score that required only one column of a punch card to record. *Normal curve equivalents* are integers ranging from 1 to 99, with a mean of 50 and a standard deviation of 21.06; they are most commonly used for Chapter 1 evaluation.

In addition to these normative scores, other derived scores may be reported. *Level, category,* or *proficiency* scores are sometimes reported, as is the case with the National Assessment for Educational Progress (NAEP), where a student may be categorized as belonging in one of four categories, such as Proficient. These scores generally have descriptors associated with them that describe what a student receiving a particular classification is likely able to do. *Developmental scores* show a student's position on a developmental continuum; an illustration is grade equivalents, which try to establish a score scale that ranges across multiple grade levels, thus facilitating the tracking of a student over time. Grade equivalents appeal to teachers and parents, who seem to have an intuitive understanding of what they mean. There are potential problems with interpreting grade equivalents, especially extreme scores, such as when a third grader receives a grade equivalent of 8.2, but parents and educators seem pretty savvy about not over-interpreting these results in practice.

Often a test developer creates an original score scale for an assessment, either building in some normative meaning at the time the scale is developed or building it to have particular properties. For example, SAT scores are reported on a scale from 200 to 800, originally scaled to have a mean of 500 and a standard deviation of 100 on a

particular sample of examinees. ACT Assessment scores are reported on a scale of 1 to 36, which was developed to try to equalize error variability along the score scale. Both of these scales have developed additional interpretations over time, such as what scores may indicate a student is ready for initial placement into a standard English composition course at a particular college.

Many tests report multiple scores. For example, the Iowa Tests of Basic Skills provides a raw score; a developmental standard score, intended to indicate the student's location on an achievement continuum; a grade equivalent, which also indicates the student's location on an achievement continuum, but one with equal rates of yearly growth between each pair of grades; national and local percentile ranks; stanines; and a normal curve equivalent score (See Hoover et al, 2001, pp. 13–14).

Petersen, Kolen, and Hoover (1989), and Angoff (1971) provide extensive discussions of creating and maintaining score scales including primary and auxiliary score scales, linear and nonlinear transformations of raw to scale scores, and methods of incorporating additional information into a score scale when developing it. For example, Petersen, Kolen and Hoover provide an example of creating a score scale to provide content meaning (where a particular score is interpreted as an indication of what a student knows or can do) or normative meaning (an example might be a grade equivalent, where a score is interpreted relative to what a typical student at that grade can do), and of incorporating score precision information into scores. The increased use and capabilities of computers in recent years has led to many technical and sophisticated types of scores, particularly those based on item response theory and computer-based testing (see Thissen & Wainer, 2001). Mehrens and Lehmann (1991) also provide examples and discussion of several types of reported scores.

Types of Test Score Interpretations

There are two basic types of score interpretations: norm-referenced and criterion-referenced.

Norm-referenced interpretations provide meaning by comparing a student's performance to that of a well-defined group of examinees, such as a nationally representative sample of fifth graders from public and private schools in the United States. How informative the comparison is depends in part on how representative the norm group is, how relevant it is to the comparison one is interested in making, and

how recently the data were gathered. Other issues also come into play, such as how motivated the examinees in the norm group were, whether the data were gathered under standardized conditions, and whether the norms are empirical, versus interpolated or extrapolated from other data. For example, if one is interested in being selected for a special program for which there are limited slots, one is probably most interested in comparing one's score to the scores of other applicants. A comparison with the general public may be of less interest and relevance. Percentile ranks are easy to identify as norm-referenced scores. The nature of other scores, such as grade equivalents, may be harder to identify. For example, is a particular grade equivalent established using empirical data or using judgmental methods? It is important to remember that norm-referenced interpretations indicate how students actually performed, not how they *should* perform. A student's norm-referenced scores indicate simply how the student scored compared to other students, not whether the student is functioning at an acceptable level.

To address the issue of performance quality, *criterion-referenced* interpretations provide score information based on a set of criteria, generally skills or knowledge. Such a score represents what a student knows or can do. An example would be a score from a writing assessment that is linked to a rubric detailing what skills a student receiving that score has demonstrated and failed to demonstrate (e.g., "Used strong voice"; "Lack of subject-verb agreement").

The difficulty in developing criterion-referenced interpretations is to define clearly the domain or skill. If the ability to add two single-digit non-negative integers is the skill of interest, one could write out all the possible problems (i.e., 0 + 0; 0 + 1; 1 + 0 . . . 9 + 9), randomly select some to be placed on a test, and use the percentage correct on the test as an estimate of the percentage of the entire domain the student knows. Other skills and content areas are much more difficult to define accurately, however; consider "appreciates literature," "demonstrates appropriate grammar," or "understands Shakespeare's tragedies."

Rarely is a test score interpretation purely norm-referenced or purely criterion- referenced. That is, generally one is not interested in a normative comparison without addressing content, nor is one interested in criterion-referenced interpretations without knowing what reasonable expectations are. For example, a parent of a young child is interested in assessing both whether the child can read successfully (criterion-referenced) and whether the child is progressing in line with his or her peers (norm-referenced).

Information Needed to Interpret Test Results

A test developer has the responsibility to inform a test user of the characteristics of the test, such as content specifications, reliability, validity for particular score uses, and how score scales are developed. The onus is also on the test developer to describe how the test user should use the scores. The test user is responsible for adhering to the cautions, qualifiers, and limitations provided by the test developer. The test user is also responsible for following the administration conditions and for maintaining the integrity of the test. For example, if a test user ignores instructions not to allow calculators, does not time the test as instructed, or allows students to work collaboratively when the directions forbid it, the scores reported for the user's students will not be comparable to scores obtained when the instructions were followed. This will affect both norm-referenced and criterion-referenced interpretations.

The test developer needs to provide good descriptions of the norm group for any normative scores, so the test user will be able to determine if the normative comparison is appropriate for his or her test takers. The test developer also needs to describe how domains and skills were defined, how levels were established (if level scores are reported), and how score scales were developed for scale scores. The test developer should also provide information regarding how accurate scores are likely to be, either as classification consistencies or conditional standard errors of measurement, as well as reliabilities.

The test developer and the test user share responsibility for providing validity evidence for particular score uses. Whereas the test developer is responsible for providing evidence for any uses he or she recommends, the test user is responsible either for ensuring that his or her specific use of the test is encompassed by the test developer's recommendations or for providing additional validity evidence for the specific use. Test developers are also responsible for cautioning test users against likely misinterpretations of test results—such as taking a percentile rank table developed for use with individual student scores and using it to try to find a percentile rank for an entire school, based on an aggregated school mean score.

Test scores should never be interpreted in a vacuum, but instead considered in light of pertinent factors: how the scores are computed, who the norm groups consist of, the test content, whether the test is speeded, the administration conditions, the standard error of measurement, and so on. The type of decision to be made also influences

how the test score is interpreted: The same score earned by two very disparate students might be interpreted differently; for example, as exceptional progress for one and average progress for the other.

Interpreting Results from a Modified Test

There are excellent reasons for modifying an existing test to accommodate practical considerations of assessment or for a particular goal, as in the following examples:

- changing the administration conditions to allow a student with visual impairment the use of a reader;
- extending the time limits for a student who works unusually slowly (for example, a student with a hand in a cast);
- changing the mode of delivery by allowing a test to be delivered on a computer or permitting the use of calculators on a mathematical reasoning test;
- eliminating some items to decrease the amount of time spent away from classroom instruction; or
- translating the test into a different language to allow students with limited English proficiency to take it in their native language.

Any or all of these modifications may improve the validity of the assessment scores for the particular use the test user has in mind. That is, a math test given in Spanish may be a more valid measure of math ability for a particular student than a math test given in English. However, test scores that have been derived based on standard conditions must be interpreted with caution when those conditions have been altered. This applies to normative scores when the standardized conditions under which the norms were obtained are altered, and to derived scores when the raw-to-scale-score conversions were obtained under standardized conditions. Context effects have been found to affect test scores in ways that appear unpredictable, and therefore caution must be exercised in interpreting scores from a test that has been altered in any way. For example, switching the order of the tests in a battery to accommodate a school lunch schedule may or may not affect the test scores. Small differences—such as changing the order of administration of the tests, or deleting some items and modifying the time limits accordingly—have been shown to have unanticipated effects. It is wise to err on the side of caution when using data from modified tests for decision making. As the *Standards for Educational and Psychological*

Testing (AERA, APA, & NCME, 1999, p. 61) state, "Although accommodations are made with the intent of maintaining score comparability, the extent to which that is possible may not be known."

Rationale and Procedures for Setting Performance Standards

Numerous procedures exist for setting performance standards (see, e.g., Cizek, 2001), but the Angoff method is perhaps the most widely used (Angoff, 1971). This method requires a group of trained panelists to estimate the probability that a "minimally acceptable person" would answer items on the test correctly. Generally, the first step in an educational setting is to develop narrative descriptors of what content a student at each level should know and what skills the student should possess. The second step is to select panelists to participate in the process. Next, the panelists are trained in internalizing the descriptors. This is an extremely important step, as panelists cannot be expected to determine how a Basic-level student would perform on a given item if they do not really understand what "Basic" means.

Once panelists understand what skills or knowledge is typical of a category, they are asked to picture a student who minimally meets that category of requirements and to judge how this student would respond to items on a test. For example, what is the probability that a minimally Basic student would get a particular item correct? These probabilities are then averaged across panelists and across items to arrive at a cutoff score for the Basic category.

The setting of performance standards is generally an iterative process in which panelists receive feedback data, which might include other panelists' ratings; empirical data on how students actually performed on the test items; and impact data, or what percentage of students in a particular group would be classified in each category based on the proposed cutoff scores. Setting performance levels is a very complicated procedure calling for a great deal of judgment. Decisions regarding the selection and training of the panelists, the number of rounds of ratings to hold, how to derive the ratings themselves, what feedback to provide to panelists, and others all require human judgment. Cizek (2001), Green (1996), and Hansche (1998) provide a great deal of additional detail for the interested reader.

Information Needed to Interpret Test Scores Correctly

To use test scores as one piece of data in making a well-informed decision, the test user must be clear on what the test scores mean. The most important consideration is the test content: what knowledge and skills are being tested, and how they are being tested. For example, does a reading comprehension test use novel material or material a student would likely have seen before? Is the test administered under somewhat hurried conditions, or would almost all students have enough time to complete the assessment? For a math test, are calculators allowed or prohibited?

Knowing what is tested is the most important aspect of interpreting test scores, but it is by no means the only consideration. Many scores (such as percentile ranks) are derived using a norm group. In order to interpret these scores accurately, one must be knowledgeable regarding the conditions under which the data were collected. Was it an operational administration or a special study where the examinees were unlikely to be motivated? How were the data edited? Other important information concerns score precision: how accurate a score is likely to be. For example, if an examinee is reported as being Proficient, how likely is the examinee to be classified as Proficient again, if the same test, or an alternate form of the test, were administered a second time? Most tests are accompanied by some type of reliability or score precision information, but the user must discern whether the information provided is relevant to his or her needs. For example, knowing the internal consistency of a test may not be of as much interest as knowing the classification consistency for a particular score use. Some test score scales, such as the ACT Assessment 1 to 36 scale, have tried to incorporate score precision information into the actual score scale, or to report scores as bands instead of single points, to illustrate measurement error.

Test users need continually to remember that the norms provided with test results are not standards that students must achieve. Not all students will score above the median for a test; not all students will show one year's growth in 12 months' time on a particular score scale. Many scores do not have equal units, meaning that progressing from one score point to another will indicate different amounts of change in different parts of the score scale.

A final point to remember is that scores from one test are not necessarily comparable to scores on another test, even if both scores are termed "grade equivalents" or "national percentile ranks." Different

test developers use different norming samples and calculate grade equivalents using different methodologies. Different tests also generally cover different content, have different administration conditions, and are scored and scaled in different ways. One must be cautious when trying to compare scores from different tests.

Timely Provision of Test Results

Tests are administered to obtain data to inform decision making. Therefore, it is important to obtain those data in a timely manner, before the decisions need to be implemented or the data become so dated they are no longer of value. If a student takes a college entrance exam, the results are needed quickly enough to allow the student time to consider the results, in conjunction with other information, and decide whether or not to apply to a particular college prior to the application deadline. How quickly results are needed will depend on particular circumstances, but sooner is better than later. Faster scanners, electronic scoring, computerized score reports, and electronic delivery of score reports all have the potential to speed up the delivery of test results without sacrificing quality.

An additional consideration is who receives test result information. For many educational decisions, there are numerous stakeholders: students, parents, teachers, counselors, administrators, and the public in general. Who receives test score information, and of what type, depends on legal, confidentiality, practical, and situational factors. For example, a young child may not be capable of understanding what a particular derived score means so need not be given that information; a school board may receive aggregate score information for the district but not the particular score information for Pat Smith.

Basing Decisions on Multiple Sources of Information

Because no single test is likely to be comprehensive enough to encompass all the content one is interested in assessing, or to be reliable enough to measure a student's true ability without error, it is important to rely on multiple measures when making decisions, particularly if the decisions are virtually irreversible, long term, or high stakes. The *Standards for Educational and Psychological Testing* (AERA et al., 1999, p. 146) makes this explicit, stating in Standard 13.7: "In educational settings, a decision or characterization that will have major impact on a student should not be made on the basis of a single test

score."

Most test developers appear to be in agreement with the standards on this point, also cautioning test users not to rely on a single measure when making a decision. For example, the interpretive guide for the *Iowa Tests of Educational Development* (University of Iowa, 1994, p. 95) cautions, "Throughout this *Guide,* stress has been placed on the necessity of interpreting test results in relation to other available information about students. Any profile of test scores either for an individual student or for a group of students can be misleading if considered without regard to other factors such as classroom performance, interests, expectations, and aspirations." Teachers, counselors, administrators, parents, and the students themselves have knowledge that cannot be obtained from a test score. Likewise, test scores provide information not readily obtainable from other sources. Pieces of knowledge pooled from multiple sources augment each other, and the result is more complete information on which to base a decision.

Some Final Aspects of Test Score Interpretation

For some uses, one is not interested just in the scores from a particular test, but instead wishes to compare scores across different forms of the test. For example, one may wish to compare scores for this year's fourth graders with those from last year, or one may wish to test students before and after they receive an intervention. In order to make these types of comparisons, it is necessary that the scores on the different forms be comparable, usually through a statistical adjustment called *equating.* Test developers who offer different forms of a test should discuss how they ensure that scores from the different forms may be used interchangeably.

In addition to investigating the technical characteristics of the test scores that are reported, the test user needs to ensure the integrity of the scores obtained by the students. This requires adhering to the administration conditions prescribed by the test developer (e.g., regarding timing or use of calculators and dictionaries), and preventing examinees from obtaining inappropriate scores through fraudulent means (such as copying). It also means attempting to motivate students to try their best on the test.

Tests are given to obtain data to inform educational decisions. To the extent that the test user understands the scores from those tests, and the scores are appropriate for the decisions he or she is attempting to make, the decisions will be well informed. By relying on test scores in

conjunction with additional information; by ensuring that the test developer has provided complete information regarding how the tests were constructed, how the score scales and norms were developed, and how the scores should be used; and by becoming familiar with all this information, the test user becomes able to make better educational decisions using test results.

References

AERA, APA, & NCME [American Educational Research Association, American Psychological Association, & National Council on Measurement in Education]. (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. I. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Cizek, G. J. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Green, B. F. (1996, Nov. 6). *Setting performance standards: Content, goals, and individual differences.* Paper presented at the second annual William H. Angoff Memorial Lecture, Princeton, NJ, Educational Testing Service.

Hansche, L. N. (with Hambleton, R. K., Mills, C. N., Jaeger, R. M., & Redfield, D.). (1998). *Handbook for the development of performance standards: Meeting the requirements of Title I.* Washington, DC: U.S. Department of Education and the Council of Chief State School Officers.

Hoover, H. D., Dunbar, S. B., Frisbie, D. A., Oberley, K. R., Bray, G. B., Naylor, R. J., et al. (2001). *Iowa Tests of Basic Skills Complete/ Core Battery: Spring norms and score conversions with technical information.* Form A, Levels 5–14. Itaska, IL: Riverside Publishing.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). Fort Worth, TX: Holt, Rinehart and Winston.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., p. 262). New York: Macmillan.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring.* Mahwah, NJ: Lawrence Erlbaum Associates.

University of Iowa. (1994). *Iowa Tests of Educational Development: Interpretive guide for school administrators.* Forms K and L, Levels 15, 16, and 17/18. Chicago, IL: Riverside.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
X	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").