Evaluating Psychological and Psychoeducational Tests

Melissa A. Messer, MHS

PAR®

Executive Summary

There are countless psychological and psychoeducational tests on the market, available from test publishing companies, published in journal articles, or in the public domain. Professionals are responsible for critically evaluating every test before using it for clinical decision making to ensure it meets the necessary requirements. This white paper presents a framework that can be used when evaluating assessments. A fillable form that accompanies this white paper is available for download and may be reproduced for future use by individuals evaluating tests.

An assessment often provides a way to better understand a person in order to make informed decisions. More specifically, educational and psychological assessments aid in gathering important information that have been developed to be effective at measuring a particular trait, behavior, disorder, skill, etc. A standardized psychological test is a task or set of tasks given under standard, set conditions. It is designed to assess some aspect of a person's knowledge, skill, or personality. The tests that will be discussed here are normed (or standardized) tests. Norm-referenced tests are designed to compare and rank test takers in relation to each another. Norm-referenced tests report how test takers performed relative someone similar to the test taker (e.g., comparing them to others their age, education, sex).

Educational and psychological assessments aid in gathering important information that have been developed to be effective at measuring a particular trait, behavior, disorder, skill, etc.

Reliable and valid tests should have well-documented evidence of their development as well as theoretical and technical underpinnings. There are countless tests available from test publishing companies, published in journal articles, or in the public domain. However, it is important for professionals to keep in mind that ***not all tests are created equal***. It is the responsibility of the professional to critically evaluate every test before using it for clinical decision making to ensure it meets the necessary requirements. There are several standards that should be considered. The following is a framework to use when evaluating assessments. A version of this framework may also be downloaded and reproduced for future use.

Considerations in Test Evaluation

General Information

The information needed to fulfill the criteria below can typically be found on a publisher's website, in a test's professional manual, or in a test review (additional resources for finding this information are provided at the end of this white paper). This is important information to have access to and become familiar with prior to deciding to use a test in your practice.

Title of test: The title of the test is important; however, the version or edition of the test is just as significant. Standard 9.08 of the *APA Ethical Principles of Psychologists and Code of Conduct* indicates that professionals should not base their recommendations on tests and measures that are obsolete (American Psychological Association, 2017).

Author(s): It can be helpful to be aware of who the authors are when doing additional research on the measure.

Publisher: There are several reasons why knowing the publisher is important. First, many publishers distribute other publishing company's proprietary measures. However, if you have specific questions about an assessment, it is likely best obtained from the measure's publisher as opposed to a distributor. Additionally, digital versions of an assessment are typically only accessible from the publisher. To find the publisher of a measure, you can check the title page. This information is frequently printed on either the front or back cover, as well.

Publication year: Just like noting the version of the assessment, it is important to note the publication year. An older publication date does not always imply that a test is outdated. Professionals need to evaluate if the year of publication is in line with the test content. For example, if a test was designed to align with *DSM* criteria, and those criteria change, then a test could be considered outdated. However, a performance task of abstract reasoning could be 10 years old and not require updating. Additionally, publication dates may vary for test items, normative data, forms, and supplementary materials. It is important to understand the implications of these dates for each of the components.

Time required to administer: Administration times for psychological or educational testing vary. Age and attention span of the examinee should be important considerations. Professionals should consider the impact of lengthy testing times (e.g., it may be difficult for a young child to be administered a 4-hour test).

Qualification level: In accordance with the *Standards for Educational and Psychological Testing* (2014), test publishers provide qualification guidelines that limit who may purchase and administer the test. Eligibility to purchase an assessment product is typically based on training, education, and experience in the field. Most tests and materials are only available to those professionals who are appropriately trained to administer, score, and interpret psychological tests. Each publisher has its own process in place for establishing a consumer's qualification level. To establish a qualification level, a consumer must complete the application process with each publisher prior to purchase.

Cost: Cost is an important factor to consider. Publishers may offer a variety of pricing and packaging options. Users may find discounted pricing is offered for research or training purposes. For example, PAR offers discounts on certain products to instructors and graduate students through the University Partnership Program (UPP).

Access to training materials: New tests may be challenging to learn. The professional manual is often the best place to obtain the information needed to administer and score a test. However, additional training materials may be available. For example, PAR maintains a Training Portal that offers free, on-demand courses on select products. PAR also offers additional resources such as white papers, PowerPoint presentations, bibliographies, and educational materials designed for training programs.

Type of test: Within the category of norm-referenced tests, there are generally two types of formats: rating scales (or questionnaires) and performance-based tests (e.g., intelligence or achievement tests). Performance-based tests typically require much more training and practice to administer compared to rating scales.

Item characteristics: A variety of different item types exists across tests. Rating scales commonly include an item stem with possible responses. Some variation includes selecting one statement from a list of statements or selecting between two statements. Some important considerations are the reading level of the items and the response options (discussed in more detail below). Performance-based tests tend to have more variability in item types; this is typically related to the content of the test. For example, a test examines phonological awareness will likely have item types that require the examinee to listen to words read by the examiner and produce a rhyming word. An achievement test, however, may include math reasoning items that are read aloud by the examiner.

Test response format: Tests are available with many different response formats, such as multiple choice, true/false, observation, open-ended verbal response, written responses, and more. Again, some formats require more training to administer. For example, a spelling task requires the examiner study the list in advance to ensure they are able to correctly pronounce each of the words during administration. For rating scales, the response options can sometimes be confusing for examinees and may need explanation. Examiners should be able to explain these options to test takers as needed.

Population for whom it was designed: Who is the intended respondent or examinee? What is the appropriate age? What is the reading level? For normed-reference tests, these considerations are imperative because the scores compared to clearly defined normative groups.

Nature of the content: Some tests cover a wide breadth of constructs, whereas others are designed to home in on a single construct. Examples of test content include spatial perceptual skills, reading fluency, internalizing behavior, verbal memory, executive functioning, etc.

Subtests and separate scores: As mentioned earlier, not all tests are created equal. Two different tests may purport to measure the same construct; however, they may have very different subtests and scores to do so.

Qualitative features: This includes more subjective concepts, such as the design and ease of use of the test booklets, attractiveness, appropriateness for intended test taker, face validity, etc.

Ease of scoring and administration: Professionals should evaluate challenges with administration and scoring, what format options are available (digital vs. print), and the time and complexity involved in scoring.

Multicultural considerations: The *Standards* emphasizes fairness in all aspects of testing across diverse populations and contexts. According to the *Standards*, there are four general areas to consider:

- Test bias
- Equitable treatment in testing process
- Equality in outcomes of testing
- Opportunity to learn

Technical Evaluation

The technical evaluation of a test is critical in deciding if it is appropriate for use. The *Standards* provide specific test design and development criteria that should be used when developing a test, including principles for test specifications, item development, test administration, scoring procedures, test materials, and test revisions. Below is a high-level overview of the technical characteristics that should be considered. For more comprehensive information, refer to the *Standards*.

Test development methodology: Each test is unique and, therefore, requires a different set of criteria used during development. However, there are some commonalities across most tests in terms of best practices. For example:

- Item development
  - The procedures used to develop, review, and try out items, as well as how the final items were selected, should be documented.
  - Use of external reviewers as well as statistical methods should be utilized to determine the final item set.
- Administration
  - The administration instructions used to standardize the test should be clearly documented and easily accessible for use when administering the test.
  - Allowable deviations from test administration should be described.
- Scoring
  - The development of scoring criteria should be documented and detailed instructions on how to derive scores should be clear, this is especially important with extended-response item types (e.g., essays).

Normative data: A norm-referenced test is based on a set of normative groups. The demographic characteristics of the normative groups as well as the size and representativeness of each group should be scrutinized. Most well-developed tests will be matched to the current U.S. Census proportions with the goal of representation across multiple characteristics (e.g., age, sex, education, race/ethnicity, region).

Type of scores: Standardized tests should include standardized scores (e.g., standard scores, scaled scores, *t* scores, percentiles).

Reliability and validity: According to the *Standards*, test developers and publishers should document steps taken during the design and development process to provide evidence of reliability and validity.

- ***Reliability:*** There are multiple methods for establishing evidence of and assessing the reliability of a test. It is important to evaluate the methods used to assess reliability and determine the level of precision needed for the types of decisions that are being made. Reliability analyses include internal consistency, standard error of measurement, split-half, test-retest, and interrater reliability.
- ***Validity:*** A valid test is one that accurately measures the psychological construct for which it is intended. Test validity is multidimensional in nature and should be evaluated using a variety of different sources and methodologies, each providing unique evidence that supports the validity of the test. Sources of validity evidence include evidence based on test content, evidence based on theoretical constructs, evidence based on internal structure, and evidence related to other variables.

You can access a fillable template that contains the above criteria by visiting this link.

Resources

There are several sources available when evaluating psychological or educational tests. It is often a red flag if any of the information included in the evaluation above is not readily available to test users. Examples of resources include:

- Professional/technical manual
- Publisher content (e.g., customer services, training materials, website)
- Test reviews (e.g., Mental Measurement Yearbook/ Buros, journal articles)
- Books:

  Sherman, M. S., Tang, J. E., & Hrabok, M. (2020). *A Compendium of Neuropsychological Tests* (4th ed.). Oxford University Press.

  Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological Assessment* (5th ed.). Oxford University Press.

  Sattler, J. M. (2018). *Assessment of Children: Cognitive Foundations and Applications* (6th ed.). Jerome M. Sattler Publisher.

  Stoltz, K. B., & Barclay, S. R. (2019). *A Comprehensive Guide to Career Assessment* (7th ed.). National Career Development Association.

Additional Readings and Resources

Groth-Marnat, G. & Wright, J. A. (2016). *Handbook of Psychological Assessment* (6th ed.). Wiley.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice Hall.

Essentials of Psychological Assessment series from Wiley.

Council of National Psychological Associations for the Advancement of Ethnic Minority Interests. (2016). *Testing and assessment with persons and communities of color*. Washington, DC: American Psychological Association. Retrieved February 14, 2020, from https://www.apa.org/pi/oema.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Psychological Association. (2017). Ethical principles of psychologists and code of conduct (2002, Amended June 1, 2010 and January 1, 2017). Retrieved February 14, 2020, from https://www.apa.org/ethics/code.

Melissa A. Messer, MHS

Director of
Product Development
mmesser@parinc.com
1.800.331.8378

Melissa A. Messer, MHS, is the Director of Product Development. She oversees all new product acquisitions and development and manages a team of project directors and quality control specialists. Melissa received her bachelor's and master's degrees from the University of Florida.

Melissa joined PAR in 2002. During her time at PAR, she has worked on the development of more than 30 products. Melissa is the author of several PAR products, including the Academic Achievement Battery (AAB) series, the Working Styles Assessment (WSA), and the Work Values Inventory (WVI). In addition, she has coauthored multiple editions of John Holland's Self-Directed Search (SDS).

PAR • 16204 N. Florida Ave. • Lutz, FL 33549 • 1.800.331.8378 • www.parinc.com

To cite this document, use:
Messer, M. A. (2020). *Evaluating psychological and psychoeducational tests* (white paper). Lutz, FL: PAR.