11-14-2015

Strategies and Resources to Enhance Test Evaluation and Selection

Janet F. Carlson

Nancy Anderson

Strategies and Resources to Enhance Test Evaluation and Selection

Janet F. Carlson & Nancy Anderson | Buros Center for Testing | University of Nebraska - Lincoln

Introduction

Testing serves an important function for SLPs in offering an evidence base that is useful in screening, diagnosing, monitoring progress, and documenting outcomes. Tests are used to measure diverse constructs such as communication, literacy, oral and written language, receptive and expressive vocabulary, articulation, phonological awareness and processing, and auditory perception and processing. In addition, specific impairments may require specialized measures to evaluate conditions such as stuttering and orthographic competence.

When using tests to diagnose language impairments, Betz, Eickhoff, and Sullivan (2013) suggest that SLPs consider carefully a test's psychometric properties, particularly because of the "increasing emphasis on evidence-based practice, specifically, the requirement to validate clinical decisions regarding assessment and treatment" (p. 142). Kirk and Vigeland (2014) echo these sentiments in stating, "It would be helpful to have evidence-based practice guidelines that provide recommendations for determining the psychometric adequacy of norm-referenced tests" (p. 375). At the 2014 ASHA conference, Pavello and Ireland reviewed psychometric and other considerations that influence test selection.

For nearly 80 years, the Buros Center has published independent reviews of commercially available English language tests in its Mental Measurements Yearbook (MMY) series, currently in its nineteenth volume (Carlson, Geisinger, & Jonson, 2014). Each volume provides reviews of tests across a wide range of categories, including Language and Speech & Hearing. Cizek, Koons, and Rosenberg (2011) observed that "the MMY series is unique in that it serves as an independent source of evaluations of specific tests . . . [and] is widely considered to be the most accurate, complete, and authoritative source of information about published tests" (p. 123). However, as the MMY editors carefully note, "The [MMY] series was developed to stimulate critical thinking and assist in the selection of the best available test for a given purpose, not to promote the passive acceptance of reviewer judgment" (Carlson et al., 2014, p. xiii). In a similar vein, Thorndike (1999) advised that MMY reviews "must be supplemented by a thorough knowledge of the situation for which a test is desired and by mature professional judgment on the part of the prospective user" (p. 50).

This poster session presents a framework for test evaluation and selection to inform decisions about standardized tests used by SLPs within their practices or research. Examples of reviews of tests illustrate best practices in test evaluation. Resources to promote the application of critical thinking skills to test evaluation and selection are provided.

Financial Disclosures

Janet Carlson and Nancy Anderson are full-time salaried employees of the Buros Center for Testing. These relationships are ongoing. They have been employed by Buros for about 5 and 3 years, respectively. Each had prior part-time experience at the Buros Center. The Center receives royalties from sales of its publications, which provide reviews of and information about commercially available tests. Royalties accrue to print volumes and electronic subscriptions. In addition, revenues are generated from direct sales of test reviews and information through an e-commerce site.

Non-financial Disclosures

Janet Carlson and Nancy Anderson are Associate Director and Managing Editor, respectively, at the Buros Center for Testing. These relationships are ongoing. Both Carlson and Anderson are members of the editorial team that edits test reviews submitted for publication in the Mental Measurements Yearbook series. In addition to this series, the Buros Center also publishes the Tests in Print series, and Pruebas Publicadas en Espanol. Janet also serves as a manuscript reviewer for a few journals; some submissions involve specific tests or assessment practices. She has authored about 20 test reviews, most of them for the Buros Center for Testing and all of them before 2006.

Excerpts From Test Reviews

DESCRIPTION

Excerpted from Hutchins & Cannizzaro (in press)

Describes test in general terms; explains test purposes

Identifies target population and method of administration

Provides detail about conducting an evaluation

May compare current edition with previous edition(s)

Describes how test is organized

Describes process and time needed for administration

Indicates types of scores and reports available

The Clinical Evaluation of Language Fundamentals–Fifth Edition (CELF-5) is offered as a norm- and criterion-referenced measure that can aid in the screening and identification of language disorders. It is also intended as a tool to guide curricular modifications and treatment planning. The CELF-5 is an individually administered assessment intended for individuals 5–21 years of age.

The CELF-5 proposes a stepwise evaluation process that reflects best practices in education for identifying struggling students as well as potential classroom supports. The suggested process begins with the Observational Rating Scale (ORS), which guides the observation of naturalistic speech, language, and communication behaviors. ORS results can then be used to initiate individualized instruction and guide decisions about which portions of the CELF-5 would yield the most useful information for a particular student.

Changes to the content of the fifth edition are extensive and include new, revised, and expanded sections for the assessment of reading, writing, and pragmatic abilities. The CELF-5 includes updated content and revised normative data, and growth scale scores have been added that can be used to document student progress over time. A number of sections found in the previous version (i.e., the CELF-4) have been removed based on customer feedback and to help focus the overall scope of the battery. Specifically, the Expressive Vocabulary, Familiar Sequences, Number Repetition, Phonological Awareness, Word Associations, and Rapid Automatic Naming tests have been eliminated.

Subsections of the CELF-5 are now referred to as tests (as opposed to subtests) and have been expanded to strengthen the floors and raise ceilings (particularly in the 13-21 age group). The tests vary depending on the student's age (i.e., between 5–8 years or 9–21 years). ...

The CELF-5 encourages flexible and individualized starting points and uses clear discontinuation rules. Tests that make up the CELF-5 can be administered in five to 15 minutes. Total testing time will depend on the number and type of tests ultimately employed (a table with mean administration time by test and child age is presented in the examiner's manual for assessment planning purposes). The average time to administer the core tests (i.e., those that make up the Core Language Index) is 34 minutes for students ages 5–8 and 42 minutes for students 9–21.

The CELF-5 yields growth scale values, age equivalents, percentile ranks, normal curve equivalents, stanines, and standard (scaled) test (M = 10, SD = 3) and index (M = 100, SD = 15) scores. For the purpose of identifying language disorders, three or four tests (depending on the age group) contribute to the Core Language Score. Other index (i.e., composite) scores can be calculated including the Receptive Language Index, the Expressive Language Index, the Language Content Index, the Language Structure Index (ages 5-8 only), and the Language Memory Index (ages 9–21 only). Difference scores (i.e., Discrepancy Comparisons, discussed below) for the Receptive-Expressive Language Index and other major indices also are provided.

CELF-5 test components include an examiner's manual, a technical manual, two stimulus books, the Observational Rating Scale, two record forms (ages 5–8 and 9–21), and two reading and writing supplements (ages 8–10 and 11–21). First-time users will need some time and practice with the CELF-5 to ensure correct administration, scoring, and interpretation. Fortunately, stimulus materials and score forms are well designed to facilitate fluid administration. For example, stimulus books (which employ color illustrations depicting characters from a variety of ethnic backgrounds) use an easel with stimuli presented on one side and examiner prompts on the other. In addition, score forms have primers for each test to remind the examiner of key procedures (e.g., starting and stopping rules, repetitions allowed or not allowed). The examiner's manual is well-organized and clearly written, and several instructive case studies are offered as examples. ...

DEVELOPMENT

Excerpted from Spenciner (in press)

Offers cautionary notes for prospective users

Discusses underlying assumptions or theory used to define the construct to be measured

Provides details on item development

Discusses pilot testing and selection of final items

Describes theoretical reasoning and technical procedures used to craft the instrument

Reviews steps taken to evaluate the appropriateness of selected items

The Vocabulary Assessment Scales–Expressive and Vocabulary Assessment Scales–Receptive were published in 2013, following an extensive development process. This process is described in detail in the professional manual. Briefly, the test author developed a market research survey and assembled an initial expert review team to identify the key features of a picture vocabulary test and to identify improvements that could be made to existing vocabulary assessments. The individuals who served as expert reviewers and their affiliations are listed in Appendix A of the professional manual. Using the information gathered, the test author developed a list of potential words that represented the following categories: actions and activities; descriptors and numbers; animals; body parts; buildings, art, and architecture; clothing and accessories; foods; geographic scenes; household objects; musical instruments; shapes and symbols; plants; tools; vehicles; people and workers; books and money; and toys and recreation. Colored photos were matched to the word list for both expressive and receptive vocabulary.

The development process also included the selection of foils; details about how foils were chosen are included in the professional manual (p. 28). In addition, the new assessments also underwent local and national pilot testing, additional expert reviews, and an item analysis that provided statistical estimates (p value and Rasch difficulty parameter) regarding how difficult a test item was for examinees. Test items were also reviewed for possible bias by individuals who participated in a bias review panel. Their affiliations are listed in Appendix B of the professional manual. ...

For RESOURCES and REFERENCES, please see handouts

TECHNICAL

Excerpted from Moyle & Long (in press)

Focuses on three main points: standardization and/or norms, reliability, and validity evidence using documents and statistical data provided by test publisher

Presents information about norm sample; evaluates how well the sample matches intended population

Describes evidence of score consistency

Presents types of reliability estimates and their magnitudes

Evaluates test content and adequacy of test for measuring intended construct

Evaluates differential validity across gender, racial, ethnic, and cultural groups

The TOLD-P:4 was normed on 1,108 children in 2006 and 2007 from 16 states across four major regions of the United States. The demographics of the sample closely resemble those of the 2005 U.S. school-aged population of children with regard to gender, geographic region, race, Hispanic ethnicity, exceptionality status, family income, and education level of parents. The test manual does not specify whether any children in the sample spoke nonmainstream dialects or were English language learners. Thirteen percent of the sample consisted of children with disabilities. ...

The test authors present three types of evidence to demonstrate reliability of the TOLD-P:4—coefficient alpha, test-retest, and scorer differences (i.e., interscorer reliability). Coefficient alpha was used to evaluate content sampling error, or internal consistency reliability of the test. The test authors consider coefficients of .80 to be minimally reliable and .90 or higher to be most desirable. The average coefficients across ages exceeded .80 for the nine subtests, with seven exceeding .90. The average coefficients for the composites were .90 or greater. This evidence suggests that the TOLD-P:4 is internally consistent. Test-retest reliability was examined using a sample of 89 children who were retested one to two weeks after initial testing. Overall, the vast majority of correlation coefficients fell between .80 and .90, suggesting acceptable test-retest reliability. The reliability of scorer differences was examined by having two staff members from the test publisher independently score 50 test protocols drawn from the normative sample. Correlation coefficients ranged from .97 to .99 for the subtests and composites. This method does not appear to be a rigorous test of interscorer reliability given that the raters did not record children's responses during the actual testing process.

Three types of validity evidence are presented in the test manual: Content-description, criterion-prediction, and construct-identification. The test authors first provide qualitative evidence in support of content-description validity. Next, the results of conventional item analyses are presented. Both item discrimination (i.e., the degree to which an item accurately differentiates test takers in terms of the measured behavior) and item difficulty (i.e., the percentage of examinees who pass an item) were examined. Items that did not meet acceptable levels of discrimination or difficulty were deleted. Finally, the test developers compared item functioning between three pairs of groups: male vs. female, African American vs. non-African American, and Hispanic American vs. non-Hispanic American. A logistical regression procedure was used to test all items contained in the TOLD-P:4. Results indicated that no more than three items were associated with significant effect sizes within each comparison (e.g., male vs. female), and the effect sizes were regarded as "negligible" (manual, p. 53). These results suggest that the test functions in a manner that is nonbiased with respect to gender, race, and Hispanic ethnicity. In addition, the means for the six composite scores for all racial and ethnic subgroups were in the normal range (defined by the test authors as 90 to 110).

Reports validity coefficients obtained when test scores are compared to other tests measuring the same construct

Describes acceptability of evidence presented by test publisher to support test interpretation and use

Describes evidence related to construct identification, such as raw scores increasing with age (for constructs expected to show developmental progression), moderate relationships among subtests, and confirmatory factor analyses

Criterion-prediction validity was investigated by correlating TOLD-P:4 scores with those of three oral language measures: the Pragmatic Language Observation Scale (PLOS; Newcomer & Hammill, 2009), the Test of Language Development—Intermediate: Fourth Edition (TOLD-I:4; Hammill & Newcomer, 2008), and the Verbal Comprehension composite of the Wechsler Intelligence Scale for Children—Fourth Edition (WISC-IV; Wechsler, 2003; T8:2903). The average correlation coefficients between the TOLD-P:4 and scores from the three tests ranged from .52 to .77 across all subtests and .61 to .81 for the composites. The test authors describe these coefficients as "large or very large" (manual, p. 55) in magnitude based on guidelines from Hopkins (2002). ...

Criterion-prediction validity was also examined through positive predictive outcome analyses. Three indices were calculated: A sensitivity index (correct identification of children with language disorder), a specificity index (correct identification of children with normal language), and positive predictive value (the proportion of children who truly have a language problem among all those whom the measure identifies as having a language problem). Based on several previous studies, the test authors consider index values of .70 or greater to be acceptable. Results indicated that the TOLD-P:4 provided acceptable levels of sensitivity, specificity, and positive predictive value (≥ .70).

Construct-identification validity "relates to the degree to which underlying traits of a test can be identified and the extent to which these traits reflect the theoretical model on which the test is based" (manual, p. 60). The test authors predicted that they would observe a positive correlation between test performance and age, given that the TOLD-P:4 was designed to measure language development. Results generally supported this hypothesis. The test authors also predicted that performance would differ among ability groups. Compared to the larger normative sample, mean scores were considerably higher for the gifted and talented group and considerably lower for the language impaired group. The means of all minority subgroups were within the normal range, although they tended to fall below the European American group (except for Asian/Pacific Islanders), particularly on the composite scores. Next, relationships among the TOLD-P:4 subtests were examined. The test authors predicted that correlations would be moderate, suggesting overlap but not redundancy. Correlation coefficients among the six core subtests were statistically significant, ranging in strength from .34 to .68, indicating that the TOLD-P:4 subtests measure related, but not redundant language skills. For the supplemental subtests, the median coefficients for Word Discrimination, Phonemic Analysis and Word Articulation were .41, .42, and .20, respectively. These results suggest that Word Articulation measures a unique construct, lending support for the test authors' decision not to combine the three supplemental subtests into a phonological composite. An additional factor analysis indicated that all six core subtests loaded onto one factor (Factor 1), while the supplemental subtests loaded onto a different factor (Factor 2). These results lend additional support for keeping the three phonology subtests separate instead of including them in the Spoken Language composite. ...

Resources that Support Test Evaluation and Selection

Information about Tests

ERIC Institute of Education Sciences	*http://eric.ed.gov/*
ETS Test Collection Database	*http://www.ets.org/test_link/about*
*Mental Measurements Yearbooks**	*http://buros.org/test-reviews-information*
*Pruebas Publicadas en Espanol**	*http://buros.org/test-reviews-information*
Test Reviews Online	*https://marketplace.unl.edu/buros/*
*Tests in Print**	*http://buros.org/test-reviews-information*

*Available through database subscription services offered by EBSCO and/or Ovid. Check with your reference librarian.

Evaluating Tests

- Questions to Ask When Evaluating Tests http://buros.org/questions-ask-when-evaluating-tests
- Using a Mental Measurements Yearbook Review to Evaluate a Test (includes list and definitions of key psychometric terms and concepts) http://buros.org/using-mental-measurements-yearbook-review-evaluate-test#
- Using a Mental Measurements Yearbook Review and Other Materials to Evaluate a Test (includes list and definitions of key psychometric terms and concepts) http://buros.org/using-mental-measurements-yearbook-review-and-other-materials-evaluate-test
- Glossaries available via the Buros Center's Assessment Literacy pages *http://buros.org/glossaries* includes Glossary of Important Assessment Measurement Terms (National Council on Measurement in Education); Glossary of Testing, Measurement, and Statistical Terms (Riverside Publishing); and Glossary of Standardized Testing Terms (Educational Testing Service)

Testing Standards, Codes, and Guidelines

The Buros Assessment Literacy page http://buros.org/standards-codes-guidelines lists 20 documents promulgated by various organizations concerned with testing. Access and/or links to ordering information is provided as well. Among the organizations listed:

American Association of School Administrators
American Counseling Association
American Educational Research Association
American Federation of Teachers
American Psychological Association*
Association for Assessment in Counseling*
International Test Commission
Joint Committee on Testing Practice
National Association of Elementary School Principals
National Association of Secondary School Principals
National Association of School Psychologists*
National Council on Measurement in Education
National Education Association
Society for Industrial and Organizational Psychology*

*Document available as PDF download.

References

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

Anastasi, A. (1989). What is test misuse? Perspectives of a measurement expert. In Educational Testing Service (Ed.), *The uses of standardized tests in American education* (pp. 15-25). Princeton, NJ: Educational Testing Service.

Betz, S. K., Eickhoff, J. R., & Sullivan, S. F. (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools, 44,* 133-146.

Carlson, J. F., & Geisinger, K. F. (2008). Psychological diagnostic testing: Addressing challenges in clinical applications of testing. In R. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 67-88). Washington, DC: American Psychological Association.

Carlson, J. F., & Geisinger, K. F. (2012). Test reviewing at the Buros Center for Testing. *International Journal of Testing, 12,* 122-135.

Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (2014). *The nineteenth mental measurements yearbook.* Lincoln, NE: Buros Center for Testing.

Camara, W. J. (1997). Use and consequences of assessments in the USA: Professional, ethical and legal issues. *European Journal of Psychological Assessment, 13,* 140-152.

Cizek, G. J., Koons, H. K., & Rosenberg, S. L. (2011). Finding validity evidence: An analysis using the *Mental Measurements Yearbook.* In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K-12 settings* (pp. 119-138). Washington, DC: American Psychological Association.

Cizek, G. J., Rosenberg, S. L., & Koons, H. K. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68,* 397-412.

Eyde, L. D., Robertson, G. J., & Krug, S. E. (2010). *Responsible test use: Case studies for assessing human behavior.* Washington, DC: American Psychological Association.

Geisinger, K. F., & Carlson, J. F. (2009). Standards and standardization. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 99-111). New York, NY: Oxford University Press.

Gerhardstein Nader, R. (2013). *Vocabulary Assessment Scales–Expressive/Vocabulary Assessment Scales–Receptive.* Lutz: FL: Psychological Assessment Resources.

Hutchins, T. L., & Cannizzaro, M. S. (in press). [Test review of Clinical Evaluation of Language Fundamentals–Fifth Edition]. In J. F. Carlson, K. F. Geisinger, & J. L. Jonson (Eds.), *The twentieth mental measurements yearbook.* Retrieved from http://marketplace.unl.edu/buros/

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64), Westport, CT: American Council on Education and Praeger.

Kirk, C., & Vigeland, L. (2014). A psychometric review of norm-referenced tests used to assess phonological error patterns. *Language, Speech, and Hearing Services in Schools, 45,* 365-377.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Moyle, M., & Long, S. (in press). [Test review of Test of Language Development–Primary: Fourth Edition]. In J. F. Carlson, K. F. Geisinger, & J. L. Jonson (Eds.), *The twentieth mental measurements yearbook.* Retrieved from http://marketplace.unl.edu/buros/

Newcomer, P. L., & Hammill, D. D. (2008). *Test of Language Development—Primary: Fourth Edition.* Austin, TX: PRO-ED.

Nitko, A. (n.d.). Using a Mental Measurements Yearbook Review and Other Materials to Evaluate a Test. Retrieved from http://buros.org/using-mental-measurements-yearbook-review-and-other-materials-evaluate-test

Pavello, S., & Ireland, M. (2014, November). Beyond reliability and validity: Evidence-based practices in test selection. Paper session presented at the 2014 ASHA Convention, Orlando, FL.

Rudner, L., & Impara, J. C. (n.d.). Questions to Ask When Evaluating Tests. Retrieved from http://buros.org/questions-ask-when-evaluating-tests

Spenciner, L. J. (in press). [Test review of Vocabulary Assessment Scales–Expressive/Vocabulary Assessment Scales–Receptive]. In J. F. Carlson, K. F. Geisinger, & J. L. Jonson (Eds.), *The twentieth mental measurements yearbook.* Retrieved from http://marketplace.unl.edu/buros/

Thorndike, R. M. (1999). Book review: Conoley, J. C., & Impara, J. C. (Eds.). (1995). *The twelfth mental measurements yearbook.* Lincoln, NE: The Buros Institute of Mental Measurements. *Journal of Psychoeducational Assessment, 17,* 50-55.

Wiig, E. H., Semel, E., & Secord, W. A. (2013). *Clinical Evaluation of Language Fundamentals—Fifth Edition.* San Antonio, TX: Pearson.