Agency for Healthcare Research and Quality • www.ahrq.gov

Evidence Report/Technology Assessment

Number 52

Criteria for Determining Disability in Speech-Language Disorders

Summary

Overview

Approximately 42 million people (1 in 6) in the United States have some type of communication disorder. Of these, 28 million have communication disorders associated with hearing loss, and 14 million have disorders of speech, voice, and/or language not associated with hearing loss. The personal and societal costs of these disorders are high. On a personal level, such disorders may affect nearly every aspect of daily life. Estimates of annual societal costs in the United States range from $30 billion to $154 billion in lost productivity, special education, and medical costs.

Over the last several decades, researchers and clinicians have developed a vast array of assessment instruments for speech, voice, and language; one source reviewing commercially available assessment instruments includes more than 140 tools in its most recent edition. Important clinical decisions follow from the assessment of a person with a communication disorder. These clinical decisions affect an individual's access to services and funding (e.g., eligibility for special education services, third-party payer coverage of treatment, and Social Security disability income).

Thus, the quality of the evaluation procedures on which such decisions are based is an important issue for individuals with a communication disorder, the clinicians involved in their evaluation and treatment, and the policymakers with fiscal responsibilities for services to individuals with these disorders. This evidence report, prepared by staff of the Research Triangle Institute-University of North Carolina at Chapel Hill Evidence-based Practice Center (RTI-UNC EPC) is directed to audiences who must grapple with this set of issues.

Reporting the Evidence

The clinical questions in this report were developed in conjunction with the Social Security Administration (SSA) to assist the agency in reviewing its criteria for determining disability in individuals with speech or language disorders, or both. Currently, disability determination depends on the functional limitations individuals experience, either with respect to employment in adults or with respect to the major life activities of children or adolescents (for example, school or play).

Therefore, in evaluations of individuals with speech and language disorders, the SSA is concerned with the concurrent relationship between the degree of impairment as measured by the assessment instrument and functional limitations associated with the speech or language impairment. Another commonality in the definitions of disability in children and adults is that the disability must be expected to last for at least 12 months or to result in death during that period. This criterion leads to a second important concern for the SSA, which is to know what evidence is available for various speech and language assessment instruments regarding their predictive power for future functioning of an individual. The SSA is interested in children and adults who (1) are English-speaking and have normal hearing, with or without normal cognition; (2) are non-English-speaking and have normal hearing, with or without normal cognition;

(3) are mentally retarded; (4) have learning disorders; and (5) are hard of hearing.

Based on concerns related to the criteria and process for determining disability in children and adults, the SSA outlined two key questions as the basis for this report. First, do the 18 reviewed instruments have demonstrated reliability, validity, and normative data? Second, do the instruments have demonstrated predictive validity for the individual's communicative impairment and performance?

Methodology

Search Process and Inclusion Criteria

The task of synthesizing the available evidence on all speech and language evaluation instruments was clearly too large an undertaking to complete within the scope of this project. Thus, EPC staff had to select and prioritize instruments in such a way as to address the critical informational needs of the SSA while also limiting the scope to fall within the contractual boundaries of the project. To do this, we assembled a panel of 10 national experts, our Technical Expert Advisory Group (TEAG). They, along with Agency for Healthcare Research and Quality (AHRQ) and SSA staff, identified 19 instruments for literature review and evidence analysis—three each for adult language, adult speech, child speech, and voice, and eight for child language disorders. One speech instrument can be used with both adults and children and thus was counted twice. We later excluded one instrument because it was not a single instrument but instead was an approach to conducting more comprehensive clinical analysis of phonological patterns for which standard "diagnostic test characteristics" would be hard to determine.

The RTI-UNC EPC review team conducted detailed searches of the relevant English-language literature from 1966 (or the initiation of the specific electronic database) to October 2000 using the MEDLINE®, CINAHL, PsycLIT®, ERIC, Health and Psychosocial Instruments (HAPI), and Cochrane Collaboration databases. We initially excluded all gray literature. After reviewing abstracts for eligibility, however, we recognized that, for many instruments, data on reliability and validity could be found only in the instrument manuals. Thus, we expanded efforts to include instrument manuals in the review. We also examined reference lists of all included articles and instrument manuals to identify additional studies.

The EPC team applied a series of inclusion and exclusion criteria to the literature searches. Essentially, we included all English-language research on the selected instruments in children and adults (ages 18 through 62) in which the study evaluated the instrument's reliability, validity, or ability to predict future communicative impairment and/or functioning (i.e., predictive validity). Articles reporting the efficacy or effectiveness of speech or language therapy that did not provide information relevant to the key questions were excluded. Because of the need to address issues facing the SSA in establishing disability criteria in the United States, we excluded articles providing normative data from populations other than the United States.

The EPC team selected studies for inclusion from among 1,238 citations using a process of duplicate but independent review of titles, abstracts, and, where necessary, full papers. Discussion leading to consensus was used to resolve disagreements. The number of citations reviewed ranged from three, for the Dysarthria Examination Battery (DEB) and Voice Handicap Index (VHI), to 256, for the Test of Language Development (TOLD).

The team abstracted data, using single abstraction with subsequent review by clinical and methodological experts, from 92 articles whose abstracts met inclusion criteria. Two reviewers with expertise in quantitative psychology and experience in the validation and standardization of educational tests abstracted the data. During the data abstraction phase, we eliminated 53 articles because they did not meet inclusion criteria or did not address the version of the instrument selected by TEAG members.

The EPC study director and clinical experts completed a quality rating for each article and manual. The quality rating scales evaluated research design and conduct, measurement of reliability and validity, development of instrument norms, justifications for conclusions, and external validity concerns. Six additional items evaluated aspects of instrument development or revision for the instrument manuals.

The team compiled the data into a series of five evidence tables for each instrument. The first of these tables provides information on the study design and conduct and the quality scores assigned by the methodologist and the expert clinicians. The subsequent four tables describe the reliability, validity, predictive validity for future communicative functioning, and available normative data found in the reviewed articles and manuals.

Subsequently, the team graded the evidence summarized in the tables, assessing whether the evidence met thresholds for acceptable reliability, validity, and availability of normative data. Where relevant, we used classic criteria for clinical decisionmaking about *individuals*, not groups of subjects. The criteria employed were:

- **Reliability**—the criterion for reliability is "strictly" met if the following three conditions are *all* met:

- Internal consistency reliability, measured using either Cronbach's coefficient alpha or Kuder-Richardson statistics (K-R 20), is greater than or equal to 0.90;
- Test-retest/intra-rater reliability is greater than or equal to 0.90 if measured using a correlation coefficient, or greater than or equal to 0.80 if measured using Cohen's Kappa; and
- Inter-rater reliability is greater than or equal to 0.90 if measured using a correlation coefficient, or greater than 0.80 if measured using Cohen's Kappa.

- **Validity**—the criterion for validity is met if the following conditions are *all* met:
  - Instrument developers examine relationships between subtests, composite scores, and total scores, establishing hypotheses a priori for these relationships and for patterns of scores for individuals belonging to various groups of import;
  - These relationships are all statistically significant at $p < 0.05$; and
  - In the case of correlation coefficients, the magnitude of the relationship is at least 0.30, thus providing evidence of a moderate correlation.
- **Normative Data**—the criterion for normative data is met if the following conditions are *all* met:
  - Data are available for the population targeted by the instrument;
  - An adequate sample size is used (i.e., at least 100 per group); and
  - Evidence is provided on how well the sample represents the population.

Some might reasonably argue that we set the criterion for internal consistency reliability too high given the complexity of speech and language functioning and disorders. Additionally the variability in daily performance that arises from these different speech and language disorders suggests that our criterion for test-retest reliability or intra-rater reliability was also set too high. Thus, we defined a "relaxed" criterion, which differs from the strict criterion in that internal consistency reliability may be as low as 0.80 and/or test-retest/intra-rater reliability may be as low as 0.80 (correlation) or 0.70 (Cohen's Kappa). The relaxed criterion is at a level suitable for having confidence in group, rather than individual comparisons.

After grading the psychometric properties of the individual instruments, we graded the strength of the overall body of evidence for groups of instruments identified by age group and disorder. We graded instrument manuals and peer-reviewed literature separately employing the following definitions for both.

- Acceptable: research or analyses were well conducted, had representative samples of reasonable size, and met our psychometric evaluation criteria discussed earlier.
- Unacceptable: studies were poorly conducted, used small or nonrepresentative samples, or had results that did not meet or only partially met the psychometric criteria.

Findings

Reliability, Validity, and Availability of Normative Data

The EPC team evaluated the strength of evidence describing the reliability, validity, and availability of normative data separately for instruments assessing adult language, child language, adult speech, child speech, and voice disorders.

- **Adult Language Instruments**—The Porch Index of Communicative Ability (PICA) met our relaxed standards of evidence for both reliability and validity, as did the original version of the Western Aphasia Battery (WAB); however, one small study suggested that the WAB might not consistently classify patients with aphasia. The Boston Diagnostic Aphasia Examination, 2nd Edition (BDAE-2) met neither the reliability nor validity criterion.

  Although normative data are available for two of the instruments, these data were derived from individuals treated at single institutions. Information was insufficient to assess whether they are representative of typical aphasics.

- **Child Language Instruments**—Three tests—the Clinical Evaluation of Language Fundamentals, 3rd Edition, Spanish Edition (CELF-3Sp), the Test of Language Development, Primary, 3rd Edition, (TOLD-P:3), and the Test of Language Development, Intermediate, 3rd Edition, (TOLD-I:3)—met the standards we established for reliability, validity, and the availability of representative normative data.

  The Preschool Language Scale, 3rd Edition (PLS-3) met the relaxed reliability criterion for all age groups except children between 0 and 8 months of age; the Clinical Evaluation of Language Fundamentals, 3rd Edition (CELF-3) met the relaxed criterion for total score but not for composite scores.

  With the exception of the Spanish version of the PLS-3, all instruments provided normative data derived from nationally representative populations. The CELF-3

(Spanish version) derived norms representative of the US Hispanic population.

Only the developers of the TOLD-P:3 and TOLD-I:3 provided evidence of the reliability and validity for use with four of the five populations specifically targeted by the SSA.

- **Adult Speech Instruments**—None of the adult speech disorder instruments met the standards of evidence we established for both reliability and validity. The Stuttering Severity Instrument for Children and Adults, 3rd Edition (SSI-3), however, met the validity criterion.

  No instrument met normative data standards. Although normative data were available for the SSI-3 and the Assessment of Intelligibility in Dysarthric Adults (AIDS), these data had been derived from individuals treated at single institutions. Instrument developers provided insufficient information to assess whether these patients were representative of adults with speech disorders.

- **Child Speech Instruments**—Neither the Goldman-Fristoe Test of Articulation, 2nd Edition (GFTA-2) nor the SSI-3 met our relaxed criteria for reliability and validity. The GFTA-2 met our relaxed criterion for internal consistency reliability. Developers of both instruments employed nonstandard statistical methods to test other forms of reliability.

  GFTA-2 provided normative data derived from nationally representative populations; the SSI-3 also provided normative data but gave no information on its representativeness.

- **Voice Instruments**—Both the Voice Handicap Instrument (VHI) and the Kay Elemetrics Multi-Dimensional Voice Program (MDVP) met our criteria for reliability, validity, and availability of normative data.

Prediction of Future Communicative Functioning

We found only four studies providing evidence about prediction of future functioning; thus, we consider the evidence incomplete on this point. Of the 18 instruments we reviewed, information on predictive validity was available for only four—one for adult language disorders, two for child language disorders (but not for versions *directly* reviewed in this report), and one for child speech disorders. None of the instruments we reviewed for either adult speech disorders or voice disorders had evidence of predictive validity.

Future Research

Further research is needed to evaluate and demonstrate the reliability, validity, and availability of normative data for instruments used to assess speech and language functioning and disorders. Instrument developers must be encouraged to document all types of instrument reliability (internal consistency, test-retest or intra-rater, and inter-rater reliability) and validity (content, construct, and concurrent validity) and to use currently accepted statistical procedures for psychometric analyses. Normative samples need to be representative of the population(s) of interest and of sufficient size that instruments can be shown to provide valid, interpretable results.

Funding agencies can facilitate this process by providing resources for the development and validation of new and existing instruments. Likewise, journal editors can help by encouraging the submission of reports on instrument reliability and validity, identifying peer reviewers who are qualified to evaluate the quality and rigor of these types of reports, and then publishing such data in their journals.

With the increasing cultural, linguistic, and racial diversity of the U.S. population, the applicability of assessment instruments to individuals who are members of different subpopulations is of crucial importance to clinical diagnosis and the process of disability determination. Despite the existence of a large number of speech and language assessment instruments, we still lack appropriate instruments for reliably and validly assessing speech and language in many subgroups defined in terms of language, dialect, or cultural differences. Thus, future research funding and priorities should be directed at addressing these serious deficiencies. Funding sources should encourage research teams that represent collaborations among professionals with expertise in speech and language disorders, cultural experts for the demographic subpopulations of interest, professionals with expertise in disorders that often co-occur with speech and language impairment, and psychometric experts.

In addition to demographic subpopulations, research is needed on the applicability of speech and language assessment instruments for assessment of individuals with different disorders, such as severe physical impairment, mental retardation, learning disorders, and hearing impairment. Including representative numbers of members of these subgroups in normative samples during instrument standardization is important, but improving the evidence base requires analyses examining reliability and validity of instruments for subpopulations, not just for the total normative sample. Researchers and instrument developers should be encouraged to fill this gap.

Further, large-scale research also is needed on the ability of speech and language assessment instruments to predict future performance. Such investigations should not be

limited to the predictive value of instruments in assessing specific intervention programs or in predicting future performance of a restricted subgroup. Rather, in terms of concern about disability, prediction of future test performance and future adaptive performance in everyday life is also critical. Such a "real world" research agenda would not only assist the SSA in decisions about disability but also contribute to the "ecological validity" of all speech and language assessments. We need both more instruments providing direct measurement of activity limitations and participation restrictions and more research demonstrating the relationship between speech and language impairment and activity limitations or participation restrictions.

Information on costs and burden to patients and to those in health care delivery settings should also be assembled, as it will likely be valuable in helping SSA or clinicians to select among otherwise seemingly similar instruments. A related area for future research is to compare the relative sensitivity and specificity of different approaches to disability determination for different types and degrees of speech and language impairment and to determine when the relative costs and benefits justify the addition of standardized instruments to the assessment process rather than relying solely on clinical judgments.

Important future research in this area includes investigation of the societal costs of speech and language disorders and the societal benefits of treating them. A good deal of work is needed simply on amassing data on costs of illness and costs of treatment. Combined with better information on efficacy and effectiveness of treatment, as called for above, such information would help researchers, clinicians, and policymakers better understand the cost-effectiveness of alternative therapeutic modalities.

Virtually no literature is available on the adverse effects or harms of diagnostic testing or disability evaluation. We urge that researchers take a broader perspective on the investigation of speech and language instruments, so as to shed some light on the likelihood that adults or children may be mislabeled (in both positive and negative ways) and on the consequences of such labeling.

Finally, we see a rich portfolio of research concerning appropriate ways to manage speech, language, or voice disorders in both adults and children. A necessary part of such investigations involves tracking patients' progress over time, and obviously the types of instruments reviewed here could play a part in such outcomes assessments. However, the deficiencies in many of these popular and well-known instruments need to be addressed before they can be used with confidence in treatment trials or studies. Apart from the basic measurement issues, methodological work is needed on the responsiveness of these instruments (that is, on their sensitivity to change and on the calculation of appropriate effect sizes that reflect change over time for individuals and groups). One strategy for those engaging in or supporting research on the management of patients with speech and language disorders is to build solid methodological research directly into treatment and rehabilitation studies, thereby strengthening both the given studies and the measurement field as a whole.

Availability of the Full Report

The full evidence report from which this summary was taken was prepared for AHRQ by the Research Triangle Institute—University of North Carolina at Chapel Hill Evidence-based Practice Center under contract No. 290-97-0011. It is expected to be available in the spring of 2002. At that time, printed copies may be obtained free of charge from the AHRQ Publications Clearinghouse by calling 800-358-9295. Request Evidence Report/Technology Assessment No. 52, *Criteria for Determining Disability in Speech-Language Disorders*. When available, Internet users will be able to access the report online through AHRQ's Web site at www.ahrq.gov.

www.ahrq.gov

AHRQ Pub. No. 02-E009
January 2002

ISSN 1530-440X