

## **UTAH LD TEST SELECTION COMMITTEE TEST RATING PROCEDURES**

### **Background**

Implementing regulations for The Education of All Handicapped Children Act of 1975, PL 94-142 (Education of All Handicapped Children Act of 1975, 1975) specified that, “(a) A team may determine that a child has a specific learning disability if, “(1) The child does not achieve commensurate with his or her age and ability levels.” Thus, one criterion for specific learning disability qualification was that the child exhibits a severe discrepancy between aptitude (intelligence) and achievement.

At the time PL 94-142 was being implemented, there was much discussion about how severe discrepancy should be measured but no general agreement. Several formulas were proposed as a national standard but none were agreed upon. In the end, each state was left to adopt its own standard.

As an aid to states the U. S. Department of Education formed the Special Education Programs Workgroup on Measurement Issues in the Assessment of Learning Disabilities. This group of measurement experts, headed by Cecil Reynolds, was charged with determining how best to measure severe discrepancy. After appropriate deliberations, it arrived at an agreed upon formula (Reynolds, Berk, Gutkin, Boodoo, Mann, Cox, Page, & Willson, 1983; Reynolds, 1984-85).

Utah responded to the severe discrepancy requirement in two ways. First, it adopted the Special Education Programs Workgroup on Measurement Issues in the Assessment of Learning Disabilities formula as the state standard and developed software, ESTIMATOR, to make the necessary calculations. Second, it established the LD Test Selection Committee. This group of special educators, psychologists, and speech pathologists was charged with reviewing aptitude and achievement tests to determine if they were technically sound and appropriate to be used in conjunction with the adopted formula as part of determining if a student has a specific learning disability. The committee recommends tests to the Utah State Office of Education – Special Education. Those approved are incorporated into the ESTIMATOR software program.

Over the years, Utah’s severe discrepancy formula has been modified to make technical improvements. Also, federal regulations and state rules pertaining to learning disabilities qualification have expanded. Nevertheless, the LD Test Selection Committee continues to do its work and the ESTIMATOR program is continually updated to provide districts wishing to employ a severe discrepancy qualification model with appropriate tools.

This document describes the standards by which the LD Test Selection Committee determines if tests are technically adequate for learning disabilities qualification in Utah.

### **Introduction**

The committee acts on recommendations to review the suitability of tests used as part of eligibility determinations for students to receive special education services under the category Specific Learning Disability. Recommendations to review tests may come from local education agencies (LEA), test publishers, and school psychologists, among others. Further, committee members often suggest consideration of tests that are newly published. The committee meets monthly to review tests. Recommendations concerning the use of tests are made to the Utah State Office of Education/Special Education Section which has final approval authority.

When a test is to be considered by the committee, a reviewer is assigned. The reviewer reads the test manual and assembles the documentation necessary to address the following: standardization, administration, reliability, validity and, for aptitude tests, measurement of general intelligence, “g”. The review is presented to the full committee at the monthly meeting. Following review and discussion, the committee votes to approve or not approve. The discussions and rationales for decisions are documented in the meeting minutes and posted on the Utah ESTIMATOR web page, <https://ESTIMATOR.srlonline.org>.

The following describes in detail the standards considered by the committee when reviewing tests. It is intended to serve as a guide to those presenting reviews of tests, an indication of what evidence is considered, and the standards of quality required for a favorable recommendation from the committee.

The basis for the test standards applied by the LD Test Selection Committee are those of Reynolds (1984-85) but the committee does not rigidly adhere to Reynolds’ recommendations. Deviations from Reynolds’ standards are noted in the sections below under Committee Practice.

### **A. Standardization**

*Reynolds’ Standard (2) Normative data should meet contemporary standards of practice and be provided for a sufficiently large, nationally stratified random sample of children [American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999].* Standardization samples are crucial to establishing levels of performance for comparison purposes in both normative and appositional interpretive models. Samples must be representative of a larger, well defined population, and large enough to provide a high

level of stability of measures of central tendency and dispersion. While for such categories of handicap as mental retardation and emotional disturbance, the preference for national over local norms is less convincing, in the case of LD, national norms are the preferred standard.

Reynolds' standard. (3) *Standardization samples for tests whose scores are being compared must be the same or highly comparable.* Under the best conditions, the aptitude, achievement, and other measures on which children are being compared to themselves or others should be co-normed (i.e., their standardization samples should consist of the *same* children, not just be sampled from the same population). When this is not possible, the norms for each test should be based on comparable samplings of the same population that all meet the requirements stated under (2) above. In this latter case, standardization of the scales should have been undertaken in the same general time period or else equating studies should be done. Scales normed on different samples and at different times are likely not to have the same mean and standard deviation across samples even though they may be scaled to a common metric *within* their respective samples. There is ample evidence to demonstrate that general levels of performance on aptitude and achievement measures vary in the population across time. As just one example, the population mean level of performance on the 1949 WISC is now very close to 110 and the 1974 revision (the WISC-R) now has a mean of nearly 103, though both are scaled, within their respective normative samples, to a mean of 100. While using highly similar samples tested at a similar time or with equating studies completed is acceptable in the face of the paucity of co-normed scales, co-norming will always be superior provided the sample meets the conditions of point (2).

Reynolds' standard. (6) *Age-based standard scores should be used for all measures and all should be scaled to a common metric.* The formulas for deriving severe discrepancies require the use of at least interval data. Score systems such as age or grade equivalents should be meticulously avoided whenever score comparisons are to be made. For purely descriptive purposes, such scores may be helpful but they are unacceptable for comparing scores of individuals or groups except under special circumstances that infrequently exist. Scores that are ratios of age and/or grade equivalents, such as an intelligence quotient derived from the traditional formula of  $(MA/CA) \times 100$ , are also inappropriate. Grade-base standard scores are inappropriate as well. The criteria for LD given in PL 94-142 specifically denote a discrepancy in achievement for age and ability. Age is properly considered in age-based standard scores. The scores should be age-corrected at appropriate intervals. Two to six months

are reasonable ranges of time in age groupings for the derivation of standard scores, but in no case should groups extend more than six months for children below age 6 years or more than 12 months for children above age 6 years.

### **Committee Practice**

**Step 1: Describe the standardization sample(s), including stratification. Describe the characteristics of the scaled scores from the test, i.e., indicate means, standard deviations, and what ages and age intervals are used. List the ages for which the test is normed, the lowest scaled score on the test for each age, and the standard error of measurement for each age.**

**Generally, sample sizes below 100 for each year of age are not considered adequate. In some cases, tests are approved for some ages and not for others because of sample size inadequacies.**

**The problem of comparability of norming samples between aptitude and achievement tests should be considered by test administrators but is not a responsibility of the Committee. It is important to use recently normed tests with large nationally stratified random samples of children.**

### **B. Administration**

Reynolds' standard. (1) *Tests should meet all requirements stated for assessment devices in the rules and regulations implementing PL 94-142.* This is not only a requirement of law, but is consistent with good professional practice. For example, administering a test in accordance with the instructions provided by the test maker is prerequisite to interpretation of test scores. If a standardized test is not given explicitly according to the instructions provided, inestimable amounts of error are introduced and norm-referenced scores are no longer interpretable. All personnel evaluating children with educational problems must be conversant with the requirements of PL 94-142 in this regard and adhere closely to these standards.

Reynolds' standard. (4) *For the purpose of arriving at a diagnosis, individually administered tests should be used.* For purely screening purposes, to determine the need for referral for comprehensive evaluation, group administered tests may be appropriate, though individual screening of young children is a better procedure (Reynolds & Clark, 1983). For all children, but especially handicapped children, there are simply too many uncontrolled and unnoticed factors that can affect performance in an adverse manner

(Chronback, 1970). These factors are more likely to be detected under the conditions of individual assessment, where close observation of the child is possible. Central to the proper assessment of learning problems for children at all ages is careful clinical observation of the child while performing a variety of academic and intellectual tasks (Kaufman, 1979; Lutey & Copeland, 1982; Reynolds & Clark, 1983). Special adaptations and testing procedures may be required as well that are best provided through individual assessment. Individual assessment, generally, affords better opportunity to maximize the child's performance.

### **Committee Practice**

**Step 2: The Committee considers only those tests which are designed for individual administration and which have norms based on data from individual administrations.**

**Conformance with requirements for proper selection, administration, and interpretation of tests is the responsibility of local education agencies and is not addressed by the committee.**

### **C. Measurement of "g"**

Reynolds' standard (5) *In the measurement of aptitude, an individually administered test of general intellectual ability should be used.* Such a test should sample a variety of intellectual skills but should nevertheless be recognized as a good measure of 'g,' the general intellectual ability that permeates performance on all cognitive tasks. If ability tests are too specific, a single strength or weakness in the child's ability spectrum may have an inordinate influence on the estimation of aptitude. It is also considered important to assess multiple abilities in deriving a remedial or instructional plan for a handicapped student and in preventing ethnic bias in the assessment (Reynolds, 1982). Highly specific ability measures (e.g., Bender-Gestalt, Columbia Mental Maturity Scale, Peabody Picture Vocabulary Test-Revised), while a necessary complement to a good assessment, are inadequate for estimating the general ability level of handicapped children.

### **Committee Practice**

**Step 3: Describe the scope of cognitive tasks measured by scales of aptitude.**

**In general, it is desirable to use a measure of aptitude that assesses performance on a variety of abilities. Determination of which test and score is the best measure of 'g' for a particular student is the**

responsibility of the examiner.

#### **D. Reliability**

Reynolds' standard (7) *The measures employed should demonstrate a high level of reliability and have appropriate studies for this determination in the technical manual accompanying the test.* The specific scores employed in the various discrepancy formulas should have associated internal consistency reliability estimates (where possible) of no less than .80 and preferably of .90 or higher. Coefficient *alpha* is the recommended procedure for estimating reliability, and should be routinely reported for each age level in the standardization sample of the test at not more than one year intervals. It is recognized that *alpha* will not be appropriate for all measures. Test authors and publishers should routinely use *alpha* where appropriate and provide other reliability estimates as may be appropriate to the nature of the test. Authors and publishers should be careful not to spuriously inflate reliability estimates through inappropriate sampling or other computational methods (Willson & Reynolds, in press, Chapter 3, and Stanley, 1971).

#### **Committee Practice**

**Step 4: Present data (usually tables) to indicate evidence of internal consistency reliability for each age level under consideration.**

**The Committee is reluctant to approve tests with alpha or split-half coefficients less than .90 (Anastasi & Urbina, 1997). Coefficient alpha is based on the inter-item correlations of all test items. When split-half reliabilities and coefficient alpha are both available, coefficient alpha is considered a better estimate of internal consistency. Reliability measures based on item response theory (IRT) are considered comparable to alpha. They are analogous to coefficient alpha in that they are based on all test items. They differ from alpha in that they are based on estimates of item error rather than inter-item correlations.**

**The Committee has no standard for test-retest reliabilities, but questions scales with  $r_{xx} < .80$ .**

#### **E. Validity**

Reynolds' standard (8) *The validity coefficient,  $r_{xy}$  representing the relationship between the measures of aptitude and achievement should be based on an appropriate sample.* This should consist of a large, stratified, random sample of normally functioning children. A large sample is necessary to reduce the sampling error in  $r_{xy}$  to an absolute minimum, since variations in  $r_{xy}$  will affect the calculation of a severe discrepancy, and affect the difference score distribution the most at the extremes of the distribution, the area of greatest

concern in this case. Normally functioning children are preferred since we are defining a severe discrepancy in part based on the frequency of occurrence of the discrepancy in the normal population. When co-norming of aptitude and achievement measures is conducted, this problem is simplified greatly since  $r_{xy}$  can be based on the standardization sample of the two measures (which should meet the standards in point 2 above) without any handicapped children included.

Reynolds' standard (9) *Validity of test score interpretations should be clearly established.* Though clearly stated in the rules and regulations for PL 94142, the Group felt that this requirement should receive special emphasis, particularly with regard to Cronbach's (1971) discussion of test validation. Validation with normal samples is insufficient for application to diagnosis of handicapping conditions and validity should be demonstrated for exceptional populations (though, for use of Formulas 20 and 21,  $r_{xy}$  should

again be based on a normal sample). This requirement is an urgent one, especially in certain areas of achievement where a paucity of adequate scales exists. To determine deviations from normalcy, validation with normal samples should typically be sufficient. This requirement also does not require separate normative data for each handicapping condition. The generalizability of norms and of validity data is in part a function of the question one seeks to answer with the test data and is ultimately an empirical question. Contemporary discussions of these latter problems may be found in Reynolds (in press) and Reynolds, Gutkin, Elliot, and Witt (1984).

Reynolds' standard. (10) *Special technical considerations should be addressed when using performance-based measures of achievement (e.g., writing skill).* For some measures, such as written expression, special problems of reliability and validity are present and must be specifically address. For example, interrater reliability of scoring on any measure calling for judgments by the examiner should be studied, reported, and be high (i.e., .85 to .90 or higher). This would hold for such tasks as the Wechsler vocabulary and comprehension measures as well, where examiners are frequently called upon to make fine judgments between the levels of quality of a response. Highly speeded and primarily memory-based tasks will also pose special technical problems that must be addressed.

Reynolds' standard. (11) *Bias studies on the instruments in use should have been conducted and reported.* Criterion-related validity should receive emphasis in this regard, but not to the exclusion of other studies of bias. Bias should be addressed with regard to

appropriate demographic variables that may moderate the test's validity, but at a minimum should include race, sex, and SES (though not necessarily simultaneously). In the assessment and diagnosis of LD in particular, sex bias needs to be investigated since boys outnumber girls in classes for the learning disabled by about 3.5 to 1. The procedure for evaluating bias in all aspects of a test are presented in a comprehensive form in Berk (1982) and in Jensen (1980). While measures that exhibit little or no statistical bias are the measures of choice, other measures can be used with the appropriate corrections.

### **Committee Practice**

**Step 5a: Present data (usually tables) to indicate evidence of correlation between aptitude and achievement tests. Include a description of the sample on which the correlations are based.**

**Step 5b: Present data (usually tables) to indicate evidence of correlation between the test and other tests of the same or similar construct.**

**Step 5c: Present evidence of content validity. Examples of test items should demonstrate that the test samples an appropriate domain. Since students taking a test, are ordinarily given different items depending on age and ability level, it is important to consider content validity (and test reliability) at different age levels.**

**Step 5d: If studies of test bias are available, summarize the findings.**

**The Committee is reluctant to approve tests with correlations between aptitude and achievement of less than .50. Although large national samples for such studies are desirable, often only small regional samples are available. If no empirical estimates of correlations can be found,  $r = .50$  is assigned as a default value.**

**Any available evidence of factorial, concurrent, and discriminate validity is considered. The Committee has no specific standards for these lines of evidence.**

**Evidence of content validity is examined at different ability levels. Tests are frequently approved for some age levels and not for others, based on the content validity and reliability. The Committee has no specific standards for these lines of evidence.**

**Evidence of test bias is examined. The Committee has no specific standards for evaluating test bias.**

Anastasi, A. & Urbina, S. (1997). *Psychological testing: Seventh edition*. Upper Saddle River, NJ: Prentice Hall, Inc.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, (1999). Standards for educational and psychological testing. Washington DC: American Educational Research Association

Education of All Handicapped Children Act of 1975, Pub. L. No. 94-142, Chapter 33, §1400-1485.

Reynolds, C. R., Berk, R. A., Gutkin, T. B., Boodoo, G. M., Mann, L., Cox, J., Page, E. B. & Willson, V. L. (1983). *Critical measurement issues in learning disabilities: Report of the United States Department of Education, Special Education Programs work group on measurement issues in the assessment of learning disabilities*. Washington, DC: United States Department of Education.

Reynolds, C. R. (1984-85). *Critical measurement issues in learning disabilities*. *The Journal of Special Education*, 18 (4), 451-476.