DOCUMENT RESUME

ED 218 348 TM 820 389

AUTHOR Davis, W. Alan; Shepard, Lorrie A.
TITLE The Use of Tests by LD Teachers, School Psychologists, and Speech/Language Specialists in the Diagnosis of Learning Disabilities.
PUB DATE Mar 82
NOTE 35p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Diagnostic Tests; Elementary Secondary Education; *Handicap Identification; *Learning Disabilities; Psychometrics; School Psychologists; Special Education Teachers; Surveys; *Testing Problems; *Test Interpretation; *Test Selection; Test Use
IDENTIFIERS Standards for Educational and Psychological Tests; *Test Appropriateness

ABSTRACT

The purposes of this study were to determine (1) which tests are most frequently used in the identification of learning disabilities, (2) how knowledgeable specialists are about the technical properties of the tests, and (3) what practices are used to safeguard valid diagnoses when psychometrically inadequate tests are used clinically. A two stage cluster sample of learning disabilities teachers (n=674), school psychologists (n=176), and speech/language teachers (n=240) was selected and surveyed by questionnaire. Although tests with high reliability and validity were generally preferred, poor tests were frequently used when superior substitutes were available. All classes of specialists tended to overrate the tests they used, and generally indicated a lack of familiarity with the psychometric properties of commonly used tests. Although a majority of specialists valued clinical judgment over test scores for diagnosis, substantial numbers appeared to lack knowledge of procedures to ensure the validity of such judgments. One fourth to one half of specialist groups did not interpret ability-achievement score discrepancies correctly. (Author)

Table 2 (continued)

Lists of Familiar Tests, Frequency of Use, Indices of Quality, and Correlations Between Test Use and Test Quality for Three Groups of Professionals

Category of Professional	Familiar Tests	Median Frequency of Use[a]	Shepard-Smith Index	Thurlow-Ysseldyke Index
LD Teachers	Peabody Individual Achievement Test	4.0	3	3
	WISC-R	3.9	4	3
	Beery VMI	3.8	2	0
	KeyMath Diagnostic Arithmetic Test	3.6	2	0
	Woodcock Reading Mastery	3.3	3.5	3
	Wide Range Achievement Test	3.2	2	1
	Detroit Tests of Learning Aptitude	2.9	0	0
	Wepman Auditory Discrimination	2.9	2	0
	Peabody Picture Vocabulary	2.7	0	2
	Illinois Test of Psycholinguistic Abilities	2.6	0	0
	Correlation between test use and quality		r = .77	r = .42
Speech/ Language Specialists	Peabody Picture Vocabulary	4.7	0	2
	Carrow Tests for Auditory Comprehension	3.8	2	0
	Detroit Tests of Learning Aptitude	3.5	0	0
	Wepman Auditory Discrimination	3.4	2	0
	Boehm Test of Basic Concepts	3.3	1	not rated
	Goldman-Fristoe Test of Articulation	3.1	3	2
	Illinois Test of Psycholinguistic Abilities	2.9	0	0
	Goldman-Fristoe-Woodcock Test of Auditory Discrimination	2.9	not rated	not rated
	Spencer Memory for Sentence	2.8	1	not rated
	Northwestern Syntax Screening	2.5	not rated	not rated
	Token Test	2.2	not rated	not rated
	Correlation between test use and quality		r = .32	r = .42

[a]5 = always used.

Algozzine, Regan, and Potter, 1979). Coles (1978) reviewed validation research on the ten tests of cognitive processing most frequently recommended for inclusion in the LD test battery. Each test failed to correlate with a diagnosis of learning disabilities (that is, the tests could not discriminate LD from normal learners). Arter and Jenkins (1979) reviewed validation research on 12 extensively used tests of perception and psycholinguistic abilities and found them too low in reliability to use for individual decisions. Most of the tests lacked sufficient diagnostic or predictive validity. Larsen, Rogers, and Sowell (1976) found that only one of four of the most widely used tests of cognitive processing was able to differentiate between two groups of 30 normal and 59 LD students. In general, LD children received higher, not lower scores on the four tests, although the LD students had lower mean IQ. Thurlow and Ysseldyke (1979) found that of 30 instruments used by three or more Child Service Demonstration Centers, only 8 had technically adequate norms, 10 had adequate reliability, and 9 had adequate validity. Only 7 met all three standards.

One of the questions to be addressed in this study is whether professionals involved in the identification of LD children in the schools are aware of the psychometric limitations of the tests they use. Historically, measurement specialists have known that regular classroom teachers have very poor knowledge about standardized tests, such as requirements for validity and the meaning of IQ scores (Goslin, 1967; Hastings, Runkel, and Damrin, 1961). One presumes, however, that specialists such as school psychologists, speech/language specialists, and learning disabilities teachers would have had considerably more technical training than regular classroom teachers. The research about what specialists actually know about tests is sparse. Ysseldyke, Algozzine, Regan, and Potter (1979) found that school personnel involved in placement and

identification decisions for LD students appear not to differentiate between technically adequate and technically inadequate tests. In a simulated decision-making exercise involving regular education teachers, special education teachers, administrators, school psychologists, and support personnel who had participated in at least two placement team meetings (N = 159), devices with technically adequate reliability and validity were selected as often as devices with inadequate reliability and validity. However, less than half of their sample was made up of specialists who routinely administer and interpret standardized tests.

Purpose

The purposes of this study were to determine: (1) how widespread is the use of psychometrically inadequate tests for the identification of LD, (2) how knowledgeable are learning disabilities teachers, school psychologists, and speech/language specialists about the technical properties of tests used most frequently, (3) how knowledgeable are professionals about interpreting pertinent test statistics such as ability-achievement discrepancy scores and years below grade level, and (4) what practices are used to safeguard valid diagnoses when inadequate tests are used clinically.

Methods

The data analyzed for this study were gathered in the course of a larger study evaluating the identification of children with learning disabilities in Colorado (Shepard, Smith, Davis, Glass, Riley, and Vojir, 1981). The study had several parts including two large-scale data collection components. Of primary interest for the present study was a survey of learning disabilities

teachers, school psychologists, speech/language specialists, and school social workers. The first three groups are particularly influential in the assessment and diagnosis of learning disabilities, and make the greatest use of tests for this purpose. School social workers, who make little or no use of tests, provide a comparison group, particularly to examine the use of clinical judgment. Supportive data for the present study is from the second component of the larger study, an analysis of the case files of individual pupils currently classified as LD.

Sampling

For both the survey of professionals and the study of case files, a stratified two-stage cluster sampling design was used. In the first stage, the 48 special education units in Colorado were categorized by type (districts or Boards of Cooperative Services) and were divided into three size categories. Twenty-two units were sampled at random from within the size and type categories. All of the sampled units agreed to participate. The sampling frame for the second stage consisted of lists of personnel in the sampled units supplied by the special education directors. All members of the school psychologist and speech/language specialist groups were selected because relatively small numbers (176 and 240, respectively) were involved. An 80% sample of LD teachers was drawn at random resulting in 674 selected cases. The return rates of LD teachers, school psychologists and speech/language specialists were 80%, 74%, and 75%, respectively. In addition to follow-up mailings, non-returns in an *a priori* 20% random subsample (called the core sample) were contacted by phone to obtain answers to key questions and to learn the reason for non-response. Statistical comparisons between core and non-core sample responses, and correlational analyses for the entire sample between day of return and opinions expressed were done to investigate non-response bias.

A similar sampling procedure was used in the study of student files. After 22 administrative units were selected in the first stage, 790 individual files were randomly sampled in proportion to stratum populations. Exact proportionality in deriving population estimates was achieved by post hoc weighting.

Instruments

Three questionnaires were developed, specifically tailored for the three categories of professionals surveyed. Issues regarding identification of LD were derived from the literature and from interviews with special education directors and formulated into questionnaire items. Several items were taken with permission directly from "A Survey of Attitudes Concerning Learning Disabilities," by S. A. Kirk, P. B. Berry, and G. M. Senf (1979). Additional items were modified versions of questions extracted with permission from the study by Applied Management Sciences, "Study to Evaluate Procedures Undertaken to Prevent Erroneous Classification of Handicapped Children."

All three questionnaires included a list (derived from pupil files, a survey of the literature, and interviews with directors of special education) of 52 tests and instruments used in the identification of LD. Respondents were asked to indicate how frequently they used each test and to rate the reliability and validity of each. Other questions investigated professionals' knowledge about the interpretation of test scores and test statistics of import in the diagnosis of learning disabilities. Such questions dealt with identifying a discrepancy between ability and achievement, allowing for increased variation in achievement at higher grade levels, and interpreting subscale scatter. Other questions dealt with the relative roles of tests and clinical judgment in diagnosing learning disabilities. Additional questions

dealt with topics outside the scope of the present study, including definition of LD and instructional interventions provided to LD pupils.

The questionnaires were pilot tested with small samples of specialists in administration units which were not part of the randomly selected sample. A few ambiguous or otherwise objectionable items were deleted. The questionnaires were also reviewed by two state advisory committees.

Analysis and Results

Test Use

Frequently used tests were identified through the survey of professionals and the study of student files. The instructions to professionals on each questionnaire were as follows: "Think of all of the staffings and assessments you have participated in in the last two years which led to LD placement (as well as those who were potentially LD but were staffed and not placed). Indicate below in approximately what % of the cases you used the following tests." Respondents chose among five categories of response: Never (0%); Rarely (1-15%); Sometimes (16-50%); Often (51-85%); Nearly Always (86-100%). Fifty-two tests were listed. From the survey responses a list of 18 tests was compiled which at least 40% of a professional group reported using often or nearly always. A second list was drawn up of tests reported used in the placement staffing of at least 10% of pupil files. Fifteen tests met this criterion; all but one, the Slosson, were in the professionals' list. The combined list of 19 tests is presented in Table 1.

The single best review of the technical quality of tests used in the diagnosis of learning disabilities is found in a monograph by Thurlow and Ysseldyke (1979). They evaluated 30 tests used nationally in Child Service

Table 1

Typical Tests Administered to LD Pupils in Colorado as Part of Their Initial Assessment and Staffing

Tests	Professional Using Test[a]	Percentage of LD Pupils Administered Test
Intelligence Tests		
Detroit Tests of Learning Aptitude	LD; Sp/Lang	38.1%
Peabody Picture Vocabulary Test	LD; Sp/Lang	46.6%
Slosson Intelligence Test	LD; Psych.	10.7%
WISC-R	LD; Psych.	58.6%
Achievement Tests		
KeyMath Diagnostic Arithmetic Test	LD	15.3%
Peabody Individual Achievement Tests	LD	38.8%
Wide Range Achievement Test	LD; Psych.	36.7%
Woodcock Reading Mastery Tests	LD	16.0%
Personality Tests		
Draw-A-Person	Psych.	26.0%
Kinetic Family Drawing	Psych.	13.8%
Sentence Completion	Psych.	13.9%
Perceptual and Processing Tests		
Beery Developmental Test of Visual-Motor Integration	LD; Psych.	45.9%
Bender (Visual-Motor) Gestalt Test	Psych.	46.3%
Spencer Memory for Sentences Test	Sp/Lang	8.1%
Wepman Auditory Discrimination Test	Sp/Lang	29.0%
Speech and Language Tests		
Boehm Test of Basic Concepts	Sp/Lang	4.7%
Carrow Tests for Auditory Comprehension of Language	Sp/Lang	5.1%
Goldman-Fristoe Test of Articulation	Sp/Lang	3.0%
Illinois Test of Psycholinguistic Abilities	Sp/Lang	32.5%

[a]Used by more than 40% of these professionals.

Demonstration Centers and rated the adequacy of their norms, reliability, and validity. In the larger evaluation study from which this study derives, Shepard and Smith et al. (1981) elaborated on the Thurlow and Ysseldyke ratings, drawing upon additional technical reviews regarding the psychometric adequacy of each test for the purpose of LD identification. They arrived at an overall rating for each of 19 tests on a 5-point scale. Only tests rated "3" or above were judged adequate for making placement decisions about individuals. To achieve this rating, tests were required to have test-retest or parallel form reliability of at least .85, some evidence of diagnostic validity, and at least some normative data. The KeyMath test, for example, was judged to have quite high content validity but does not even have standard deviations for the purpose of normative comparisons. Therefore, it received the highest rating for instructional planning purposes but was rated as inadequate for the diagnosis of LD.

Of the 19 tests identified as most frequently used in Colorado, only 4 were given acceptable ratings. None of these were tests of perception or cognitive processing. This finding supports what is generally known about the technical adequacy of the typical LD battery of tests used nationally.

Professionals' Knowledge of Test Adequacy

If professionals are aware of the psychometric strengths and weaknesses of the tests they use, we should find relatively greater use of better tests. To examine the relationship between use and technical adequacy, we expanded the set of tests to include all tests with which a group of professionals expressed familiarity. (Respondents were asked to rate each test's reliability and validity or indicate "Don't Know." Any test which was actually rated by a majority of respondents was judged "familiar.") A test was eliminated from a

list if it was considered outside of that group's professional domain (e.g., the WISC-R was not included in the list of tests for speech/language specialists because they rarely make use of an IQ score even though they rate the test highly). Correlations were obtained for median use with technical adequacy using the 5-point Shepard-Smith index and a 4-point composite index from the Thurlow-Ysseldyke ratings. They are summarized in Table 2.

The correlations reported in Table 2 are not offered as definitive measures of association since they are obtained from relatively crude measures. They are convincing in their consistency, however, since they refer to three relatively distinct sets of tests. The positive relationship between technical adequacy and use suggests that school psychologists, LD teachers, and speech/language specialists are not insensitive to technical quality in their selection of tests.

The relatively low correlations between test use and test quality could be explained in part if professionals make frequent use of tests, which they know to be low in technical quality, to aid in hypothesis formation when high quality alternatives do not exist. We examined this hypothesis in each of its parts. To do so, we asked professionals themselves to rate the reliability and validity of the tests they use.

All three categories of respondent were given the following instructions:

> Please indicate below which tests have adequate reliability evidence and are valid <u>for the purpose of identifying learning disabilities</u> [underliring in original]. Of course, any test could be invalid if used inappropriately. But which tests have research evidence of their validity when used appropriately? Check the columns that apply: 1=Adequate 2=Inadequate 3=Don't Know.

Reliability and validity were rated separately.

Only two instances were found of tests which were extensively used even though they received more ratings of "inadequate" than "adequate" from a

Table 2

Lists of Familiar Tests, Frequency of Use, Indices of Quality, and Correlations Between Test Use and Test Quality for Three Groups of Professionals

Category of Professional	Familiar Tests	Median Frequency of Use[a]	Shepard-Smith Index	Thurlow-Ysseldyke Index
School Psychologists	WISC-R	4.6	4	3
	Bender Visual-Motor Gestalt	4.3	1	0
	Draw-A-Person	3.9	1	not rated
	Kinetic Family Drawing	3.3	1	not rated
	Sentence Completion	3.2	1	not rated
	Wide Range Achievement Test	2.8	2	1
	Beery VMI	2.6	2	0
	Stanford-Binet	2.2	3	1
	Peabody Individual Achievement Test	2.2	3	3
	KeyMath Diagnostic Arithmetic Test	2.1	2	0
	Peabody Picture Vocabulary	2.0	0	2
	Slosson Intelligence	1.7	2	0
	Illinois Test of Psycholinguistic Abilities	1.4	0	0
	Detroit Tests of Learning Aptitude	1.4	0	0
	Correlation between test use and quality		r = .33	r = .35

[a]5 = always used.

Table 2 (continued)

Lists of Familiar Tests, Frequency of Use, Indices of Quality, and Correlations Between Test Use and Test Quality for Three Groups of Professionals

Category of Professional	Familiar Tests	Median Frequency of Use[a]	Shepard-Smith Index	Thurlow-Ysseldyke Index
LD Teachers	Peabody Individual Achievement Test	4.0	3	3
	WISC-R	3.9	4	3
	Beery VMI	3.8	2	0
	KeyMath Diagnostic Arithmetic Test	3.6	2	0
	Woodcock Reading Mastery	3.3	3.5	3
	Wide Range Achievement Test	3.2	2	1
	Detroit Tests of Learning Aptitude	2.9	0	0
	Wepman Auditory Discrimination	2.9	2	0
	Peabody Picture Vocabulary	2.7	0	2
	Illinois Test of Psycholinguistic Abilities	2.6	0	0
	Correlation between test use and quality		r = .77	r = .42
Speech/ Language Specialists	Peabody Picture Vocabulary	4.7	0	2
	Carrow Tests for Auditory Comprehension	3.8	2	0
	Detroit Tests of Learning Aptitude	3.5	0	0
	Wepman Auditory Discrimination	3.4	2	0
	Boehm Test of Basic Concepts	3.3	1	not rated
	Goldman-Fristoe Test of Articulation	3.1	3	2
	Illinois Test of Psycholinguistic Abilities	2.9	0	0
	Goldman-Fristoe-Woodcock Test of Auditory Discrimination	2.9	not rated	not rated
	Spencer Memory for Sentence	2.8	1	not rated
	Northwestern Syntax Screening	2.5	not rated	not rated
	Token Test	2.2	not rated	not rated
	Correlation between test use and quality		r = .32	r = .42

[a] 5 = always used.

particular group: school psychologists made extensive use of the Kinetic Family Drawing and the Sentence Completion tests but rated them relatively low in reliability and validity. In two other instances, tests were rated "adequate" by small pluralities. Thirty-one percent of the psychologists expressed reservations about the Goodenough-Harris Drawing Test. The same percentage of speech/language specialists rated the Wepman Auditory Discrimination Test inadequate, yet 81% said they used it at least some of the time. When these four tests are omitted, correlation between test use and test quality (Shepard-Smith index) increases slightly for the two groups: from .33 to .49 for psychologists; from .32 to .34 for speech/language specialists. Correlations between use and each group's <u>own</u> rating of its tests were, however, much higher: .79 for school psychologists; .70 for LD teachers; .48 for speech/language specialists (suggesting that they do not have the same understanding of the tests' limitations as the measurement specialists).

We now turn to an examination of the hypothesis that some low quality tests are used because no high quality alternatives exist. To do so, we consider four types of tests separately: IQ tests, achievement tests, perceptual and cognitive processing tests, and speech and language tests.

Among intelligence tests for children, the WISC-R is superior. Split-half reliabilities are quite high: .96, .93, and .97 for verbal, performance, and full-scale IQ, respectively. More than a thousand research studies, taken together, provide compelling evidence of construct validity. In a review of LD tests, Coles (1978) concluded that the WISC-R and Stanford-Binet are the preferred individual IQ measures and are the only tests in the typical LD battery which have strong enough validity to warrant consistent use. The WISC-R was the only intelligence test for children rated adequate in norms, reliability, and validity by Thurlow and Ysseldyke (1979).

Among school psychologists, the WISC-R was the overwhelming choice (83% used it in half or more of their assessments, compared to 17% for the Peabody Picture Vocabulary Test). On the other hand, LD teachers reported greater use of the Peabody Picture Vocabulary Test than the WISC-R (51% and 47%, respectively), which is reflected in the finding from the study of pupil cases that 23% of students placed in LD were given only low quality intelligence tests rather than the WISC-R or Stanford-Binet. Speech/language specialists made far greater use of the PPVT and the Detroit Tests of Learning Aptitude than the WISC-R, which is explained in part by the fact that this group is interested in indications of language processes rather than IQ. Harder to justify is the finding that speech/language specialists as a group rated both of these tests higher in reliability and validity than the WISC-R, when the Detroit tests in particular are lacking evidence of subtest reliability or discriminant validity to support their use as measures of language processing (Silverstein, 1978).

Unlike intelligence tests, where individually administered tests are preferred, group achievement tests have greater reliability and validity for making decisions about individuals than most of their individually administered counterparts. In particular, the Peabody Individual Achievement Test (PIAT) and the Wide Range Achievement Test (WRAT), the most frequently administered tests in Colorado for LD placement decisions, rate lower in reliability and validity than many of their group competitors such as the California Test of Basic Skills (Lyman, 1971). By trying to cover subject matter over many years of curriculum, only a few test items actually measure at each child's level of skill. The effect is the same as trying to make accurate assessments with tests that are only four or six questions long. Subtest reliability is inadequate for both tests (Salvia and Ysseldyke, 1978; Thurlow and Ysseldyke, (1979). However, distinctions can be made among individual achievement tests. The

PIAT has better content validity and normative data than the WRAT. The Woodcock Reading Mastery Test has substantial evidence of content validity and adequate reliability.

In fact, group achievement tests were hardly ever used for LD placement decisions. LD teachers showed a strong preference for the PIAT; school psychologists preferred the WRAT.

The category of perceptual and processing tests provides fewer clear-cut choices. While some have adequate reliability, none has convincing evidence of empirical validity (Coles, 1978; Ysseldyke and Salvia, 1974; Arter and Jenkins, 1979). Use of such tests to generate hypotheses with a proper degree of caution might be defensible. One is justifiably concerned, however, if professionals place unwarranted confidence in the validity of processing test scores.

Extensive use is made of perceptual and processing tests in the diagnosis of LD. In Colorado, we found that 79% of students placed in LD had been given at least one. School psychologists made extensive use of the Bender Visual-Motor Gestalt Test and the Beery Developmental Test of Visual-Motor Integration. Nearly half of speech/language specialists regularly used the Wepman Auditory Discrimination Test.

As has been stated previously, the use per se of inadequate instruments does not undermine the validity of diagnoses so long as specialists are fully cognizant of the shortcomings of the measures. Unfortunately, the validity of the Bender, the Beery, and the Wepman tests was rated highly by the professional groups who use them often. The adequacy of the Bender and Beery was endorsed, respectively, by 57.4% and 48.1% of school psychologists. The Beery was rated adequate by 46.2% of LD teachers; the Wepman by 31% of speech/language teachers. Such misplaced confidence in these tests can certainly contribute to misidentification of LD.

Among speech and language tests, it is important to draw attention to the Illinois Test of Psycholinguistic Abilities (ITPA) as being especially bad. The test has been severely criticized as lacking in concurrent validity (Newcomer and Hammill, 1976), discriminant validity (Waugh, 1975), and subtest reliability (Lumsden, 1978). It was used in the placement of an estimated 33% of the LD students in Colorado, however, and was rated adequate in validity by 33.6% of LD teachers, 39.2% of speech/language specialists, and 23.8% of school psychologists.

In general, we found that professionals are not widely familiar with the technical properties of the tests they use. LD teachers select inferior intelligence tests over superior ones. School psychologists and LD teachers make extensive use of inadequate achievement tests, when adequate ones are available. Substantial numbers of all three professional groups (from one-third to one-half) have misplaced confidence in the validity of tests that they use to measure language, perceptual, and cognitive processing.

Discrepancy Score Interpretation

A significant discrepancy between ability and achievement is the primary identifier of specific learning disabilities in the federal definition (U.S.O.E., 1977, p. 65083) and is central to the Colorado definition of LD. It is operationalized by administering an IQ test and an achievement test and determining whether a child's level of achievement is "significantly" below what one would expect based on his ability. "Significant" discrepancy can be interpreted to mean a difference which is reliably non-zero, evaluated by comparing the obtained difference with the standard error of measurement of the difference (see Salvia and Ysseldyke, 1978); or a difference which is large compared to others with the same obtained IQ, evaluated by comparison with the

standard error of estimate. Either approach requires the achievement score to fall well below the IQ score when each is converted to a common z-score or percentile metric.

The data in Table 3 are evidence that clinicians' instincts may not always be accurate in discerning a true or reliable discrepancy. Professionals were asked how low an achievement score would have to be to be significantly discrepant from an IQ score of 90. Since a 90 IQ is at the 25th percentile, achievement must be well below the 25th percentile to be significantly discrepant. An achievement score at the 21st percentile (Option C) is neither reliably different (achievement would have to be below the 9th percentile to be reliably different at $\alpha = .10$) nor unusual (44% of students with IQ 90 have achievement scores as low or lower than this). Only Option D, achievement at the 12th percentile or lower, could be correct. (In fact, even this seemingly extreme score results in the identification of 27% of students with IQ of 90 as having a significant IQ-achievement discrepancy.) The correct answer to the question should have been obvious without the need for computation or normative tables so long as clinicians realized that an IQ of 90 is roughly at the 25th percentile. Large percentages of each professional group over-estimated the significance of the less discrepant scores. Thirty-eight percent of the LD teachers, 33% of the school psychologists, and 45% of the speech/language specialists selected answers that were incorrect. Overestimation of the significance of IQ-achievement discrepancy could easily result in the over-identification of learning disabilities.

The tendency for many special education professionals to consider any score below grade level as a significant discrepancy for children with IQ's near 90 is consistent with practices observed in the study of LD pupil files. The habit of many professionals is to consider any IQ score in the 90 to 109

Table 3

Percentage of Professionals Selecting Various Cut-Offs on a Specific Question about a Significant Discrepancy

Question 28. If a third grade child had a WISC-R IQ score of 90, in your opinion, how low should his or her reading grade equivalent score be (in October) to be a significant discrepancy?

A. 2.7 (35th percentile) or lower
B. 2.5 (28th percentile) or lower
C. 2.2 (21st percentile) or lower
D. 2.0 (12th percentile) or lower

Professionals	Option Selected				
	A	B	C	D[a]	Blank
PCD Teachers	3.8%	9.0%	25.5%	51.1%	10.6%
School Psychologists	0.5%	8.5%	23.6%	54.4%	13.0%
Speech/Language Specialists	4.9%	9.7%	30.6%	35.1%	19.7%

[a]Correct answer.

range to be "average" and therefore to expect achievement to be at grade level, that is, at the 50th percentile (for that grade and month of school). This subgroup of professionals (which according to Table 3 is one-third of the psychologists and one-half of the speech/language specialists) is unaware that an IQ of 90 is fully two-thirds of a standard deviation below the mean and therefore consistent with achievement at roughly this same level. Further, they are unaware that once pupils are beyond the earliest grades, the normal variability in achievement can be great enough to place a score of -.67σ (minus .67 of a standard deviation) below the actual grade placement. These observations of practices, suggesting that professionals may not realize how much variability there is in normal children, are consistent with Kaufmann's (1976a, 1976b) findings that clinicians interpret WISC-R profiles as deviant that are quite frequent in the normal population.

Clinical Judgment

Tests are not the only means of assessment. Professionals frequently draw on their intuitions and experience to determine whether a given child has a learning disability. Known as the professional judgment, clinical judgment or medical model of assessment, this is a process wherein a clinician observes a pattern of symptoms or behaviors of a child and matches that pattern with mental conceptions and ideas of an underlying trait or disease. The clinician hypothesizes that the child has that particular disease, then goes on to look for other confirming or disconfirming symptomatic evidence. By this rationale, many signs or test scores, that would be unreliable and insufficient in themselves to produce valid diagnoses, may be combined to produce valid diagnoses.

Many Colorado specialists believe in and use clinical judgment in the identification of LD. The data in Table 4 are the opinions of specialists

Table 4

Professionals' Opinions about the Use of Clinical Judgment in the Identification of LD

Question 35.[a] It is possible to make valid diagnoses of LD from invalid tests if they are only used as stimuli to test clinical hypotheses.

Professionals	Strongly Agree	1	2	3	4	5	Strongly Disagree
LD Teachers		5%	25%	33%	21%	13%	
Social Workers		1%	8%	21%	45%	22%	
School Psychologists		7%	37%	24%	15%	11%	
Speech/Language Specialists		5%	36%	30%	17%	11%	

Question 36.[a] Test results should be clearly secondary to clinical judgments in arriving at an LD diagnosis.

Professionals	Strongly Agree	1	2	3	4	5	Strongly Disagree
LD Teachers		11%	30%	28%	25%	4%	
Social Workers		5%	23%	26%	33%	10%	
School Psychologists		16%	25%	23%	25%	7%	
Speech/Language Specialists		8%	36%	29%	22%	4%	

Question 37.[a] If you agree or strongly agree, describe what steps should be taken by professionals to ensure the validity of clinical judgments. [Written responses were read twice, first to develop categories and then to check consistency of classification. Categories and sample answers are reported here for LD teachers and school psychologists.]

Professionals	% of Those Responding	Categories of Response
LD Teachers	2%	Clarify the definition --"How can you diagnose what you can't define?"

[a]Question numbers correspond to LD teachers' questionnaire.

Table 4 (continued)

Professionals' Opinions about the Use of Clinical Judgment in the Identification of LD

Professionals	% of Those Responding	Categories of Response
LD Teachers (cont.)	2%	Use valid tests --"need more valid perceptual tests"
	50%	Gather informal data that can't be gotten from tests --"check functional level in classroom" --"analysis of errors, observe patterns in errors, diagnosis of learning style" --"writing samples, work behaviors, teacher anecdotes, parental anecdotes" --"spend time observing and getting to know child"
	8%	Observe on more than one occasion
	3%	Better documentation of observations --"observational reports which empirically list measurable characteristics"
	4%	Tests should support clinical judgments --"validation of test scores with class performance"
	12%	Confirmation of hypotheses, convergence --"that they be consistent" --"confirmation by other judges" --"other than test results to corroborate judgments"
	2%	Follow-up --"trial placement" --"diagnostic teaching"
	5%	Team decisions --"multi-disciplinary assessment" --"input from all members of staffing team"
	5%	Better training of professionals, experience --"train new teachers to make valid judgments" --"keep up with research as to reliability and validity"

Table 4 (continued)

Professionals' Opinions about the Use of Clinical Judgment in the Identification of LD

Professionals	% of Those Responding	Categories of Response
LD Teachers (cont.)	5%	Trust professionals --"Since so many of the tests are considered invalid, why should clinical judgments be considered less valid?"
	2%	Not Classified
	100%	
School Psychologists	7%	Clarify the definition --"written guidelines" --"concept is too fuzzy to yield highly valid judgments"
	9%	Use valid tests --"standard series of recommended tests"
	4%	Gather informal data --"classroom observation"
	9%	Better documentation of observations --"objective data" --"quantifying" --"thorough written descriptions of the observed behaviors"
	20%	Confirmation of hypotheses --"test results should be consistent with observations" --"data from various professionals should converge on the problem"
	9%	Follow-up --"trial placement and diagnostic teaching process"
	15%	Team decisions --"have a variety of professionals participate with a lot of sharing"

Table 4 (continued)

Professionals' Opinions about the Use of Clinical Judgment in the Identification of LD

Professionals	% of Those Responding	Categories of Response
School Psychologists (cont.)	21%	Better training of professionals --"inservice" --"certification of professionals" --"hire competent professionals" --"know research"
	4%	Can't be done--opposed to validity --"validity of clinical judgments would mechanize and sterilize a human process"
	2%	Not Classified
	100%	

abou the use of clinical judgment. To item 36, "Test results should be clearly secondary to clinical judgments in arriving at an LD diagnosis," between 28% and 44% of the specialist groups agreed. It is reasonable to assume that a larger percentage would support a statement which made test results and clinical judgments equal in importance for making diagnoses. Social workers were asked these questions pertaining to clinical judgment although they had not been asked questions about specific tests. Their responses are anomalous. Social workers give formal tests very infrequently and therefore rely on clinical judgments in their own diagnostic work. But they, more than any other group, rejected the idea that test results should be secondary to clinical judgment. They also expect their test-giving colleagues to administer only valid tests, i.e., they soundly reject the position (item 35) that invalid tests could be used to generate clinical hypotheses.

If specialists agreed that tests should be secondary to clinical judgments in arriving at a diagnosis of LD, they were asked to write in what steps should be taken to ensure the validity of those judgments. The answers from each group were read twice, first to develop categories and then to check the consistency of answer classifications. The categories of answers with illustrative quotations are presented in Table 4 for LD teachers and school psychologists. These two groups best typify the range of responses (reflecting the greatest differences). Speech/language specialists answered more like the LD teachers, and social workers more like the psychologists.

When used correctly, as a means for generating hypotheses that may then be confirmed or disconfirmed, clinical judgment may be an appropriate method of LD diagnosis. However, it has many critics. Historically it has been shown to be much less reliable than statistical methods for the diagnosis of psychological disorders (Meehl, 1954). Clinical judgments are apt to be overly

influenced by first impressions (Poulton, 1968) and by information which is readily available but not necessarily reliable (Tversky and Kahneman, 1974; Kahneman and Tversky, 1973). Information that conflicts with a previously held hypothesis tends to be ignored (Wason, 1968). Clinicians overlook the normal variability of traits and behaviors, fail to understand randomness, over-interpret small correlations, and mistake correlations for causes (Smedslund, 1963; Tversky and Kahneman, 1974; Peterson and Beach, 1967).

The questionnaire items had not been designed to test the knowledge of specialists about weaknesses in the professional judgment paradigm. Some post hoc analyses were prompted, however, by patterns in the data, especially the pronounced differences between LD teachers and psychologists. Questions about whether clinicians understood the hypothesis testing model were also raised by a separate study of LD pupil files.

As a group, school psychologists were much more aware that clinical judgment should require reconciling information from various sources, interpreting only the confirmed signs and seeking explanations for divergent data. Twenty percent of the psychologists answered directly with "confirmation of hypotheses" answers as the means to ensure validity; an additional 9% proposed a trial placement diagnostic-teaching model (follow-up) which also implies confirmation. The majority of LD teachers (50%) gave answers that equated clinical judgment with informal data collection rather than a method for synthesizing formal and informal evidence. Only 4% of psychologists gave answers of this type. The categories of LD teachers' response that reflected a need for confirmation were: "tests should support clinical judgments" (4%), "confirmation of hypotheses" (12%), "follow-up" (2%)--accounting for a total of 18% of those responding. For both groups the category "team decisions" did not include answers reflecting the need for convergence of signs. Rather,

answers were classified in this category if they mentioned bringing together many different specialists but did not say anything about agreement or consistency among them. Some answers, in fact, conveyed the opposing view that seeking consistency would hinder the divergent insights of various specialists. As a group, LD teachers seemed more sanguine about the validity of clinical judgments. Only 5% said that better training of professionals was needed (compared to 21% for the psychologists). In addition, 5% of the LD teachers said that professionals should just be trusted (0% of the psychologists said this).

The most pessimistic reading of the free-response question leads to the conclusion that less than one-fifth of the LD teachers and only one-third of the school psychologists understand the model of hypothesis testing and verification. To be sure, more specialists would have demonstrated some knowledge if the question had been posed more directly, "Can you identify the steps in hypothesis testing?" Nevertheless, they were asked how to ensure the validity of clinical judgments. The answers from 50% of the LD teachers who viewed clinical judgment as merely an informal data collection method did not reflect any sort of model for consistency or verification (i.e., the meaning of these classroom observations was expected to be self-evident).

The pessimistic conclusion that specialists do not have adequate knowledge of a clinical judgment model was further supported by data from a different source. As part of the larger study of the identification of learning disabilities in Colorado (Shepard and Smith, 1981), a representative sample of 790 LD pupil files were read and coded by trained coders. One set of coded variables dealt with the consistency of clinicians' diagnoses of processing deficits. Areas of deficit were defined broadly so that if problems were cited in the same general area, e.g., auditory problems, the

case was coded as having evidence of agreement. On the basis of individual specialist's reports (for each case) the following degrees of consistency were observed: 26% of the LD cases had at least some agreement between professionals, i.e., at least two professionals agreed on at least one problem area, although other unconfirmed problems might also have been cited; 11% had different processing disorders cited by different professionals (e.g., visual problems by the psychologist and memory deficits by the LD teacher), but no confirmations; 5% had contradictory evidence, i.e., what one clinician cited as an area of strength was cited as a deficit by another; 39% had a processing problem cited by only one clinician which was therefore not confirmed; 19% of the LD cases did not have a processing deficit cited by any clinician.

The more important results, germane to the question of whether clinicians understand the need for confirmatory evidence, was the rating of congruence between individual reports and the final diagnosis stated in the staffing minutes as the basis for placement in LD. The rating scale reflected the degree to which professionals in the staffing team sought confirmation of the inferred disability and attempted to reconcile and integrate the observations and conclusions of various team members. The three highest ratings describe only 7% of the LD population; in these cases the basis of handicap cited in the placement decision reflected a coherent picture of the child's intellectual functioning put together from the several separate clinicians' reports. The more prevalent practice, however, was for the staffing minutes to include all possible deficits observed by any clinician, with no attempt to reconcile inconsistent conclusions or seek confirmation; this occurred in 23% of the cases. Sixteen percent of the LD cases were said to have a processing disorder or perceptual problem in the staffing minutes when none had been cited by individual clinicians, perhaps because this is part of the legal definition of

LD, in Colorado. (Forty-eight percent of the cases were not rated either because a processing disorder was not stated in the determination of handicap or because there were no staffing minutes.)

Many clinicians prefer to use clinical judgment in lieu of inadequate tests. For example, when asked, "What percentage of LD children have a processing deficit diagnosed by clinical observation rather than test scores?" the average percentage reported by psychologists was 23% (σ = 20%), 25% by LD teachers (σ = 26%) and 30% by speech/language specialists (σ = 45%). The evidence from both the survey of professionals and the study of LD pupil files suggests, however, that many professionals do not know the steps essential to ensure the reliability and validity of clinical judgments. Therefore, although clinical judgment may be the best alternative when tests are inadequate, we can place very little confidence in professional judgment as currently practiced because only a minority of professionals follow a model of confirmation and verification.

Conclusions

Low but consistently positive correlations between indices of technical adequacy and frequency of use indicate that school psychologists, LD teachers, and speech/language specialists are somewhat sensitive to the psychometric properties of the instruments they use to diagnose learning disabilities. To a large extent, however, better tests are not preferred. Although psychologists consistently select superior tests of intelligence, LD teachers and speech/language specialists do not. Achievement tests with insufficient reliability for making placement decisions for individuals are widely preferred to more reliable alternatives. Among perceptual and cognitive processing tests, and

tests of language processes, there is a shortage of measures with consistent evidence of adequate diagnostic validity. All three groups of professionals express misplaced confidence in the adequacy of these measures, giving high ratings to those they use most often.

When asked technical questions about interpretation of test scores, from one-third to one-half of the professional groups could not correctly identify a significant discrepancy between IQ and achievement test scores. This criterion is essential to the definition of LD. Specialists apparently made the error of expecting grade level performance (50th percentile) from children with IQs of 90, not realizing that this score is at the 25th percentile. The effect of this misconception would be to misidentify normal but below-average pupils as learning disabled.

Large numbers of professionals, ranging from 28% of social workers to 44% of LD teachers, agreed that tests should be secondary to clinical judgments in arriving at an LD diagnosis. However, school psychologists were the only group for which a sizeable proportion gave evidence of understanding that the process of clinical judgment involves confirmation of findings among independent sources to ensure validity. To large numbers of LD teachers, clinical judgment meant informal assessment and the inclusion of different professional perspectives, or faith in the correctness of professional diagnoses but without a requirement for convergence among those observations.

Taken together, the findings of this study suggest that the validity of the identification process for learning disabled students is reduced by a lack of technical knowledge on the part of the professionals involved. They point up the need for more effective dissemination of psychometric knowledge, particularly in programs of professional preparation. Because professionals rely on clinical judgment in the absence of reliable and valid tests,

professional preparation must also emphasize the role of hypothesis testing through reconciling findings from diverse sources as central to a model of clinical judgment.

References

American Psychological Association. Standards for educational and psychological tests. Washington, D.C.: APA, 1974.

Arter, J. A., and Jenkins, J. R. Differential diagnosis--prescriptive teaching: A critical appraisal. Review of Educational Research, 1979, 49, 517-555.

Coles, G. S. The learning disabilities test battery: Empirical and social issues. Harvard Educational Review, 1978, 48, 313-340.

Goslin, D. Teachers and Testing. New York: Russell Sage Foundation, 1967.

Hastings, J. T.; Runkel, P.; and Damrin, D. Effects on use of tests by teachers trained in a summer institute (Cooperative research project No. 702). Urbana: University of Illinois, Bureau of Educational Research, 1961.

Kahneman, D., and Tversky, A. On the psychology of prediction. Psychological Review, 1973, 80, 237-251.

Kaufman, A. S. A new approach to the interpretation of test scatter on the WISC-R. Journal of Learning Disabilities, 1976, 9, 160-168. (a)

Kaufman, A. S. Verbal-performance IQ discrepancies on the WISC-R. Journal of Consulting and Clinical Psychology, 1976, 44, 739-744. (b)

Kirk, S. A.; Berry, P. B.; and Senf, G. M. A survey of attitudes concerning learning disabilities. Journal of Learning Disabilities, 1979, 12, 239-245.

Larsen, S. C.; Rogers, D.; and Sowell, V. The use of selected perceptual tests in differentiating between normal and learning disabled children. Journal of Learning Disabilities, 1976, 9, 32-37.

Lumsden, J. Review of Illinois Test of Psycholinguistic Abilities, Revised Edition. In O. K. Buros (Ed.), The eighth mental measurements yearbook (Vol. I). Highland Park, N.J.: The Gryphon Press, 1978.

Lyman, H. B. Review of Peabody Individual Achievement Test. Journal of Educational Measurement, 1971, 8, 137-138.

Meehl, P. E. Clinical versus Statistical Prediction. Minneapolis: University of Minnesota Press, 1954.

Newcomer, P. L., and Hammill, D. D. Psycholinguistics in the Schools. Columbus, Ohio: Charles E. Merrill Publishing, 1976.

Peterson, C. R., and Beach, L. R. Man as an intuitive statistician. Psychological Bulletin, 1967, 68, 29-46.

Poland, S.; Ysseldyke, J.; Thurlow, M.; and Mirkin, P. Current assessment and decision-making practices in school settings as reported by directors of special education (Research Report No. 14). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1979.

Poulton, E. C. The new psychophysics: Six models for magnitude estimation. Psychological Bulletin, 1968, 69, 1-19.

Salvia, J., and Ysseldyke, J. E. Assessment in Special and Remedial Education. Boston: Houghton-Mifflin, 1978.

Shepard, L. An evaluation of the regression discrepancy method for identifying children with learning disabilities. Journal of Special Education, 1980, 14, 79-91.

Shepard, L.; Smith, M. L.; Davis, W. A.; Glass, G. V.; Riley, A; and Vojir, C. Evaluation of the Identification of Perceptual-Communicative Disorders in Colorado. Laboratory of Educational Research, University of Colorado, Boulder, Colorado 80309, February 1981.

Silverstein, A. B. Review of Detroit Tests of Learning Aptitude. In O. K. Buros (Ed.), The Eighth Mental Measurements Yearbook (Vol. I). Highland Park, N.J.: The Gryphon Press, 1978.

Smedslund, J. The concept of correlation in adults. Scandinavian Journal of Psychology, 1963, 4, 165-173

Smith, M. L. How Educators Decide Who Is Learning Disabled. Springfield, Ill.: Charles C. Thomas, 1982.

Thurlow, Martha T. Views of placement team members subsequent to their participation in meetings. In J. Ysseldyke, B. Algozzine, and M. Thurlow (Eds.), A Naturalistic Investigation of Special Education Team Meetings (Research Report No. 40). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1980.

Thurlow, M. L., and Ysseldyke, J. E. Current Assessment and Decision-Making Practices in Model Programs for the Learning Disabled (Research Report No. 11). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1979.

Tversky, A., and Kahneman, D. Judgment under uncertainty: Heuristics and biases. Science, 1974, 185, 1124-1131.

U.S.O.E. Assistance to states for education of handicapped children: Procedures for evaluating specific learning disabilities. Federal Register, 1977, 42, 65082-65085.

Wason, P. C. On the failure to eliminate hypotheses: A second look. In P. C. Wason and P. N. Johnson-Laird (Eds.), Thinking and Reasoning Baltimore: Penguin Books, 1968.

Waugh, R. P. The I.T.P.A.: Ballast or bonanza for the school psychologist? Journal of School Psychology, 1975, 13, 201-208.

Weatherly, R. A. Reforming Special Education. Cambridge, Mass.: MIT Press, 1979.

Ysseldyke, J.; Algozzine, B.; Regan, R.; and Potter, M. Technical Adequacy of Tests Used by Professionals in Simulated Decision Making (Research Report No. 9). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1979.

Ysseldyke, J., and Salvia, J. A. Diagnostic-prescriptive teaching: Two models. Exceptional Children, 1974, 41, 181-186.