Davis, W.A. & Shepard, L.A. (1983). "Specialists' Use of Tests and Clinical Judgement in the Diagnosis of Learning Disabilities." Learning Disability, 6, 128.138.

SPECIALISTS' USE OF TESTS AND CLINICAL JUDGMENT IN THE DIAGNOSIS OF LEARNING DISABILITIES

W. Alan Davis and Lorrie A. Shepard

Abstract. **The purposes of this study were to determine (a) which tests are most frequently used in the identification of learning disabilities, (b) how knowledgeable specialists are about the technical properties of the tests, (c) how knowledgeable professionals are about interpreting discrepancy scores, and (d) what practices are used to safeguard valid diagnoses when psychometrically inadequate tests are used clinically. A representative sample of learning disabilities teachers (n = 542), school psychologists (n = 130), and speech/language teachers (n = 179) in Colorado was selected and surveyed by questionnaire. Although subjects generally preferred tests with higher reliability and validity, poor tests were still used frequently even when superior substitutes were available. All groups 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-third to one-half of each specialist group could not correctly interpret ability-achievement score discrepancies.**

Tests play a major part in the identification and placement of students with learning disabilities (LD). For example, in their national survey of federally funded Child Service Demonstration Centers, Thurlow and Ysseldyke (1979) found that norm-referenced tests were used more often than any other source of information in making screening, classification, and placement decisions.

Many of the tests commonly included in the learning disabilities test battery are technically inadequate. Ysseldyke and Salvia's (1974) evaluation of standardized tests frequently used in LD identification, according to the criteria established jointly by the APA, AERA and NCME (1974), revealed that the majority did not meet minimum standards. Similarly, lists of tests with inadequate norms or inadequate reliability or validity evidence have been published (Ysseldyke, Algozzine, Regan, & Potter, 1979). Coles (1978) reviewed validation research on the 10 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). The validation evidence is particularly

W. ALAN DAVIS, M.A., is a doctoral student, Research and Evaluation Methodology, University of Colorado, and Resource Specialist in Program Evaluation, Jefferson County Public Schools.

LORRIE A. SHEPARD, Ph.D., is Associate Professor and Chairman, Research and Evaluation Methodology, School of Education, University of Colorado.

weak for tests of perception, cognitive processing, and psycholinguistic abilities (Arter & Jenkins, 1979; Larsen, Rogers, & Sowell, 1976).

In view of the above findings, one of the questions to be addressed in this study is whether professionals involved in the identification of LD students are aware of the psychometric limitations of the tests they use. Historically, measurement specialists have recognized that regular classroom teachers demonstrate very poor knowledge about standardized tests, such as requirements for validity and the meaning of IQ scores (Goslin, 1967; Hastings, Runkel, & Damrin, 1961). One presumes, however, that specialists such as school psychologists, speech/language specialists, and learning disabilities teachers have had considerably more technical training than regular classroom teachers. The research regarding specialists' knowledge about tests is sparse. Ysseldyke et al. (1979) found that school personnel involved in placement and identification decisions about LD students appear not to differentiate between technically adequate and technically inadequate tests. Likewise, Bennett (1981) concluded that "serious problems exist with regard to the competence of professionals involved in the educational and psychological assessment of exceptional children" (p. 444), but noted that the evidence supporting this conclusion is partly opinion. More recently, Bennett and Shepherd (1982) gave a test of basic measurement concepts to learning disabilities specialists. Results showed that these specialists could answer only half of the questions, thereby scoring significantly poorer than a group of college students enrolled in an introductory measurement course.

Purpose of the Study

The purposes of the present study were to answer the following questions: (a) How widespread is the use of psychometrically inadequate tests for the identification of LD? (b) How knowledgeable are learning disabilities teachers, school psychologists, and speech/language specialists about the technical properties of tests used most frequently? (c) How knowledgeable are professionals about interpreting pertinent test statistics such as ability-achievement discrepancy scores? and (d) 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 students with learning disabilities[1] in Colorado (Shepard & Smith, with Davis, Glass, Riley, & Vojir, 1981). The study included a large-scale survey of professionals who participate in the identification of LD, as well as an analysis of representative pupil cases. In addition to the full technical report, the research procedures are also summarized in Shepard and Smith (1983).

Subjects

The subjects were 542 LD teachers, 130 school psychologists, and 179 speech/language specialists, selected using a stratified two-stage cluster sampling design to be representative of the Colorado populations. The response rate for each group was 74% or greater. Special followup procedures including telephone surveys were undertaken to assess non-response bias.

Where pertinent, data will also be cited from an analysis of a representative sample of 1,000 LD pupil files.

Instruments

Three questionnaires were administered (varying in length from 10 to 11 pages). The instruments were specifically tailored for the three categories of professionals surveyed. All three questionnaires included a list 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 items investigated professionals' knowledge about the interpretation of test scores and test statistics important in the diagnosis of learning disabilities. Such questions dealt with identification of 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 the diagnosis of learning disabilities.

ANALYSIS AND RESULTS

Test Use

Frequently used tests were identified through the survey of professionals and the study of student files. From the survey responses, a list of 18 tests was compiled which at least 40% of a pro-

TABLE 1

Tests Typically 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 Test	LD	38.8%
Wide Range Achievement Test	LD; Psych.	36.7%
Woodcock Reading Mastery Test	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 often by more than 40% of these professionals.

fessional group reported using *often* (in 51-85% of their assessments) or *nearly always* (in 86-100% of their assessments). A second list was drawn up of tests reportedly used in the placement staffing of at least 10% of pupil files. Fifteen tests met this criterion; all but one, the *Slosson*, were included in the professionals' list. The combined list of 19 tests is presented in Table 1.

Of the 19 tests used most frequently in Colorado, only 4[2] were judged acceptable by the APA test standards (American Psychological Association, 1974). The negative ratings of tests in the Shepard and Smith (1981) study were consistent with numerous other reviews (Arter & Jenkins, 1979; Coles, 1978; Lumsden, 1978; Lyman, 1971; Silverstein, 1978; Thurlow & Ysseldyke, 1979). Furthermore, the high-frequency use of technically inadequate tests was also consistent with studies conducted nationally (Thouvenelle, Rader, & Hanley, 1982; Thurlow & Ysseldyke, 1979; Ysseldyke et al., 1979).[3]

Professionals' Knowledge of Test Adequacy

In addition to determining which tests are used, a major purpose of the present research was to assess how knowledgeable professionals are about the technical adequacy of frequently used measures. The simplest way to analyze this question was to examine the relationship between use and technical 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. Simple correlations were obtained between frequency of use and both the Shepard and Smith (1981) and the Thurlow and Ysseldyke (1979) composite ratings of technical quality. The correlations for the three groups of professionals involved in assessment were on the order of .3 to .4 indicating a positive tendency to use better tests more often.[4]

The positive correlations between test use and test quality are nonetheless relatively low. This finding may 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. To examine this hypothesis, we asked professionals to rate the reliability and validity of the tests they use, specifically for the purpose of identifying learning disabilities (underlining in original instructions).

In two instances were tests used extensively but rated "inadequate" by a particular group. Thus, school psychologists made extensive use of the *Kinetic Family Drawing* and the *Sentence Completion* tests but rated them relatively low in reliability and validity, indicating that they recognized the limitations of the measures. Usually, however, specialists tended to give high ratings to the tests they used often whether or not the tests show evidence of reliability and validity. Correlations between use and each group's *own* rating of its tests were generally much higher than the correlations between use and the external ratings of quality reported earlier. This finding suggests that professionals involved in LD assessment do not have the same understanding of the tests' limitations as measurement specialists.

Another explanation for the use of some low-quality tests is the lack of high-quality alternatives. To examine this hypothesis, we considered 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 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. For 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*). However, LD teachers reported greater use of the *Peabody Picture Vocabulary Test* (PPVT) than the WISC-R (51% and 47%, respectively). This practice 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 either 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 this group's focus on 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, despite the *Detroit* tests' lack

of subtest reliability or discriminant validity evidence to support their use as measures of language processing (Silverstein, 1978).

The four achievement measures used most often in Colorado (see Table 1) are also the most popular nationwide (Thouvenelle et al., 1982). The *Peabody Individual Achievement Test* (PIAT) and the *Wide Range Achievement Test* (WRAT) were considered to have adequate reliability for individual placement decisions by 40% of the LD teachers and 49% of the school psychologists, respectively. Yet, subtest reliability is inadequate for both tests (Salvia & Ysseldyke, 1978; Thurlow & Ysseldyke, 1979). Furthermore, although the PIAT has better content validity and normative data than the WRAT, the WRAT is the test traditionally administered by school psychologists and continues to be rated higher by them. The poor reliability and validity of the PIAT and WRAT is to be expected since, by their very nature, they span such a "wide range" of achievement levels that only a few items specifically measure each level. The effect is the same as trying to make accurate assessments with tests consisting of only four or five questions. Ironically, professionals do not consider giving more reliable and valid standardized achievement tests such as the *Comprehensive Test of Basic Skills* (Lyman, 1971) even when the standardized administration would be appropriate for the nature of the student's handicap. The *Woodcock Reading Mastery* test is one of the technically best instruments in the LD battery, yet it was used in only 16% of the pupil cases and was rated as technically adequate by only 30% of the LD teachers and even fewer psychologists. Perhaps many professionals are not familiar with this relatively new instrument. Conversely, 60% of LD teachers said the *KeyMath* test has adequate validity evidence *for the identification of LD*, despite the lack of any normative data.

The category of perceptual and processing tests provides fewer clearcut choices. While some demonstrate adequate reliability, none has convincing evidence of empirical validity (Arter & Jenkins, 1979; Coles, 1978; Ysseldyke & Salvia, 1974). Use of such tests to generate hypotheses with a proper degree of caution might be defensible. One is justifiably concerned, however, if professionals use processing test scores without recognizing the lack of validity.

Although a "mental process" or "perceptual-processing" model is no longer well regarded among LD experts (Arter & Jenkins, 1979; Hammill, Leigh, McNutt, & Larsen, 1981), extensive use is still made of perceptual and processing tests in the diagnosis of LD in school settings. In Colorado, we found that 79% of students placed in LD had been given at least one (see footnote 3). School psychologists made extensive use of the *Bender Visual-Motor Gestalt Test* and the *Beery Developmental Test of Visual-Motor Integration*. Nearly half of the 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 used. Unfortunately, the validity of the *Bender*, the *Beery*, and the *Wepman* tests was rated high by the professional groups who used them often. Thus, the adequacy of the *Bender* and the *Beery* was endorsed, respectively, by 57.4% and 48.1% of school psychologists. Similarly, the *Beery* was rated adequate by 46.2% of LD teachers; the *Wepman* by 31% of speech/language teachers. Such misplaced confidence can contribute to misidentification of LD.

Among speech and language tests, the *Illinois Test of Psycholinguistic Abilities* (ITPA) has been severely criticized for lacking concurrent validity (Newcomer & Hammill, 1976), discriminant validity (Waugh, 1975), and subtest reliability (Lumsden, 1978). This instrument was used in the placement of an estimated 33% of the LD students in Colorado, however, and was rated as demonstrating adequate 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) misplace confidence in the validity of tests used to measure language, perceptual, and cognitive processing.

Discrepancy Score Interpretation

A significant discrepancy between ability and

achievement — the primary identifier of specific learning disabilities in the federal definition (U.S.O.E., 1977, p. 65083) — is also central to the Colorado definition of LD. It is operationalized by administering an IQ test and an achievement test and determining whether a student's level of achievement is "significantly" below what one would expect based on his/her ability. "Significant" discrepancy can be interpreted to mean (a) a difference which is reliably non-zero, based on a comparison between the obtained difference and the standard error of measurement of the difference (see Salvia & Ysseldyke, 1978); or (b) 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 2 are evidence that clinicians' instincts may not always accurately discern a true or reliable discrepancy. Professionals were asked how low an achievement score would have to be for them to consider it 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). 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 an IQ of 90 as having a significant IQ-achievement discrepancy.) The correct answer to the question should have been

TABLE 2

Percentage of Professionals Selecting Various Cutoffs 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
LD 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 anwser.

obvious without 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 overestimated 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 (plus omission rates of 11%, 13%, and 20%, respectively). Overestimation of the significance of IQ-achievement discrepancy could easily result in an overidentification of learning disabilities.

The tendency for many special education professionals to consider any score below grade level as a significant discrepancy for students with IQs near 90 is consistent with practices observed in the study of LD pupil files. Thus, many professionals consider any IQ score in the 90 to 109 range to be "average" and, therefore, 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 may be one-third of the psychologists and one-half of the speech/language specialists) is unaware that an IQ of 90 is

TABLE 3

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%	

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

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 $-.67\sigma$ (minus .67 of a standard deviation) in the year preceding the actual grade placement. These observations of score interpretations, suggesting that professionals may not realize how much variability is found in normal children, are consistent with Kaufman's (1976a, 1976b) findings that clinicians interpret as deviant WISC-R profiles that are frequent in the normal population.

Clinical Judgment

Tests do not constitute the only means of assessment. Professionals frequently draw on their intuitions and experience to determine whether a given child demonstrates a learning disability. Known as the *professional judgment* or *clinical judgment model of assessment*, this is a process wherein a clinician observes a student's pattern of symptoms or behaviors and matches that pattern with mental conceptions and ideas of an underlying trait or disorder. After hypothesizing that the child displays a particular disorder, the clinician goes on to look for other confirming or disconfirming symptomatic evidence. By this rationale, many signs or test scores, that, in themselves, would be unreliable and insufficient to produce 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 3 illustrate specialists' opinions about the use of clinical judgment. Between 28% and 44% of the specialist groups agreed to item 36, "Test results should be clearly secondary to clinical judgments in arriving at an LD diagnosis." 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 the 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 more than any other group, they rejected the idea that test results should be secondary to clinical judgment. They also expected their test-giving colleagues to administer only valid tests, i.e., they soundly rejected 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 such judgments. The answers from each group were read to develop response categories and then reread to check the consistency of answer classifications.[5] 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 (followup) which also implies confirmation. Fifty percent of LD teachers 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%), "followup" (2%) — accounting for a total of 18% of those responding. For both groups, a separate category of response, "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. They did not mention agreement or consistency among participants. 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 responded in this manner).

A somewhat 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 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 possess adequate knowledge of a clinical judgment model was further supported by data from the study of representative pupil cases (Shepard & Smith, 1981). Although at least one clinician cited a processing deficit in 81% of LD cases, only 26% of the entire sample had even one problem area confirmed by another clinician.[6] More significant to the claim that clinicians do not generally test for consistency in their conclusions were the data coded from staffing minutes. Only 7% of the LD pupil files contained summary statements that reflected an effort by the staffing team to reconcile and integrate the observations and conclusions of various team members. The more prevalent practice (23% of all cases) was for the staffing minutes to include a listing of all possible processing deficits observed by separate professionals (sometimes with each in the handwriting of the different diagnosticians). According to the staffing minutes, 16 percent of the LD cases were said to demonstrate a processing disorder or perceptual problem when none had been cited by individual clinicians. This may be because it is part of the legal definition of LD in Colorado.

Many clinicians reported that they prefer to use clinical judgment in lieu of inadequate tests. 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 tend to use technically better tests more often than they use technically inadequate ones. To a large extent, however, these professionals do not prefer better tests. 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 are widely preferred in favor of more reliable alternatives. Perceptual and cognitive processing tests and tests of language processes suffer from a lack of evidence of adequate diagnostic validity. However, significant numbers from all three groups of professionals apparently still believe these measures are reliable and valid, and are not aware of the theoretical and conceptual problems inherent in their use.

When asked technical questions about test score interpretation, 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. Apparently, specialists erroneously expected grade-level performance (50th percentile) from students 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 of which a sizable proportion evidenced an 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 without a requirement for convergence among these 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. These results emphasize the need for more effective dissemination of psychometric knowledge, particularly in professional preparation programs. Because professionals rely on clinical judgment in the absence of reliable and valid tests, their professional preparation must emphasize the role of hypothesis testing (through reconciliation of findings from diverse sources) as central to a model of clinical judgment.

REFERENCES

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

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

Bennett, R.E. Professional competence and the assessment of exceptional children. *Journal of Special Education*, 1981, *15*, 437-446.

Bennett, R.E., & Shepherd, M.J. Basic measurement proficiency of learning disability specialists. *Learning Disability Quarterly*, 1982, *5*, 177-184.

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

Davis, W.A., & Shepard, L.A. *The use of tests by LD teachers, school psychologists, and speech/language specialists in the diagnosis of learning disabilities*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, March 1982.

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

Hammill, D.D., Leigh, J.E., McNutt, G., & Larsen, S.C. A new definition of learning disabilities. *Learning Disability Quarterly*, 1981, *4*, 336-342.

Hastings, J.T., Runkel, P., & 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.

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)

Larsen, S.C., Rogers, D., & 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, NJ: The Gryphon Press, 1978.

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

Newcomer, P.L., & Hammill, D.D. *Psycholinguistics in the schools*. Columbus, OH: Charles E. Merrill Publishing, 1976.

Salvia, J., & Ysseldyke, J.E. *Assessment in special and remedial education*. Boston: Houghton-Mifflin, 1978.

Shepard, L.A., & Smith, M.L. An evaluation of the identification of learning disabled children in Colorado. *Learning Disability Quarterly*, 1983, *6*, elsewhere in this issue.

Shepard, L., & Smith, M.L., with Davis, W.A., Glass, G.V, Riley, A., & 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, NJ: The Gryphon Press, 1978.

Thouvenelle, S., Rader, J.R., & Hanley, T.V. *A study to evaluate procedures undertaken to prevent erroneous classification of handicapped children*. Contract Number: 300-79-0669, U.S. Department of Education, unpublished draft data, 1982.

Thurlow, M.L., & 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.

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

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

Ysseldyke, J., Algozzine, B., Regan, R., & 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., & Salvia, J.A. Diagnostic-prescriptive teaching: Two models. *Exceptional Children*, 1974, *41*, 181-186.

FOOTNOTES

[1] In Colorado learning disabilities are called perceptual-communicative disorders (PCD). The terms are conceptually equivalent, however, and the specific guidelines for identifying PCD are equivalent to national guidelines for identifying LD.

[2] The four tests with adequate validity, reliability and normative data were the WISC-R, PIAT, *Woodcock Reading Mastery*, and *Goldman-Fristoe Test of Articulation*. The *KeyMath test* received a high rating for instructional planning decisions, but not for diagnosis of LD, since it does not even have standard deviations for normative comparisons.

[3] Reviewers have suggested that perhaps the relatively high use of perceptual processing tests is unique to Colorado because learning disabilities are *called* perceptual-communicative disorders in Colorado (although the criteria are the same). Recently available national data (Thouvenelle, Rader, & Hanley, 1982) show that the high rate of use of processing tests is not unique to Colorado. Colorado's use of such tests as the *Beery* VMI is slightly higher than national figures. However, for the entire U.S. LD population the *Beery* and *Bender* were still the fourth and fifth most frequently used tests.

[4] Complete data for the correlational analyses are available in Davis and Shepard (1982).

[5] A tabular presentation of answer categories along with quotations to exemplify each category is available in Davis and Shepard (1982).

[6] 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 showing evidence of agreement.

This research was funded in part by the Colorado Department of Education as part of an evaluation study of LD identification.

Requests for reprints should be addressed to: Lorrie Shepard, Laboratory of Educational Research, Box 249, University of Colorado, Boulder, CO 80309.