Order Number 8904501

Empirical validation of discrepancy formulae in the diagnosis of learning disabilities

Menefee, Kevin L., Ph.D.

The University of Nebraska - Lincoln, 1988

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

Empirical Validation of Discrepancy Formulae in the Diagnosis of Learning Disabilities

by

Kevin Menefee

A DISSERTATION

Presented to the Faculty of

The Graduate College in the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Interdepartmental Area of
Psychological and Cultural Studies

Under the Supervision of Professors

Fred Grossman and Jack Kramer

Lincoln, Nebraska

August, 1988

TITLE

Empirical Validation of Discrepancy Fo mulae

in the Diagnosis of Learning Disabilities

BY

Kevin Menefee

APPROVED	DATE
Dr. Fred Grossman	8/12/88
Dr. Jack Kramer	8/12/88
Dr. Calvin Garbin	8/12/88
Dr. Reese Peterson	8/12/88
Dr. Ross Thompsen	8/12/88
Dr. David Moshman	8/12/88

SUPERVISORY COMMITTEE

GRADUATE COLLEGE **UNIVERSITY OF NEBRASKA**

GRAD 52002 300 5-66(T)

EMPIRICAL VALIDATION OF DISCREPANCY FORMULAE IN THE DIAGNOSIS OF LEARNING DISABILITIES

Kevin Menefee, Ph.D.
University of Nebraska, 1988
Advisors: Fred Grossman and Jack Kramer

The Education for All Handicapped Children Act has established a "severe discrepancy between ability and achievement" as the predominant model for diagnosing learning disabilities in the United States. The Act, however, did not operationalize some of the variables used to calculate a severe discrepancy, including appropriate measures of ability and achievement, magnitude of the necessary discrepancy, and means of comparing test results from different instruments. The present study systematically manipulates these variables to determine their effects upon the types of subjects who are diagnosed as LD, and examines the validity of diagnosis by the resultant discrepancy formulae.

Acknowledgements

This study was supported in part by U.S. Dept. of Education grant # G008435130. The author wishes to thank Dr. Mary Morris and Jan Leuenberger, co-directors of the Post-Secondary Intervention Model for Learning Disabilities, for their cooperation and participation in the collection of the data; Drs. Fred Grossman and Jack Kramer for their editorial review of the manuscript; and particularly, Dr. Calvin Garbin for his invaluable assistance with analysis and interpretation of the data.

Introduction

Despite decades of research, investigators and practitioners in the LD field have yet to reach an agreement on the definition of a learning disability. This lack of consensus may be because "no one definition of learning disabilities can meet the respective requirements of such diverse fields as education, psychology, medicine, and psychiatry" (Adelman & Taylor, 1986), or, more cynically, because "various persons and institutions involved in the educational/health delivery system benefit ... by the ambiguity of the definition and its regulations in federal and state law" (Senf, 1986, p. 31).

Whatever the reasons, there still appears to be (despite considerable overlap in concepts and terminology) great diversity in the theoretical and operational definitions of LD in use by researchers (Kavale & Nye, 1981; Olson & Mealor, 1981; Tucker, Stevens, & Ysseldyke, 1983), by state offices of education (Chalfant, 1985; Cone & Wilson, 1981; Frankenberger & Harper, 1987; McNutt, 1986; Mercer, Hughes, & Mercer, 1985), and by national advocacy and proponent organizations for the handicapped (Adelman & Taylor, 1986).

Mellard and Deshler (1984) discuss several major conceptual issues in the definition and diagnosis of learning disabilities. Possibly the most significant

issue is whether a learning disability represents a condition which is qualitatively or merely quantitatively distinct from the general, non-learning disabled (NLD) population. In other words, do learning disabled persons differ in "kind" from non-learning disabled persons (as illustrated in Figure 1), or do they simply represent a location on a continuum of learning ability which includes all people (as in Figure 2)?

Such a conceptual distinction corresponds fairly well to general categories of LD definitions found in the professional literature (Warner & Bull, 1986). Medical and developmental definitions (e.g., Hammill, Leigh, McNutt, & Larsen, 1981) tend to conceive of a learning disability as a qualitative state, while psychometric and socio-political definitions (e.g., U.S.O.E., 1977) view it more as part of a continuum.

From this conceptual distinction there naturally follows a corresponding distinction in models of diagnosing learning disabilities; or, more specifically, what Mellard and Deshler (1984) refer to as the clinical vs. statistical model distinction. Clinical models of diagnosis emphasize intensive individual evaluation leading to a clinical diagnosis based primarily upon the judgment of professionals (corresponding to the "qualitative" notion of LD). A major inherent difficulty with this model is the well-documented unreliability of decisions reached in practice (Ysseldyke, 1983). In

opposition to clinical models are statistical models of diagnosis, which emphasize relatively automatic decision making based on the application of mathematical formulae to scores derived from standardized assessment procedures (Holt, 1958, 1970; Sawyer, 1966). Statistical models are subject to a variety of technical difficulties related to inherent problems of psychoeducational measurement (Cone & Wilson, 1981; Hanna, Dyck, & Holen, 1979; Reynolds, 1984-1985), but have the advantage of being relatively objective in application (Warner & Bull, 1986). The criteria for classification of individuals under statistical models tends to be arbitrary, however, and in the case of LD, without empirical validation (Adelman, 1979; Berk, 1984; Warner, Schumaker, Alley, & Deshler, 1980; Ysseldyke, Algozzine, Shinn, & McGue, 1982). The criteria also have little utility for designing interventions (Gallagher, 1972), and, in current form, may result in socially and educationally undesirable practices in the identification and treatment of handicapped students (Council for Learning Disabilities, 1987; Galagan, 1985).

Although statistical models tend to be associated with the theoretical model of LD represented in Fig. 2, they don't require as a basis for diagnosis that the LD-NLD distinction be purely quantitative in nature. Statistical models only require that populations be discriminable on measurable characteristics. If LD is viewed as quantitatively distinctive, as in Fig. 2, this

requirement is met by definition - LD and NLD are defined by their location on the characteristics' continuum. The only problem is that the cut-off values on these characteristics are arbitrarily rather than empirically set (as are, potentially, the defining characteristics themselves). If the LD/NLD population actually forms a smooth continuum, there appears to be no empirical basis for distinguishing LD from NLD at any particular point (the cut-off criterion) as opposed to any other point on the continuum. Instead, decisions as to where to draw this cut-off line must be made on the basis of social, political, and ethical considerations, such as the percentage of the population to whom it is deemed desirable to provide services.

If LD is viewed as part of a continuum and diagnosed using a statistical cut-off formula in this way, classification will, by definition, be 100% accurate (within the reliability of the assessment procedure) in the same sense that an umpire is 100% accurate, because "what I calls 'em makes 'em what they is." If LD, however, exists as a qualitatively different state from NLD, statistical models can only attain a high degree of accuracy if the LD and NLD populations are mutually exclusive (or nearly so), as in Fig. 1, on some measurable characteristic. In this case, a statistical model could be employed, and presumably, empirically validated, to cleanly differentiate LD from NLD along a particular line

(Fig. 1, a). It should be noted, however, that the population means of LD and NLD on such a characteristic must be enormously far apart - somwhere on the order of three or four standard deviations (given approximately equal group variances). As Mellard and Deshler note, "no characteristic or attribute (exists) on which a difference of two standard deviations . . . (has been reported) for the learning disabled population and the general population (1984).

Assuming a qualitative distinction between LD and NLD, Figure 3 shows a more likely state of affairs, in which there is considerable overlap between the LD and NLD population on a given set of characteristics. In such a case, the proportion of correct classifications which can be made by a statistical model is limited by the degree of overlap between the populations. Use of any given cut-off score will result in substantial rates of either false-negative or false-positive misclassifications (or both). Realistically, the optimal location of a diagnostic cut-off score will be determined by the social, political, and ethical consequences of false-positive vs. false-negative misclassifications. Theoretically, however, there might be a value (e.g., line b) which would minimize the absolute number of misclassifications and maximize the distinctiveness of the two groups. If the overlap between the groups is not too great, this value should be subject to empirical identification (e.g., by a discriminant analytic technique). Comparison of group characteristics'

means would indicate the magnitude of the groups' distinctiveness.

The diagnostic problems presented by overlapping populations (as in Fig. 3) are likely to be compounded by the presumed low base rate for LD. Given that most estimates of the prevalence of LD are in the 1-15% range (Adelman, 1979), a more accurate representation of the LD and NLD populations may be that shown in Figure 4. Here, the overlap between the two populations may be complete despite significant mean differences. In such a case, the optimal cut-off score to minimize misclassification (line c) would correctly identify only 50% of the LD population, and misidentify an equal quantity of NLD individuals. Adjusting the cut-off score to identify a larger proportion of the LD population (as in line d) will actually result in misclassifying more NLD persons than correctly classifying additional LD persons (see Meehl & Rosen, 1955, for elaboration of this problem).

Inspection of Fig. 4 might logically lead one to wonder how this state of affairs differs from Fig. 2, in which LD and NLD are merely extreme ends of the same continuum. Theoretically, it might be argued (in support of Fig. 4) that the populations are, in fact, distinct, but that the distinguishing characteristics have not been discovered or are not presently measurable (as is the case with presumed minimal neurological dysfunction). Operationally, however, there is no difference. In either

case, the populations will be empirically indistinguishable and the location of an appropriate cut-off line will have to be set arbitrarily.

Since low academic achievement seems to be the most frequently cited characteristic in the diagnosis of LD (Mercer, Hughes, & Mercer, 1985; Olson & Mealor, 1981), the continuum depicted in Fig. 1 might be conceived of as representing the dimension of achievement. If LD is not to be construed as being identical with low achievement (LA) as some would claim (Warner, Schumaker, Alley, & Deshler, 1980; Ysseldyke, Algozzine, Shinn, & McGue, 1982), two conditions must hold true: 1.) there must exist real and meaningful differences between LD and LA on some dimension(s) (if not on achievement), and 2.) these differences must be reliably measurable. Since, as previously mentioned, clinical models of diagnosis appear to be generally unreliable, we must resort to developing a statistical model which can discriminate between LD and LA.

Clearly, the dominant statistical model in current use is that set forth by the Education For All Handicapped Children Act (U.S.O.E., 1977), which formulates LD as a "severe discrepancy between ability and achievement" (Mercer, Hughes, & Mercer, 1985; McNutt, 1986). Unfortunately, there exist a multitude of formulae for quantifying this general discrepancy model (Berk, 1984; Cone & Wilson, 1981) which, among themselves, classify not only significantly different rates of individuals as LD

(Algozzine & Ysseldyke, 1982; Epps, Ysseldyke, & Algozzine, 1983; Forness, Sinclair, & Guthrie, 1983; Sinclair & Alexson, 1986), but, even further, different kinds of individuals as LD (O'Donnell, 1980). This, in itself, raises serious questions about the validity of LD as a hypothetical construct (Cronbach & Meehl, 1955). If there is inadequate agreement between accepted measures of LD, the construct is lacking in convergent validity (Campbell & Fiske, 1959; Fiske, 1973). If, on the other hand, accepted measures of LD do not discriminate LD from other constructs such as LA (Ysseldyke, Algozzine, Shinn, & McGue, 1982) or other handicapping conditions (O'Donnell, 1980), the construct is lacking in discriminant validity (Campbell & Fiske, 1959). Either situation implies one of two things about the construct: it either is not validly defined, or the measures employed do not measure it validly (Cronbach & Meehl, 1955).

Cronbach and Meehl (1955) state that "a necessary condition for a construct to be scientifically admissable is that it occur in a nomological net, at least _some_ of whose laws involve observables." In other words, it must be possible to demonstrate some systematic relationships between the construct and some observable phenomena. LD, however, is a construct which tends to be defined in terms of other hypothetical constructs, such as intelligence, achievement, or minimal neurological dysfunction. There do not appear to be any agreed upon observable external

criteria for LD which are independent of the variables used to define LD operationally (Adelman & Taylor, 1986; Chalfant, 1985).

Assuming for the purposes of this study that LD is a valid construct, it follows that at least some of the measures (i.e., some of the various discrepancy formulae) do not measure it adequately. Unfortunately, in the absence of external criteria to provide cross-validation, it is not immediately apparent which, if any, of the measures might be (relatively) adequate. Cronbach and Meehl (1955) refer to a "bootstrap" procedure which seems relevant to such a situation, in which the construct and its measure serve as temporary validation of each other until their definitions can be refined to a point where observable phenomena are found which relate to them. Although this procedure is not without its critics (e.g., Bechtoldt, 1959), it might be worth applying to the construct of LD and the formulae used to diagnose it.

Review of the Literature

An Historical Review of Discrepancy Formulae

Intra-individual performance discrepancy was first used as a diagnostic sign 45 years ago in the investigation of reading problems (Monroe, 1932). Since then, the notion of LD being reflected by discrepancies between performance in diverse areas has gone through a variety of interpretations and means of measurement (see Weller, 1980, for a brief historical review). Early on, for instance, LD was diagnosed as a result of patterns of variable performance among different dimensions of aptitude in intellectual functioning (Bannatyne, 1968; Clements, 1966; Money, 1962), psycholinguistic processes (Kirk & Kirk, 1971), or perceptual-motor processes (Frostig, 1967; Johnson & Myklebust, 1967). This approach, however, was ultimately rejected as researchers failed to validate intra-aptitude variability as being indicative of learning problems (Arter & Jenkins, 1977; Hammill & Larsen, 1974; Larsen & Hammill, 1975; Ysseldyke & Salvia, 1974). Some of these investigators suggested, instead, the substitution of an aptitude-achievement discrepancy as the central basis for the diagnosis of LD (Ysseldyke & Salvia, 1974).

Since Monroe's (1932) original formula, numerous

others have been offered as appropriate means of quantifying a severe discrepancy between aptitude and achievement (see Berk, 1984, for a current review). Of those formulae which appear to enjoy the most widespread use (Frankenberger & Harper, 1987), all but two (Erickson, 1975; Shepard, 1980) compute discrepancy on the basis of a grade-level equivalent of the achievement component The use of grade equivalents, however, is no longer viewed as psychometrically supportable for computing such a discrepancy (Berk, 1984; Hanna, Dyck, & Holen, 1979; Reynolds, 1981, 1984-1985). Erickson (1975) sought to circumvent the problems associated with the use of grade and age equivalents by conversion of all metrics to standard-score form. Although this is generally viewed as a significant improvement over previous discrepancy formulae, direct standard-score conversion fails to take into account the correlation between measures of aptitude and achievement which have not been normed on the same subject sample. This failure leads to the over-identification of high-aptitude subjects and the under-identification of low-aptitude subjects (Reynolds, 1984-1985). Shepard (1980) attempted to compensate for this problem through use of a regression-score formula which adjusts standardized aptitude and achievement scores to correct for their correlation. At present, a substantial number of state offices of education employ a standard-score formula, while a small (but increasing) minority have adopted a version of the regression-score formula

(Frankenberger & Harper, 1987). It should be noted that, despite their psychometric advantages over other discrepancy formulae, the standard-score and regression-score versions still suffer from the inherent shortcomings of the statistical model discussed earlier, including lack of empirical validation or an empirical basis for determining a diagnostic cut-off value.

Components of Discrepancy Formulae

The presumption should not be made from the previous discussion that either the standard-score or the regression-score discrepancy formula are mathematically operationalizable in only one way. In actuality, the general formulae contain several unspecified variables or components whose optimal value for valid diagnosis has never been determined. For instance, the most appropriate size of the aptitude-achievement discrepancy is unclear. So is the best way to determine the measure of aptitude - when one of the Wechsler scales, for instance, is used, the Full Scale IQ is typically employed for this purpose; however, some researchers (e.g., Myklebust, 1968) have advocated using the higher of either the Verbal IQ or the Performance IQ when there is a significant difference between them, in case the learning disability is associated with a domain-specific processing disorder which would depress one of the Subscale IQ's and thereby the composite Full Scale IQ. The formulae also do not

delineate specific or appropriate measures of achievement. The United States Office of Education (1977) defines LD as "a disorder in one or more of the basic psychological processes involved in understanding or in using language," and lists seven areas in which an individual may be verified LD, including oral expression and listening comprehension. A review of the literature, however, reveals no reports of investigators using a statistical model to verify a language-learning disability in either of these areas of "achievement". That is, traditionally, only academic subject areas have been measured to determine the achievement component.

Many studies have been made of comparative classification rates of LD employing different discrepancy formulae with the same data (Cone & Wilson, 1981; Forness, Sinclair, & Guthrie, 1983; O'Donnelll, 1980; Sinclair & Alexson, 1986), typically resulting in widely varying percentages of individuals being verified. These studies, however, have all contrasted the effects of various kinds of formulae (e.g., expectancy formulae, grade equivalent formulae, mental age formulae) rather than the effects of different values of specific variables within a single formula. The only information of this kind was reported in a study by Epps, Ysseldyke, & Algozzine (1985) in which the interrelationships of 14 formulae were examined. Three of these formulae were in standard-score form with values for the aptitude-achievement discrepancy set at 10, 20, and 30 points, respectively. The correlation between

the 10 and 20 point formulae was .95; between the 20 and 30 point formulae, .84. When the formulae were factor analyzed, two factors emerged, with the 10 and 20 point formulae loading together heavily and the 30 point loading equally on each. The 30 point formula had correlations with several other formulae (which verified subjects as LD solely on the basis of low achievement) which were as high or higher than its correlations with the other standard-score formulae.

Given that some version of a standard-score based discrepancy formula is to be preferred for psychometric reasons, and presuming that differing values of the formula variables or components mentioned earlier (e.g., discrepancy size, aptitude measures, areas of achievement, etc.) would produce different rates/types of classifications, it remains to be seen if there is any basis for preferring any particular combinations of these variables over others. Despite the difficulty already mentioned of establishing independent criteria for the purpose of evaluating any given formula, there are a couple of validation methods referred to by Cronbach and Meehl (1955) that may be applicable; namely, investigation of internal structure and investigation of group differences.

Investigation of Internal Structure

Cronbach and Meehl (1955) state that "for many constructs, evidence of homogeneity within the test is

relevant in judging validity." For a "test" of LD (i.e., a particular discrepancy formula), it might be presumed that a formula which leads to the creation of relatively distinctive, cohesive LD and NLD groups would be preferrable to one which does not. Discriminant analysis is a statistical technique which computes an equation for maximally discriminating the members of two or more groups, and then applies it to those group members to determine how accurately the equation actually classifies subjects. This procedure might be useful for evaluating formulae based on the internal structure of the groups which are formed.

Cluster analysis is another statistical procedure that calculates a "distance score" between all possible pairs of subjects based upon measured differences between them. It then "clusters" groups of like subjects according to how "close" together they are on these measures. If a sample of subjects includes LD and NLD individuals, and the LD/NLD distinction entails significant differences between them on psychoeducational measures, one might expect the groups formed by a clustering procedure to correspond somewhat to this distinction (i.e., a formula might be preferred if it diagnoses individuals in a way which corresponds to empirically derived clusters). Spreen and Haaf (1986) performed a hierarchical clustering analysis on the test scores of adults who had been evaluated in childhood for learning problems together with those of a control group.

In this analysis, three of nine resulting clusters included all but two of the control subjects, the majority of whom fell into a single cluster. Although some LD subjects were included in the "control" clusters, there was nevertheless dramatic correspondence between cluster membership and LD/NLD status.

Pattern or profile analysis has a long and controversial history in association with the diagnosis of LD. Numerous investigators have proposed means of categorizing Wechsler Scale subtests (Bannatyne, 1974; Lutey, 1977; Myklebust, Bannochie, & Killen, 1971) and of basing diagnoses on patterns of subscale and subtest variance (Belmont & Birch, 1966; Vance, Wallbrown, & Blaha, 1978) as a way of distinguishing between LD and NLD individuals. In general, diagnosis based upon analysis of Wechsler Verbal-Performance IQ differences, subtest scatter, and "LD profiles" has not been validated (Ackerman, Peters, & Dykman, 1971; Anderson, Kaufman, & Kaufman, 1976), with one possible exception. Groups of LD subjects, diagnosed by a variety of methods, have often displayed relatively depressed scores on Wechsler subtests which are believed to be sensitive to difficulties with attention, concentration, and short-term memory - Kaufman's (1979) so-called Freedom from Distractibility or Third-Factor subtests (Ackerman, Dyckman, & Peters, 1976; Gutkin, 1979; Kaufman, 1981; Rugel, 1974; Smith, Coleman, Dokecki, & Davis, 1977). Although the limited magnitude