Journal of Educational Measurement[1]
Volume 20, No. 4 Winter 1983

TOWARD A THEORY OF CONSTRUCT DEFINITION

A. JACKSON STENNER
ComputerLand
MALBERT SMITH III
ComputerLand
DONALD S. BURDICK
Duke University

The process of ascribing meaning to scores produced by a measurement procedure is generally recognized as the most important task in developing an educational or psychological measure, be it an achievement test, interest inventory, or personality scale. This process, which is commonly referred to as construct validation (Cronbach, 1971: Cronbach & Meehl, 1955: ETS, 1979; Messick, 1975, 1980), involves a family of methods and procedures for assessing the degree to which a test measures a trait or theoretical construct.

Theorists and practitioners have not yet articulated a unified and systematic approach to establishing the construct validity of score interpretations. Contemporary measurement procedures are not, in general, supported by more convincing evidence that they are "measuring what they purport to measure" than instruments developed 50 years ago. The absence of persuasive, well-documented construct theories can be attributed, in part, to the lack of formal methods for stating and testing such theories.

Until the developers of educational and psychological instruments can adequately explain variation in item scale values[2] (i.e., item difficulty). The understanding of what is being measured will remain unsatisfyingly primitive. In contrast, most approaches to testing and elaborating construct theories focus on explaining person score variation. Cronbach (1971), for example, in his review of construct validity, emphasizes interpretation of correlations among person scores on theoretically relevant and irrelevant constructs. A related approach is the use of multimethod-multitrait matrices

---

[1] An expanded version of this paper was presented at the Fifth International conference on Educational Testing, Sterling, Scotland. June 30, 1982. We wish to thank the participants of the Scotland symposium, along with the anonymous reviewers of JEM for helpful comments and suggestions. We also thank E.B. Page, J.B. Carroll, and W.G. Katzenmeyer for their comments and suggestions.

[2] "Item difficulty" is the *p* value associated with an item and represents the proportion of respondents answering the item correctly (e.g., .38 or .74). Common sense would suggest that such an index be termed "item easiness", but tradition prevails. Under the Rasch Model, the "item scale values" and "person scores" are expressed in comparable scale units. In addition, the "item scale values" are not dependent on any particular sample of persons and, similarly, the "person scores" are not dependent on any particular sample of items.

of correlations between person scores (Campbell & Fiske, 1959). Explaining variation in item scale values has rarely been used as a source of information about construct validity.

The rationale for giving more attention to variation in item scale values is straightforward. Just as a person scoring higher than another person on an instrument is assumed to possess more of the construct in question (e.g. visual memory, reading comprehension, anxiety), an item (or task) that scores higher in difficulty than another item presumably demands more of the construct. The key question deals with the nature of the "something" that causes some persons to score higher than other persons and some items to score higher than other items. The process by which theories about this "something" are tested is termed "construct definition," where "definition" denotes specification of meaning (Kaplan, 1964).

Stenner and Smith (1982) illustrate construct definition theory using data from the Knox Cube Test (KCT). The KCT is purported to measure visual attention and short-term memory (Arthur. 1947). Five cubes are placed in a row before the examinee, who is to repeat various tap sequences, ranging from two-tap to seven-tap sequences.

To assess the extent to which the KCT measures short-term memory, item scale values based on a sample of 101 subjects were analyzed. A two-variable equation accounted for 93 percent of the variance in item scale values. The two variables, number of taps and distance covered between taps, have direct analogues in several theories about how information is lost from short-term memory (Gagne, 1977; Miller, 1956). That is, as the number of taps increases and the distance (in number of blocks) between tapped blocks increases, the combined effects of decay and interference serve to make the tap sequence more difficult.

Traditional approaches to testing construct theories analyze between-person variation on a construct. By examining relationships of other variables to construct scores, progressively more sophisticated construct theories are developed. As noted above, the study of relationships between item characteristics and item scale values is much less developed. Thurstone (1923) appears to have set the stage for a half century of neglect of this alternative when he argued:

I suggest that we dethrone the stimulus. He is only nominally the ruler of psychology. The real ruler of the domain which psychology studies is the individual and his motives, desires, wants, ambitions, cravings and aspirations. The stimulus is merely the more or less accidental fact. (p.364)

Considering the symmetry of the person and item perspectives, it is somewhat baffling that early correlationists were so successful in focusing the attention of psychometricians on person variation. The practice, adopted early in this century, of expressing the row (i.e., person) scores as raw counts and column (i.e., item) scores as proportions may have distorted the symmetry, leading early test constructors, with few exceptions, to ignore variation in item scores as a source of knowledge about a measurement's meaning.[3] In terms of construct definition theory, the regularity and pattern in

[3] Some investigators, including Witkin (1949, 1950) and Guttman (1969, 1971), have systematically manipulated item characteristics in testing construct theories. Guttman's work with facet analysis is particularly relevant to the approach advocated in this paper.

item-scale-value variation is no less deserving of exploration than that found in person-score variation. Only historical accident and tradition have blinded educators and psychologists to the common purpose in the two forms of analysis.

TERMINOLOGY

Constructs are the means by which science orders observations. Educational and psychological constructs are generally attributes of people, situations, or treatments presumed to be reflected in test performance, ratings, or other observations. We take it on faith that the universe of our observations can be ordered and subsequently explained with comparatively small number of constructs. Thurstone (1947) states:

Without this faith, no science could ever have any motivation. To deny this faith is to affirm the primary chaos of nature and the consequent futility of scientific effort. The constructs in terms of which natural phenomena are comprehended are man-made inventions. To discover a scientific law is merely to discover that a man-made scheme serves to unify, and thereby to simplify, comprehension of a certain class of natural phenomena. (p. 51)

The task of behavioral science is to recast the seemingly unlimited number of variously correlated observations in terms of a reduced set of constructs capable of explaining variation among observations. For example, "Why does person A have a larger receptive vocabulary than person B?", and "Why is the meaning of the word 'altruism' known by fewer people than the word 'compassion'?" The term "receptive vocabulary" is a construct label that refers simultaneously to the attribute and the construct theory which offers a provisional answer to these two questions. Construct definition is the process whereby the meaning of a construct, such as "receptive vocabulary," is specified. We impart meaning to a construct by testing hypotheses suggested by construct theories.

The meaning of a measurement depends on a construct theory. The simple fact that numbers are assigned to observations in a systematic manner implies some hypothesis about what is being measured. Instruments (e.g., tests) are the link between theory and observation, and scores are the readings or values generated by instruments. The validity of any given construct theory is a matter of degree, dependent, in part, on how well it predicts variation in item scale values and person scores and the breadth of these predictions.

Educational and psychological instruments are generally collections of stimuli hypothesized to be indicators of a particular attribute. An instrument (or individual item) may be characterized as being more or less well-specified depending upon how well a construct specification equation explains observed variation in item scale values. The specification equation links a construct theory to observations. As Henry Margenau (1978) stated:

If observation denotes what is coercively given in sensation, that which forms the last instance of appeal in every scientific explanation or prediction, and if theory is the constructive rationale serving to understand and regularize observations then measurement is the process that mediates between the two, the conversion of the immediate into constructs via number or, viewed the other way the contact of reason with nature. (p 199)

The construct specification equation affords a test of fit between instrument-generated observations and theory. Failure of a theory to account for variation in a set of item scale values invalidates the instrument as an operationalization of the construct theory and limits the applicability of that theory. Competing construct theories and associated specification equations may be suggested to account for observed regularity in item scale values. Which construct theory emerges (for the time) victorious, depends upon essentially the same norms of validation that govern theory evaluation in other sciences.

Often, a construct theory begins as no more than a hunch about why, after exposure to one another, items and people order themselves in a consistent manner. The hunch evolves as hypotheses about item and person variation, suggested by a construct theory, are systematically tested. The specification equation provides a direct means for hypotheses to interact with measurement such that construct theory and measurement each force the other into line (Kuhn, 1961).

Several terms to be used later require brief discussion. A response model is a mathematical function that relates the probability of a correct response on an item to characteristics of the person (e.g., ability) and to characteristics of the item (e.g., difficulty). Several such models are discussed by Lord (1980). Two of the most popular are the one-parameter Rasch Model (Wright & Stone, 1979) and the three-parameter logistic model (Birnbaum, 1968). In the former, person ability and item difficulty are viewed as sufficient to estimate the probability of a correct response. In the latter, two additional item characteristics, item discrimination and a guessing parameter, are used.

The response model does not embody a construct theory. The fit of data to a particular response model can be tested without knowledge of where the data came from, whether the objects are people, nations, or laboratory animals, or whether the indicants (i.e., items) are bar presses, scored responses to verbal analogy questions, or attitudinal responses. Nothing in the fit between response model and observation contributes to an understanding of what the regularity means. In this sense, the response model is atheoretical. Once a set of observations has been shown to fit a response model, the important task remains of ascribing meaning to scaled responses. In a way, the distinction is similar to that between classical reliability and validity. Like a well-fitting response model, high reliability suggests that "something" is being measured; but what that "something" is remains to be specified.

Under the present formulation, a construct can, in large part, be defined by the specification equation, whereas a test is simply a set of items from the structured universe bounded by the equation. This position might be interpreted as a neo-operationalist perspective on the relationship between construct and test. A more classical operationalism is exemplified in Bechtoldt's (1959) critique of construct validity: "Each test defines a separate ability; several tests involving the 'same content' but different methods of presenting the stimuli or of responding likewise define different abilities" (p. 677). Our position is that two tests measure the same construct if essentially the same specification equation can be shown to fit item scale values from both tests. Note that if the equation fits, it is irrelevant that these two instruments differ radically in name, method of presentation, scoring, scaling, or superficial appearance of the items. Similarly, two nominally similar tests that appear to the naked eye to measure the same construct may found to require different theories and construct labels. Such has been found to be the case in our research on certain norm-referenced reading comprehension tests. One publisher varies item difficulties by

manipulating vocabulary, sentence length, and syntactic complexity of the reading passages. Another publisher, however, standardizes these variables over passages within a test level, and manipulates the logical complexity of the questions asked about each passage. Substantially different specification equations are needed to explain variation in item scale values on the two kinds of tests. As this example indicates, even careful, expert examination of educational and psychological tests may lead to erroneous conclusions about what a test does or does not measure.

ADVANTAGES OF FOCUSING ON VARIATION IN ITEM SCALE VALUES

Construct definition theory assigns considerable importance to explaining variation in item scale values. Availability of a good fitting specification equation is viewed from both theoretical and practical perspectives as an absolutely essential feature of a measurement procedure. There are at least four advantages to focusing on item-scale-value variation as well as person-score variation in ascribing meaning to a construct.

1. *Stating theories so that falsification is possible.* There is an obvious need throughout the behavioral sciences for falsifiable construct theories that are broad enough to yield predictions beyond those suggested by common sense. Most verbal descriptions of constructs, or construct labels (with their trains of connotations and surplus meaning), are poor first approximations to theories regarding what a construct means or what an instrument measures. These verbal descriptions seldom lead to definite predictions and us are not susceptible to challenge or refutation. Outside of the response-bias literature (Edwards, 1953, 1970), few studies offer testable alternative theoretical perspectives on what an instrument measures or what a construct means. Yet, history reveals that this type of "challenge research" is precisely what fosters intellectual revolutions (Kuhn, 70). The behavioral sciences need more of the type of conflict such studies engender.

A suggested test interpretation is generally a claim that the measurement procedure measures a certain construct. Implicit in the use of any such procedure is a theory regarding the construct being measured. A major problem with the current state of educational and psychological measurement is that it is not at all clear how such claims about construct meaning can be falsified. Construct theories should be regarded as tentative, transient, and inherently doomed to falsification. A major reason for emphasizing item scale values is that theories about the meaning of a construct can be precisely stated in the form of a specification equation. Such an equation embodies a theory about the universe from which items have been sampled and simultaneously provides an objective means of confirming or falsifying theories about the meaning of scores.

2. *Generalizability of dependent and independent variables.* Estimated item scale values are typically more generalizable (i.e., reliable) than are person scores. In a simple person by items (p x i) generalizability design with person as the object of measurement, the error variance is divided by the number of items, whereas, if item is viewed as the object of measurement, the error variance is divided by the number of people (Cardinet, Tourneur, & Allal, 1976). Because most data-collection efforts involve many more people than items, the item scale values are typically more generalizable than are person scores. In most studies involving psychological instruments where the number of people is equal to or greater than 400, the generalizability coefficient for item scale values is equal

to or greater than .90. When items from nationally normed instruments are examined, the generalizability coefficient for item scale values generalizing over people and occasions approaches unity.

Many independent variables in construct specification equations can be measured with a high degree of precision. Thus, in studying the item face of the person by items matrix, it not unusual to analyze a network of relationships among variables where each variable has an associated generalizability coefficient (under a broad universe definition) approaching unity. For those researchers accustomed to working with error-ridden variables, this new perspective can be refreshing.

3. *Ease of experimental manipulation.*[4] Items are docile and pliable subjects. They can be radically altered without informed consent and can be analyzed and reanalyzed as new theories regarding their behavior are developed. Controlled experiments can be conducted in which selected item characteristics are systematically introduced and their effects on item behavior assessed. Because the experimenter can control items better than people, causal inference is simpler when the focus is on items. Items are passive; people are active subjects on whom only minimal experimental control can be exercised. All in all, items are better subjects for experimentation than people. Effects can be estimated and interpreted with less ambiguity, and the direct experimental manipulation is less costly and more efficient. Finally, this new perspective should breathe life into Cronbach's (1957) expressed hope that the experimental method will become a proper and necessary means of validating score interpretations.

4. *Intentions can be explicitly tested.* The notion that a test is valid if it measures what it purports to measure implies that we have some independent means of making this determination. Of course, we usually do not have an independent standard; consequently, validation efforts devolve into circular reasoning where the circle generally possesses an uncomfortably small circumference. Take, for example, Nunnally's (1967) statement, "A measuring instrument is valid if it does what it is intended to do" (p. 75). How are we to represent intention independent of the test itself? In the past, educators and psychologists have been content to represent intentions very loosely, in many cases letting the construct label and its fuzzy connotations reconstruct the intentions of the test developers. Unfortunately, when intentions are loosely formulated, it is difficult to compare attainment with intention. This is the essence of the problem faced by classical approaches to validity. Until intentions can be stated in such a way that attainment can be explicitly tested, efforts to assess the adequacy of a measurement procedure will be necessarily characterized by post hoc procedures and interpretations. The next section shows that construct specification equations offer a straightforward means of explicitly stating intentions and testing attainment.

[4] Although a correlational perspective is adopted in this paper, it should be noted that any number of other procedures including experimental designs are easily incorporated under the proposed methodology. Typically, a construct definition study of an existing instrument or instruments will employ regression or path analytic techniques, whereas a construct definition study of an instrument under development might employ more traditional experimental methods. The objective in either case is to understand the features of items responsible for the lawful behavior of item scale values.

BUILDING A THEORY OF RECEPTIVE VOCABULARY

In our daily lives, we use language to communicate with ourselves and others, to express emotions, to instruct, to abuse, to convey insight and intensity of feeling. Language is useful in enabling us to remember, in facilitating thought, in creating, in framing abstractions, and in adopting fresh perspectives.

To illustrate construct definition theory, we have started on a theory for receptive vocabulary. The theory attributes difficulty of items to three characteristics of stimulus words: (1) common logarithm of the frequency of a word's appearance in large samples of written material; (2) the degree to which the word is likely to be encountered across different content areas (e.g., mathematics, social studies, science, etc.); and (3) whether the word is abstract or concrete. The theory is restricted in application to pictorial representations of primary word meanings in the English language, and to nouns, verbs, adjectives, and adverbs. The construct theory emphasizes the importance of frequency of contact with words in context and thus is an exposure theory of receptive vocabulary. Words that appear frequently in writing or in speech are more likely to be learned, and thus are less difficult than are words that appear less frequently. Similarly, words that are less dispersed over a range of content areas (e.g., mathematics, cooking, music) are more difficult than words that are more widely distributed. Finally, for words of equal frequency and comparable dispersion, abstract words (i.e., referent cannot be touched, tasted, smelled, seen, or heard) are more difficult than concrete words.

The construct theory predicts that: (a) words can be scaled from easy to hard; (b) location on the scale will be highly predictable from a specification equation combining the three variables described above: and (c) person scores will correlate with any variable that reflects exposure to language in the environment. Research support for the last prediction can be marshalled from several suggestive studies that have found relationships between a child's verbal ability and the verbal ability of the teacher (Hanushek, 1972) and parents (Jordan, 1978; Kelly, 1963; Shipman & Hess. 1965; Wulbert, Inglis, Kriegsmann, & Mills, 1975). An illustration of how a construct specification equation is built and tested follows.

The instrument used in the illustration is the Peabody Picture Vocabulary Test-Revised (PPVT-R), Forms L and M (Dunn & Dunn, 1981).[5] The authors state: "The PPVT-R is designed primarily to measure a subject's receptive (learning) vocabulary for Standard American English. In this sense, it is an achievement test, since it shows the extent of English vocabulary acquisition" (p. 2). Each item has four high quality, black-and-white illustrations arranged in a multiple-choice format. The subject's task is to select the picture that best illustrates the meaning of a word spoken by the examiner.

The Rasch item scale values for the 350 words of the PPVT-R were used as the dependent variable in the analysis. The three predictor variables were word frequency, dispersion, and abstractness. Word frequency and dispersion of each word were obtained by referencing each keyed response word in the extensive word sample of Carroll, Davies, and Richman (1971). The Carroll et al. sample is based upon words selected from schoolbooks used by third- to ninth-grade students in the United States. The third measure, abstraction, was based upon independent ratings of each stimulus

[5] The authors thank Dr. Gary Robertson and American Guidance Service for providing norming data for the PPVT-R.

picture by two educational psychologists. Each of these predictors is described in more detail below.

1. *Log frequency.* Our early work used the common log of the frequency with which a word appeared in a running count of 5,088,721 words sampled by Carroll et al. (1971) from a broad range of written materials. We subsequently discovered that the log frequency of the "word family" was more highly correlated with word difficulty. Consequently, this variable was used in the latter stages of our work. A "word family" included: (1) the stimulus word; (2) all plurals (adding s or changing y to ies); (3) adverbial forms; (4) comparatives and superlatives; (5) verb forms (s, d, ed, ing); (6) past participles; and (7) adjective forms. The frequencies were summed for all words in the family and the log of that sum was entered in the specification equation.

2. *Dispersion.* This variable is a measure of dispersion of word frequencies over 17 subject categories specified by Carroll et al. (e.g., literature, mathematics, music, art, shop). It appears to be highly useful in differentiating words that are equally likely to appear in different kinds of verbal materials from words that tend to appear only in specialized subject matter. The values for dispersion range from 0 to 1.0. Dispersion approaches the value of 0 for words that are concentrated in only a few content areas, whereas dispersion approaches the value of 1.0 when the word is distributed fairly evenly over the entire range of contents.

3. *Abstract/concrete.* As noted above, the abstract/concrete ratings were made independently by two educational psychologists. Although several investigators have employed a Likert-type scale for the measurement of this variable (e.g., Paivio, Yuille, & Madigan, 1968; Rubin, 1980), the operational definition of abstraction lends itself to a dichotomous classification. Thc dichotomy was created to differentiate words that refer to tangible objects (e.g., car, coin, ice cream) from words denoting concepts (e.g., transportation, money, sweet). Abstract words were assigned a value of 1, and concrete words were assigned a value of 0. Inter-observer agreement for the two forms was 92 percent on Form L and 95 percent on Form M.

Table 1 presents the means, standard deviations, intercorrelations, raw score regression equation, and R^2 for the PPVT-L and PPVT-M. The construct specification equation for Form L explains 72 percent of the observed score variance in item scale values. Frequency and dispersion are highly correlated ($r = .848$) indicating that less frequently appearing words are likely to be concentrated in a small number of content areas, whereas more frequently appearing words tend to be more uniformly distributed over content areas. The abstract/concrete dichotomy is moderately correlated with item scale values ($r = .352$) but negligibly correlated with frequency ($r = -.033$) and dispersion ($r = -.081$). All relationships are consistent with the exposure theory: Less frequently appearing words are more difficult than more frequently appearing words, and abstract words are, on average, more difficult than concrete words.

A comparison of Forms L and M reveals several similarities. First, there is less than .15 standard deviation difference in the means of item scale values, frequencies, and dispersions. The largest difference appears on the abstract/concrete dichotomy where Form M includes 67 percent abstract words and Form L only 60 percent. Second, the variances are remarkably similar. Third, the pattern of intercorrelations appears to be largely invariant over forms of the PPVT. Last, the proportions of variance explained are similar (Form L, $R^2 = .722$; Form M, $R^2 = .712$).

The combined Form L and M analysis is presented in Table 2. Item scale values for 350 items were regressed on the three variables in the construct specification equation. As might be expected, the results are very similar to those obtained in the individual Form L and M analysis. The R^2 on the aggregate data set is .713 and the pattern of intercorrelations remains consistent with the previous results. The only discernible difference is in the expected reduction of the standard errors of the coefficients with the larger data set.[6]

The amount of shrinkage was examined by cross-validating each equation on data from the opposite form. When the Form L equation was applied to Form M data, the R^2 fell slightly from .722 to .702. Likewise, application of the Form M equation to Form L data resulted in a decrement from .712 to .709. As the cross-validation studies indicate, the variables in the construct specification equation are stable and reliable.

In an attempt to improve the amount of variance explained, 50 additional variables were examined for inclusion in the specification equation. Among the variables considered were: (1) part of speech; (2) phonetic complexity of the word; (3) number of letters; (4) modal grade level at which the word appears in school materials; (5) distractor characteristics; (6) experiential frequency; (7) content classification of word; (8) semantic features; (9) frequencies based upon the Kucera and Francis (1967) corpus; and (10) numerous interactions and polynomials. The results of this search were disappointing in that only negligible improvement in fit was obtained when the number of variables in the equation was increased to 8, 10, or 12 predictors.

CONCLUSION

Throughout the history of educational and psychological measurement, there has been an abundance of attention to the statistical characteristics of measurement procedures and a disturbing lack of concern for the role that theory plays or should play in measurement We have confused quality of measurement with statistical sophistication and high internal consistency coefficients. The related notions of construct theories and specification equations may provide a vehicle for restoring theory as the foundation of measurement in psychology and education. One conclusion seems certain: Much can be learned from attempts at building construct specification equations for the major instruments used in education and psychology. Our expectation is that there will be no shortage of surprises once such work begins.

---

[6] It should bc noted that heterogeneity of the item set is not a satisfactory explanation for the explanatory powers of the specification equation. we have examined the fit of the equation to items with drastically restricted ranges and found the R^2 to be smaller (as we might expect), but not dramatically so. Furthermore, the form of the equation remained invariant.

Table 1

Descriptive Data and Construct Specification Equation for Forms L and M

	Form L				Form M			
	Rasch Item Scale Value	Log of Frequency	Dispersion	Abstract/ Concrete	Rasch Item Scale Value	Log of Frequency	Dispersion	Abstract/ Concrete
Rasch Item Scale Value		-.776	-.736	.352		-.783	-.791	.359
Log of Frequency			.848	-.033			.870	-.146
Dispersion				-.081				-.188
Mean	98.94	1.72	.49	.60	98.82	1.80	.52	.67
S.D.	26.05	.86	.26	.49	25.95	.91	.27	.47
Regression Coefficients		-17.53	-21.84	16.71		-11.59	-38.65	12.48
Standard Error of Coefficients		2.30	7.68	2.15		2.38	8.18	2.31
	R^2 = .722				R^2 = .712			

Table 2

Construct Specification Equation for Combined Forms of PPVT-R

	Rasch Item Scale Value	Log of Frequency	Dispersion	Abstract/Concrete
Rasch Item Scale Value		-.779	-.763	.354
Log of Frequency			.859	-.086
Dispersion				-.130
Mean	98.88	1.76	.51	.64
S.D.	25.96	.89	.26	.48
Regression Coefficients		-14.57	-29.73	14.70
Standard Error of Coefficients		1.65	5.61	1.57
	R^2 = .713			

REFERENCES

ARTHUR, G. *A point scale of performance tests.* New York: Psychological Corp., 1947.

BECHTOLDT, H. P. Construct validity: A critique. *American Psychologist*, 1959, 14, 619-629.

BIRNBAUM, A. Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, 56, 81-105.

CARDINET, J., TOURNEUR, Y., & ALLAL, L. The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement*, 1976, 13, 119-134.

CARROLL, J. B., DAVIES, P., & RICHMAN, B. *Word frequency book.* Boston: Houghton **Mifflin, 1971.**

CRONBACH, L. J. The two disciplines of scientific psychology. *American Psychologist*, 1957, 12, 671-684.

CRONBACH, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971.

CRONBACH, L. J., & MEEHL, P. E. Construct validity in psychological tests. *Psychological Bulletin,* 1955, 52, 281 -302.

DUNN, LLOYD M., & DUNN, LEOTA M. *Manual for Forms L and M of the Peabody Picture Vocabulary Test-Revised.* Circle Pines, Minn.: American Guidance Service, 1981.

EDWARDS, A. L. *Edwards Personal Preference Schedule.* New York: Psychological Corp., 1953.

EDWARDS, A. L. *The measurement of personality traits by scales and inventories.* New York: Holt, Rinehart, & Winston, 1970.

EDUCATIONAL TESTING SERVICE. *Construct validity in psychological measurement: Proceedings of a colloquium in theory and applications in education and employment.* Princeton, N. J.: Henry Chauncey Conference Center, October 1979.

GAGNE, R. M. *The conditions of learning.* New York: Holt, Rinehart, & Winston, 1977.

GUTTMAN, L. Integration of test design and analysis. In *Toward a theory of achievement measurement: Proceedings of the 1969 invitational conference on testing problems.* Princeton, N. J.: Educational Testing Service, 1969.

GUTTMAN, L. Measurement and structural theory. *Psychometrika*, 1971, 36, 329-347.

HANUSHEK, E. A. *Education and race: An analysis of the educational production process.* Lexington, Mass.: Lexington Books, 1972.

JORDAN, T. E. Influences on vocabulary attainment: A five year prospective study. *Child Development*, 1978, 49, 1096-1106.

KAPLAN, A. *The conduct of inquiry.* San Francisco: Chandler, 1964.

KELLY, S. The social world of the urban slum child: Some early findings. American Journal of Orthopsychiatry, 1963, 33, 823-831.

KUCERA, H., & FRANCIS, W. N. *Computational analysis of present day American English.* Providence, R. I.: Brown University Press, 1967.

KUHN, T. S. The function of measurement in modern physical science. In H. Woolf (Ed.), *Quantification: A history of the meaning of measurement in the natural and social sciences.* Indianapolis: Bobbs-Merrill, 1961.

KUHN, T. S. *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press, 1970.
LORD, F. M. *Applications of item response theory to practical testing problems.* Hillsdale, N. J.: Erlbaum, 1980.
MARGENAU, H. U. *Physics and philosophy. Selected essays*. Dordrcht. Holland: Reidel, 1978.
MESSICK, S. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist,* 1975, 30, 955-966.
MESSICK, S. Test validity and the ethics of assessment. *American Psychologist*, 1980, 35, 1012-1027.
MILLER, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 1956, 63, 81-97.
NUNNALLY, J. C. *Psychometric theory*. New York: McGraw-Hill, 1967.
PAIVIO, A., YUILLE. J. C., & MADIGAN, S. A. Concreteness, imagery and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 1968, 76, 1-25.
RUBIN, D. C. 51 properties of 125 words: A unit of analysis of verbal behavior. *Journal of Verbal Learning and Verbal Behavior,* 1980, 17, 736-755.
SHIPMAN, V., & HESS, R. D. Social class and sex differences in the utilization of language and the consequences for cognitive development. Paper presented at the meeting of the Midwestern Psychological Association. Chicago, 1965.
STENNER. A. J., & SM ITH. M. Testing construct theories. *Perceptual and Motor Skills.* 1982, 55 415-426.
THURSTONE, L. L. The stimulus-response fallacy in psychology. *Psychology Review*, 1923, 30, 354 -369.
THURSTONE, L. L. *Multiple factor analysis*. Chicago: University of Chicago Press, 1947.
WITKIN, H. A. Perception of body position and of the position of the visual field. *Psychological Monographs,* 1949, 63, (1, Whole No. 302).
WITKIN, H. A. Perception of the upright when the direction of the force acting on the body is changed. *Journal of Experimental Psychology*, 1950, 40, 93-106.
WRIGHT, B. D., & STONE, M. H. *Best test design.* Chicago: MESA Press, 1979.
WULBERT, M., INGLIS, S., KRIEGSMANN, E., & MILLS, B. Language delay and associated mother-child interactions. *Developmental Psychology*, 1975, 11, 61-70.

AUTHORS

A. JACKSON STENNER Address: ComputerLand, 4125 Chapel Hill Blvd., Durham, NC 27707. Title: Vice President. Degrees: B.S., B.A., University of Missouri; ABD, Duke University. Specializations. Measurement and evaluation.
MALBERT SMITH Address: ComputerLand, 4125 Chapel Hill Blvd., Durham, NC 27707. Title: Vice President. Degrees: B.A., Duke University; M.Ed., Ph.D., University of North Carolina, Chapel Hill. Specializations: Measurement and evaluation.
DONALD S. BURDICK Address: Dept. of Mathematics, Duke University, Durham, NC 27707. Title. Associate Professor of Mathematics and Biomedical Engineering. Degrees. B.S., Duke University; M.A.. Ph.D., Princeton University. Specializations. Multivariate data analysis, linear models.