Perceptual and Motor Skills, 1982, 55, 415-426.

TESTING CONSTRUCT THEORIES

A. JACKSON STENNER AND MALBERT SMITH, III
NTS Research Corporation[1]

Summary: This paper presents and illustrates a novel methodology, construct-specification equations, for examining the construct validity of a psychological instrument. Whereas traditional approaches have focused on the study of between-person variation on the construct, the suggested methodology emphasizes study of the relationships between item characteristics and item scores. The major thesis of the construct specification-equation approach is that until developers of a psychological instrument understand what item characteristics are determining the item difficulties, the understanding of what is being measured is unsatisfyingly primitive. This method is illustrated with data from the Knox Cube Test which purports to be a measure of visual attention and short-term memory.

The process of ascribing meaning to scores produced by a measurement procedure is generally recognized as the most important task in evaluating a psychological measure, be it an achievement test, interest inventory, or personality scale. This process, which is commonly referred to in the literature as construct validation (Cronbach & Meehl, 1955; Cronbach, 1971; Thorndike & Hagen, 1577; Jensen, 1980), involves a family of methods and procedures for assessing the degree to which a test measures the trait or theoretical construct the test was designed to measure.

While there is general consensus among measurement theorists regarding the prominent role of construct validity, theorists and researchers have not articulated a unified and systematic approach to establishing the construct validity of score interpretations. As Buros (1977) noted in his assessment of the past fifty years in testing, there has been little progress in the field of testing except for tremendous advances in computer technology. His observations seem particularly accurate in terms of the current status of construct validity. Contemporary measurement procedures do not, in general, provide any more convincing evidence that they are measuring what they purport to measure than instruments developed fifty 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.

A major thesis of this paper is that until the developers of a psychological instrument can adequately explain variation in item scores (i.e., difficulty), the understanding of what is being measured is unsatisfyingly primitive. Most approaches to testing and elaborating construct theories focus on explaining person score variation.[2] Cronbach (1971), for example, in his review of

---

[1] 2634 Chapel Hill Boulevard, Durham, North Carolina 27707.

[2] Throughout this paper we distinguish a "construct theory" from a broader "data theory". A construct theory accounts for a regularity in a set of observations generated from a measurement procedure. A data theory interrelates a set of constructs in a nomological network.

construct validity, emphasizes the formal testing of nomological networks and interpretation of correlations among person scores on theoretically relevant and irrelevant constructs. Except for isolated investigations, explaining variation in item scores has not been viewed by methodologists or practitioners as a particularly protean source of information about construct validity.

The rationale for giving more attention to variation in item scores is straightforward. Just as a person scoring higher than another person on an instrument is assumed to possess more of the construct in question (i.e., visual memory, reading comprehension, anxiety), an item (or task) which scores higher (in difficulty) than another item must be viewed as demanding 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 validation."

There are essentially two routes one can travel in testing construct theories. The traditional route is the study of between-person variation on the construct. By examining relationships between potential determinants and construct scores, progressively more sophisticated construct theories are developed. The second, much less traveled, route involves the study of relationships between item characteristics and item scores. Thurstone (1923) appears to have set the stage for a half century of neglect of this latter 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 in the person and item perspectives on construct validation, it is somewhat baffling that Thurstone, Terman, Binet, and Goodenough were so successful in focusing the attention of early correlationists exclusively on person variation. The practice, adopted early in the 1900s of expressing the row (i.e., person) as raw counts and column (i.e., item) scores as proportions may have distorted the symmetry, leading early test constructors, with few exceptions, to abandon variation in item scores as a source of knowledge about a measurement's meaning. From the perspective of construct validation, the regularity and pattern in item-score variation is no less deserving, of exploration than that found in person-score variation. Only historical accident and tradition have blinded behavioral researchers to the singularity of purpose inherent in the two forms of analysis.

Perhaps the only exception to this blindness has occurred in the field of cognitive psychology. This branch of psychology has been very diligent in analyzing items as a data source in understanding the components of behavior. As Pellegrino and Glaser (1982) explain, "performance on psychometric test tasks becomes the object of theoretical and empirical analyses" (p. 275). Similarly, Sternberg (1977, 1980), Carroll (1976), and Whitely (1981) have focused upon item variation in attempting to explicate the cognitive processes required by certain psychometric tasks. As the next section indicates, there are several advantages in adopting such an approach.

Advantages in Analyzing Item-score Variation

There are advantages to studying item-score variation in the process of validating a construct theory. In this section we shall examine three such advantages.

(1) Stating theories so 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, as well as construct labels with their trains of connotations and surplus meaning, serve as poor first approximations to deductive theories regarding what a construct means or what an instrument measures. These verbal descriptions seldom lead to predictions and thus are not susceptible to challenge or refutation. Outside of the literature on response bias there are few studies indeed that offer testable alternative theoretical perspectives on what an instrument measures or what a construct means. Yet history shows that this type of "challenge research" is precisely what fosters intellectual revolutions (Kuhn, 1970). The behavioral sciences need more of the type of conflict such studies engender.

A suggested test interpretation is a claim that a procedure measures a construct. Implicit in the use of any such procedure is a theory regarding the nature of the construct being measured. A major problem with the current state of behavioral science measurement is that it is not at all clear how such claims about construct meaning can be falsified. Unless and until construct interpretations are framed in ways that are in practice as well as in principle falsifiable, the behavioral sciences will not progress beyond a disjointed collection of instrumentation and methodological curios.

A major reason for emphasizing item-score variation is that theories about the meaning of a construct can be precisely stated in the form of a construct-specification equation (discussed at length in the next section). Such an equation embodies a theory of item-score variation and simultaneously provides a straightforward means of confirming or falsifying alternative theories about the meaning of scores generated by a measurement procedure.

(2) Higher generalizability of dependent and independent variables is possible. Item scores are typically more generalizable (i.e., reliable) than are person scores. In a simple persons X items generalizability design with person as the object of measurement, the error term is divided by the number of items, whereas, if item is viewed as the object of measurement, the error term is divided by the number of people. Since most efforts to collect data involve many more people than items, the item scores are typically more generalizable than are person scores. In most studies involving psychological instruments where the number of people is greater than or equal to 400, the generalizability coefficient for item scores is greater than or equal to .95. When items from nationally normed instruments are examined, the generalizability coefficient for item scores generalizing over people and occasions approaches unity.

It is also true that most measures used as independent variables in construct-specification equations can be measured with considerable precision. In studying the item face of the person by items matrix, it is not unusual to find one's self analyzing 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) Experimental manipulation is possible. 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. Similarly, controlled experiments can be conducted in which selected item characteristics are systematically introduced and their effects on item behavior assessed. Also, causal inference is strengthened when one employs items as subjects rather than people, due to the fact that the experimenter can control items better than people. Items are passive subjects, whereas people are active subjects on whom only minimal experimental control can be exercised; threats to the validity of a study are more likely to be operative when people are studied. 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 expressed hope that the experimental method be viewed as a proper and necessary means of validating interpretations of psychological scores.

Construct-specification Equation

A "measurement procedure" requires (1) an "item format" which details the general framework of fixed item characteristics within which the item content is to be varied, (2) the "construct-specification equation" which operationally defines the construct and presents the item characteristics and associated regression weights to be used in generating items and systematically varying their difficulty, (3) a sampling plan for selecting items from the universe bounded by the specification equation, (4) a plan for selecting persons and assigning them to items, (5) administrative procedures which govern the context in which the measurement procedure is administered, and (6) rules for assigning scores to persons and items.

A construct-specification equation is developed by regressing item scores on selected characteristics of the items. The results from this multiple regression analysis yield two important sources of information. First, the R^2 index indicates the amount of variance in item scores accounted for by the model.[3] Second, the analysis provides a regression equation that indicates which characteristics of the items are important in predicting item scores.

The construct-specification equation expresses a theory regarding observed regularity in a set of observations generated by a measurement procedure. The equation sets forth a theory of item-score variation and simultaneously provides the vehicle for confirmation or falsification of the theory. Just as important, alternative theoretical formulations can be tested side by side with criteria for priority unambiguously stated.

As an example, suppose a researcher develops a measure of "career maturity" and asserts that the 100-item instrument measures "student's awareness and knowledge about the world of work and his

[3] An "item difficulty" is the *p* value associated with an item and represents the proportion of respondents answering the item correctly (i.e., .68 or .74). Common sense suggests that such an index be termed "item easiness" but tradition prevails. An "item score" analogous to a "person score" is a scaled transformation of the item difficulty. Under the Rasch model the "item scores" and "person scores" are expressed in comparable scale units. In addition, the "item scores" are not dependent on any particular sample of persons, and similarly the "person scores" are not dependent on any particular sample of items.

role in it." Further suppose that another researcher examines the instrument and concludes that it is merely a poor measure of "reading comprehension." Employing the perspective encouraged in this paper, both researchers would be charged with building specification equations that embody the two competing construct theories. Both equations would be applied to the same dependent variable (i.e., same set of item scores) and the data analyzed. If the reading comprehension theory predominated, the "career maturity" theorist could revise his instrument systematically controlling reading comprehension while at the same time either modifying the specification equation to account better for observed item scores or revising items so that they conform better to predictions generated from the construct theory. Whatever the outcome of this hypothetical research, the direct confrontation between construct theories would have sharpened both theory and instrumentation.

Note that under the present formulation a construct is not operationally defined by a particular instrument but rather by a corresponding specification equation. The construct validity of a score interpretation is assessed, in part, by how closely observations generated from a measurement procedure conform to predictions made by the specification equation. If a specification equation explains a substantial proportion of variation in item scores, then person scores can be interpreted in light of the construct theory with increased confidence that such interpretations are valid. A high R^2 increases confidence in the construct theory, as well as the particular measurement procedure under examination, whereas a low R^2 raises doubts about the construct theory, the instrument, or both.

Similarly, the construct validity of a single item is assessed by the magnitude of its residual. A small residual indicates that an item conforms well to specifications and is not measuring unspecified (i.e., unwanted) influences. A large residual suggests that unspecified sources of item variability account for an item's score. Note the constant interplay between formulation of theory and development of an instrument. Given a specification equation it is possible to develop highly valid items and discard invalid items with large residuals. Conversely, examination of item residuals may suggest new variables that can be used to modify a construct theory and improve the predictive power of the specification equation. Theory and instrument are thus bound in a dynamic interplay, with a change in one implying modification of the other.

Observations are generated through applications of measurement procedures. Such observations are meaningless unless interpreted by a construct theory. A good construct theory explains the observed regularity in a set of observations and thus imparts meaning to a set of person or item scores. Construct-specification equations aid us in making explicit our theories about what is being measured by a particular collection of items. As more comprehensive construct theories emerge, predictions about observations (on both the person and item face) become more precise, and broader generalizations are possible.

One pretender to legitimate explanation is the practice of describing the same thing in a new language and passing the translation off as a contribution to understanding. As, for example, when judges rate the complexity of a group of test items and this rating is then employed in accounting for variance in item scores (Kirsch & Guthrie, 1980). Unless the reasoning underlying the judgments is explicated, such activity contributes little to construct validation. An individual is not incapable of tenderness because he is unloving and an item is not difficult because it is complex. Such pseudo explanations are not only theoretically sterile but intellectually quite unsatisfying.

Construct-specification equations provide a useful, objective methodology for explicating the components which account for the complexity in a set of items. We turn now to an illustration of how a construct-specification equation can be built and used.

Illustration

To illustrate the development of a "construct-specification equation," we have conducted an analysis of data from the Knox Cube Test which purports to be a measure of visual attention and short-term memory (Arthur, 1947). The test uses five 1-in. cubes; four of the cubes are fixed 2 in. apart on a board, and the fifth cube is used to tap a series on the other four. To avoid confusion we will arbitrarily number the cubes from left to right as "1," "2," "3," and "4." The easiest item is a two-tap item in which the examiner taps cube 2, then cube 3. The most difficult item is a seven-tap item in which the sequence is 4, 1, 3, 4, 2, 1, 4.

Theoreticians have conceptualized short-term memory as an individual's "working" memory span. It is the immediate memory in a person's cognitive system which appears to have a very limited capacity in terms of the amount of material that can be retained and the length of time it can be retained. The limited nature of short-term memory, in terms of volume of storage, tends to be around 7 + 2 bits of information (Miller, 1956), while the limit in terms of time appears to be 18 to 20 sec. (Gagne, 1977).

Most psychologists have postulated that material is lost (forgotten) from short-term memory due to two factors, interference and/or time decay. Due to limited working capacity, the strength of memory traces is diminished as more bits of information are presented. As more demands are placed upon short-term memory, the recall of any particular memory trace becomes more difficult. That is, as the background complexity increases, the probability of retrieving a particular trace (signal) from the background (noise) is reduced.

The second determinant of loss from short-term memory is time. The probability of retrieving an item from short-term memory is negatively correlated with the passage of time. The longer an item stays in memory, the weaker it becomes, until at some point it is completely lost.

To assess the extent to which the Knox Cube Test measures the construct, "short-term memory," item scores computed on a sample of 101 subjects ages 3 to 16 yr. were subjected to the previously described analysis. First, the items were arranged in order of difficulty from the easiest to the most difficult. Second, each item was carefully examined in an attempt to ascertain whether items differed from one another in ways consistent with a "short-term memory" construct interpretation. In trying to identify characteristics of the Knox Cube Test items which accounted for their difficulty, we selected the following attributes: (1) number of taps, (2) number of reversals, and (3) total distance covered. Table 1 presents difficulty indices for the items along with the tapping sequence and data on the three identified facets of each item.

TABLE 1

DISTRIBUTIONS OF 106 CHILDREN'S RESPONSES TO THE KNOX CUBE TEST*

Item No.	Tapping Sequence	Rescaled Rasch Item Difficulties	No. of Taps	No. of Reversals	Distance Covered
2	2-3	1.0	2	0	2
3	1-2-4	1.4	3	0	4
7	1-2-3-4	3.0	4	0	4
4	1-3-4	3.2	3	0	4
6	3-4-1	4.5	3	1	5
5	2-1-4	4.7	3	1	5
8	1-4-3-2	4.7	4	1	6
9	1-4-2-3	5.2	4	2	7
10	1-3-2-4	5.2	4	2	6
15	1-2-3-4-3	5.7	5	1	5
11	2-4-3-1	6.2	4	1	6
13	2-1-4-3	6.4	4	2	6
16	1-2-3-4-2	6.5	5	1	6
14	4-2-1-3	6.7	4	1	6
12	3-1-4-2	7.1	4	2	8
18	1-3-2-4-3	8.6	5	3	7
17	1-3-1-2-4	9.2	5	2	8
19	1-4-3-2-4	9.2	5	2	8
20	1-4-2-3-4-1	10.1	6	3	11
21	1-3-2-4-1-3	10.6	6	4	11
22	1-4-2-3-1-4	10.6	6	4	12
23	1-4-3-1-2-4	11.7	6	2	10
25	3-2-4-1-3-4-2	13.0	7	4	12
24	4-1-3-4-2-1-4	13.5	7	3	13

*Taken from Best Test Design (Wright & Stone, 1979).

The descriptive statistics for these variables are presented in Tables 2 and 3. Table 2 presents the mean, standard deviation, and range for each variable, while Table 3 presents the correlation matrix.

An examination of these two tables provides several characteristics of the Knox Cube Test. First, the three identified facets (number of taps, number of reversals, and distance covered) seem to be important determinants of item difficulty. As the zero-order correlations indicate, items become harder as the distance covered increases ($r = .95$), as the number of taps increases ($r = .94$) and as the number of reversals increases ($r = .87$). Second, there is substantial multicolinearity among the independent variables. For example, as the number of taps increases, there are concomitant increases in the distance covered ($r = .90$) and the number of reversals ($r = .82$).

TABLE 2

MEANS, STANDARD DEVIATIONS, AND RANGE
OF FOUR ITEM FACETS
OF THE KNOX CUBE TEST

Variable	*M*	*SD*	Range
Item Difficulty	7.00	3.44	1 – 13.5
Number of Taps	4.54	1.32	2 – 7
Number of Reversals	1.75	1.26	0 – 4
Distance Covered	7.17	2.94	2 – 13

TABLE 3

CORRELATIONS AMONG THE FOUR ITEM FACETS
OF THE KNOX CUBE TEST

	No. of Taps	No. of Reversals	Distance Covered
Item Difficulty	.94	.87	.95
Number of Taps		.82	.90
Number of Reversals			.90

TABLE 4

RESULTS OF MULTIPLE REGRESSIONAL ANALYSIS OF
ITEM DIFFICULTIES ON ITEM CHARACTERISTICS

	df	SS	R^2	*F*	*p*	Predictors	Beta	*F*
Regression	2	253.23	.93	139.8	.01	Distance covered	.56	14.8
Residual	21	19.03				Number of taps	.43	8.8
Total	23	272.56						

To develop and refine the construct specification equation, a series of multiple regression analyses were conducted. First, hierarchical set-wise multiple regression was utilized in which the three main effects were entered as a first set, and the three two-way interactions were entered as a second set. The first set of main effects accounted for 93% of the variance in item difficulties. Although the addition of the set of interactions increased the variance explained to 96%, the incremental variance accounted for was not statistically significant ($F = 4.7$). In the second analysis, the three independent variables were entered in a stepwise fashion. The results of this analysis are presented in Table 4.

An inspection of Table 4 indicates that the construct-specification equation requires only two variables, distance covered and number of taps. With this two-variable model, we can account for 93% of the variance in item difficulties; we have a firm understanding of what is governing item difficulties on the Knox Cube Test. This model predicts that items become more difficult as the distance covered is increased and the number of taps is increased. Since there is a high degree of multicolinearity among the predictors, the regression results were subjected to commonality analysis. The commonality analysis indicated that each variable accounted for only a small amount of unique variance. The unique contribution of distance covered and the number of taps was .05 and .03, respectively. The variance shared in common between the two variables was .85.

The empirically generated two-variable model of the Knox Cube Test is consistent with our theoretical understanding of the processes which account for loss of information from short-term memory. As previously discussed, the loss of information can be attributed to two factors, interference and time decay, which are highly interrelated. Distance covered corresponds directly with interference, and the number of taps serves as a proxy for time decay. Interference, which can be conceptualized as the ratio of signal to noise, is represented by the distance covered in a sequence. A two-tap sequence such as "2-3" is easier than a two-tap sequence which is spread out over more blocks, such as "1-4." In the latter case, there is more background noise which the individual must process and filter out.

The second factor, time decay, also has its analogue in the item components in terms of the number of taps. As the number of taps increases, the string of blocks to be recalled increases the memory load not only in terms of number but also in terms of time. That is, the second individual block tapped in a four-tap sequence must be held in memory twice as long as the second block in a two-tap sequence.

The construct-specification equation derived from the Knox Cube Test provides satisfying evidence that it is measuring what it purports to measure. From this illustration, we now have a more clear and precise understanding of what characteristics govern the item difficulties. Difficulty is determined by two item characteristics (distance covered and number of taps) which conform nicely to the previously discussed construct theory. In addition to providing confirming evidence for the construct theory, the specification equation has an immediate practical application. New items, with predictable item difficulties, can be quickly and easily generated by manipulating the specification equation.

The construct-specification equation provides a means for testing different conceptualizations of a hypothesized construct. For example, if a researcher suspects that a self-concept instrument is actually measuring verbal reasoning and social desirability, then this hypothesis can be empirically tested by translating the hypothesis into a specification equation. If the new equation reflecting the parameters of verbal reasoning and social desirability of the items accounts for a large percentage or the variance, then the test is not measuring the construct its authors intended. Second, the development and utilization of construct-specification equations permits the test designer to control systematically unwanted sources of variation. If a wide range vocabulary scale was contaminated by items requiring fine discriminations among word meanings, the resulting construct-specification equation could be employed to create items which systematically controlled such influences.

Note that issues of item bias and culture-fairness are easily addressed under this procedure. LISREL (Joreskog, 1976) can be used to test the hypothesis that the same item-specification equation fits black, white, and Spanish populations. If it does not, then items with high or low residuals can be identified and either modified or discarded. Through an iterative procedure, an equation that fits all language and cultural groups might be identified, ensuring that the measurement procedure is valid (i.e., measures the same thing) for each group.

An admittedly untested intuitive leap is that for each component identified as an important determinant of item scores, there exists an analogous person characteristic. Such an inference could be tested by developing relatively pure component tests and determining how well person scores on these components predict person scores on the more componentially complex instrument. Profiles of person scores on these components might prove more useful to teachers and diagnosticians than current score profiles computed from scores of unknown composition.[4]

Conclusion

As Cronbach (1957) has noted, a test interpretation is a claim that a test measures a construct:

To decide how well a purported test of anxiety measures anxiety, construct validation is necessary, i.e., we must find out whether scores on the test behave in accordance with the theory that defines anxiety. This theory predicts differences in anxiety between certain groups, and traditional correlational methods can test those predictions. But the theory also predicts variation in anxiety, hence in the test score, as a function of experiences or situations, and only an experimental approach can test those predictions (p. 767).

We believe that much can be gained from a more balanced perspective than that suggested by Cronbach and others with its nearly exclusive emphasis on person-score variation. Let us return the stimulus (i. e., item) to its rightful place on the throne and let it share equally in the load of testing construct theories. Item-score variation is far more than an "accidental fact" and much can be learned about how and why individuals vary from an examination of how and why items vary.

---

[4] While conceptually very similar to componential analysis (Sternberg, 1977, 1980), the construct-specification equation approach differs in terms of its orientation and level of inference. Although both methodologies employ linear regression models, componential analysis focuses upon the cognitive processes (components) that occur in a task, whereas construct-specification equations focus on concrete observable characteristics of the task (i.e., item).

REFERENCES

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

BUROS, O. K. Fifty years in testing: some reminiscences, criticisms, and suggestions. *Educational Researcher*, 1977, 6, 9-15.

CARROLL, J. B. Psychometric tests as cognitive tasks: a new structure of intellect." In L. B. Resnick (Ed.), *The nature of intelligence.* Hillsdale, NJ: Erlbaum, 1976. Pp. 27-56.

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. Pp. 443-507.

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

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

JENSEN, A. R. *Bias in mental testing.* New York: Free Press, 1980.

JORESKOG, K. G. Structural equation models in the social sciences: specification, estimation and testing. Uppsala, Sweden: Uppsala Univer., 1976. (Research Report 76-9).

KIRSCH, I. L., & GUTHRIE, J. T. Construct validity of functional reading tests. *Journal of Educational Measurement*, 1980, 17, 81-93.

KUHN, T. S. *The structure of scientific revolutions.* (2nd ed.) Chicago: Univer. Of Chicago Press, 1970.

MILLER, G. A. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 1956, 63, 81-87.

PELLEGRINO, J. W.,& GLASER, R. Analyzing aptitudes for learning: inductive reasoning. In R. Glaser (Ed.), *Advances in instructional psychology.* Hillsdale, NJ: Erlbaum, 1982. Pp. 269-345.

STERNBERG, R. J. *Intelligence, information processing and analogical reasoning: the componential analysis of human abilities.* Hillsdale, NJ: Erlbaum, 1977.

STERNBERG, R. J. Factor theories of intelligence are all right---almost. *Educational Researcher*, 1980, 9, 6-13.
THORNDIKE, R. L. & HAGEN, E. P. *Measurement and evaluation in psychology and education.* New York: Wiley, 1965.

THURSTONE, L. L. The stimulus-response fallacy in psychology. *Psychological Review*, 1923, 30, 354-369.

WHITELY, S. E. Measuring aptitude processes with multicomponent latent trait models. *Journal of Educational Measurement,* 1981, 18, 67-84.

WRIGHT, B. D., & STONE, M. H. *Best test design.* Chicago: MESA Press, 1979.

Accepted July 31, 1982.