Gabriel, K. R. Simultaneous test procedures in multivariate analysis of variance. *Biometrika*, 1968, *55*, 489–504.

Gabriel, K. R. A comparison of some methods of simultaneous inference in MANOVA. In P. R. Krishnaiah (Ed.), *Multivariate Analysis II*. New York: Academic Press, 1969.

Harris, R. J. *A primer of multivariate statistics*. New York: Academic Press, 1975.

Hotelling, H. A generalised T test and measure of multivariate dispersion. In J. Neyman (Ed.), *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, 1951.

Hummel, T. J., & Sligo, J. R. Empirical comparison of univariate and multivariate analysis of variance procedures. *Psychological Bulletin*, 1971, *76*, 49–57.

Mosteller, F., & Tukey, J. W. Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (2nd ed., Vol. 2). Reading: Addison-Wesley, 1968.

Olson, C. L. Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, 1974, *69*, 894–908.

Olson, C. L. On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 1976, *83*, 579–586.

Olson, C. L. Practical considerations in choosing a MANOVA test statistic: A rejoinder to Stevens. *Psychological Bulletin*, 1979, *86*, 1350–1352.

Pillai, K. C. S. Some new test criteria in multivariate analysis. *Annals of Mathematical Statistics*, 1955, *26*, 117–121.

Pillai, K. C. S. *Statistical tables for tests of multivariate hypotheses*. Manila: Statistical Center, University of the Philippines, 1960.

Ramsey, P. H. Choosing the most powerful pairwise multiple comparison procedure in multivariate analysis of variance. *Journal of Applied Psychology*, 1980, *65*, 317–326.

Rogan, J. C., & Keselman, H. J. Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal?: An investigation via a coefficient of variation. *American Educational Research Journal*, 1977, *14*, 493–498.

Roy, S. N. On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics*, 1953, *24*, 220–238.

Schatzoff, M. Sensitivity comparisons among tests of the general linear hypothesis. *Journal of the American Statistical Association*, 1966, *61*, 415–435.

Scheffé, H. A method for judging all contrasts in the analysis of variance. *Biometrika*, 1953, *40*, 87–104.

Scheffé, H. *The analysis of variance*. New York: Wiley, 1959.

Stevens, J. Comment on Olson: Choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 1979, *86*, 355–360.

Timm, N. H. *Multivariate analysis with applications in education and psychology*. Monterey: Brooks/Cole, 1975.

Wilks, S. S. Certain generalizations in the analysis of variance. *Biometrika*, 1932, *24*, 471–494.

Received March 3, 1982
Revision received July 27, 1982 ■

Psychological Bulletin
1983, Vol. 93, No. 1, 179–197

0033-2909/83/9301-0179$00.75

Construct Validity: Construct Representation Versus Nomothetic Span

Susan Embretson (Whitely)
University of Kansas

The purpose of this article is to present a new approach to construct validation research—construct modeling. A paradigm shift from functionalism to structuralism in psychology since the original Cronbach and Meehl (1955) formulation of construct validity permits two types of research to be separated. *Construct representation* is concerned with identifying the theoretical mechanisms that underlie item responses, such as information processes, strategies, and knowledge stores. *Nomothetic span* is concerned with the network of relationships of a test score with other variables. These two types of construct validation research address different issues and require different types of data. For each type of construct validation research, appropriate methods and quantitative models are presented to test a priori hypotheses about construct validity. Several research examples are presented, and the construct modeling approach is compared with both the traditional psychometric approach and the information processing approach to establishing theoretical mechanisms in performance.

In Cronbach and Meehl's (1955) formulation of construct validity, a construct is defined as "some postulated attribute of people, assumed to be reflected in test performance" (p. 283). According to Cronbach and Meehl (1955), "*construct validation* is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not operationally defined" (p. 282). The problem faced by the investigator is, "What constructs account for variance in test performance?" (p. 282). In the summary they indicated that "the investigation of a test's construct validity is not essentially different from the general scientific procedures for developing and confirming theories" (p. 300).

However, the nature of psychological theory has changed substantially in the 25 years since Cronbach and Meehl formulated the construct validation concept. The information-processing perspective on psychological theory has led to changes that have been described as a paradigm shift (Segal & Lachman, 1972). The change has been from functionalism to structuralism. That is, the goal of psychological theorizing has changed from explaining antecedent/consequent relationships to explaining performance from the systems and subsystems of underlying processes. As a paradigm shift, the information-processing view entails changes not only in the questions that are asked but also in the type of data that are deemed relevant.

If the construct validation process is equivalent to theory construction, then paradigm changes should influence construct validation research as deeply as other psychological research. That is, the basic issues and appropriate methods for determining the constructs that account for variance in test performance are qualitatively different in the information-processing paradigm. To reflect this different perspective, it is proposed that two separate issues must be addressed by construct validation research: construct representation, and nomothetic span. In the current development, the term *construct* is given a somewhat broader definition than in Cron-

This research was partially supported by National Institute of Education grant number NIE-6-7-0156 to Susan E. Whitely, principal investigator. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred. The author would like to thank Lisa Schneider and Priscilla Hancock for their comments on an earlier draft of this article, and William T. Whitely for his encouragment and support during the development of the ideas presented in the article.

Requests for reprints should be sent to Susan Embretson, Department of Psychology, University of Kansas, Lawrence, Kansas 66045.

bach and Meehl's (1955) usage as "a postulated attribute of people." Here, *construct* refers to a theoretical variable that may or may not be a source of individual differences.

Construct representation is concerned with identifying the theoretical mechanisms that underlie task performance. In short, the goal of construct representation research is task decomposition. In the information-processing paradigm, construct representation refers to the relative dependence of task responses on the processes, strategies, and knowledge stores that are involved in performance. These underlying variables can apply to a wide span of tasks, since the information-processing paradigm has influenced many areas, including personality and social psychology as well as cognitive psychology. In cognitive psychology, for example, task decomposition has been accomplished on tasks that range from simple letter identification (Posner, 1978) to complex text comprehension (Just & Carpenter, 1980). In personality and social psychology, interpersonal inferences and social judgments have been viewed as information-processing tasks (e.g., Nisbett & Ross, 1980).

It should be noted that the constructs that are identified in task decomposition have no necessary implication for individual differences. That is, a process or strategy may be an essential aspect of task performance, but the population that is studied may not vary systematically in their ability to perform it. Alternatively, the particular tasks that are studied are so easy on the element in question that virtually all members of the population can perform without difficulty. Construct representation research is concerned with task variability rather than subject variability.

Nomothetic span refers to the network of relationships of a test to other measures. Nomothetic span is supported jointly by the strength, frequency, and pattern of significant relations with other measures, such as other traits, criterion measures and so forth. It is assessed by individual differences data. The network of correlations between measures arises from many possible communalities that jointly influence development (i.e., genetic and environmental communalities), as well as common construct representation. Nomothetic span, unlike construct representation, indicates the importance of a test as a measure of individual differences.

When Cronbach and Meehl (1955) initially elaborated the construct validation concept, the functionalist paradigm that guided psychological research was not compatible with construct representation as a separate research goal. In experimental psychology the main emphasis was antecedent/consequent relationships, with little interest in the intervening mechanisms that are relevant to construct representation. In correlational psychology, researchers in the factor analytic paradigm were interested in task decomposition, but construct representation was completely confounded with nomothetic span. That is, the "components" that are decomposed by principal factor analysis are based on correlations of individual differences between tasks. Unfortunately, as has been pointed out by many writers, correlations between tasks reflect many influences, including knowledge prerequisites, educational communalities, and genetic communalities as well as common underlying theoretical components. The factors are "functional unities," (e.g., Thurstone, 1935) because they represent influences that cannot be separated in a given set of variables, but this does not necessarily imply that they represent elementary theoretical mechanisms. Factor analysis is unable to separate one or more unique "components" in a task from error variance.

Two additional points about the distinction between construct representation and nomothetic span should be emphasized. First, it is possible to obtain strong support for one but not for the other. For example, in cognitive psychology the construct representation of the Posner (1978) task as a measure of verbal encoding is well supported, but the relatively low correlations of the task with other measures indicate weak nomothetic span (see, for example, the Hunt, Lunneborg & Lewis, 1975, data). In contrast, the nomothetic span of intelligence tests is quite strong, but the construct representation remains unclear (Carroll & Maxwell, 1979). Second, the distinction between construct representation and nomothetic span is not equivalent to other distinctions that have been proposed (see Messick, 1980, for a summary). For example, although nomothetic span is somewhat similar to Campbell's (1960) nomothetic validity in that it concerns correlations with other measures, it is somewhat broader because it also includes correlations with other measures of the same construct or constructs when observed under different conditions. Construct representation, on the other hand, shares some features with Loevinger's (1957) substantive and structural components of validity, but it proposes a more explicit relationship between theory and quantitative models.

Assessing Construct Representation

In this section, some methods for determining the construct representation of test items are presented. The methods assume that the stimulus characteristics of the test items determine the components that are involved in its solution. Four general criteria for evaluating a methodology for assessing construct representation seem appropriate. First, performances of persons on the test must be related to the stimulus characteristics of items. Although item characteristics provide a means of assessing constructs that are involved on the test, the goal of psychological tests clearly is to measure person differences, not item differences, on important constructs. Therefore, the relation of person performance to item characteristics must be clearly specified in the methodology.

Second, the methodology must provide for the comparison of alternative theories of the constructs that are involved in the task. As Cronbach and Meehl (1955) note, assessing construct validity is equivalent to building a theory. Thus, there must be a clear method of comparing alternative theories of task performance in the construct representation phase of construct validation. One effective method for comparing theories is to operationalize the competing theories in quantitative models and then to compare model fits.

Third, the methodology must provide for quantification of the theoretical constructs in specified items. If a test is to be successfully decomposed into more basic theoretical constructs, it is necessary to have parameters for each item on the constructs that are involved in its solution. It is well known in test theory (e.g., Lord & Novick, 1968) that the properties of the test score depend on the item statistics. Modern tests utilize item banking procedures, such that item characteristics are stored with the item. A new test can be built by selecting items with a specified set of characteristics that have direct implications for person measurement. Therefore, if construct representation is to be successfully linked to the concerns of testing, item indices that have similar implications for test scores should be an outcome of the methodology. Not only would such parameters permit item banking but it also would be possible to design a test with specified theoretical properties. Further, item parameters would permit tests to be compared by item difficulty parameters that represent the theoretical properties that are embedded in the item.

Fourth, the methodology should provide for the measurement of persons on the constructs. That is, the methodology should yield person parameters that measure individual differences on the constructs. Ideally, an ability parameter should be available to measure persons on each construct that is involved in the solution of the item.

The primary purpose of this section is to propose an approach to assessing construct representation. However, three different approaches to assessing construct representation will be presented and then evaluated according to these four criteria for an effective methodology. These approaches are (a) mathematical modeling, particularly as implemented in cognitive psychology; (b) psychometric modeling, as exemplified by latent trait modeling; and (c) the multicomponent latent trait modeling, which combines features of the other two approaches. It will be shown that the proposed approach, multicomponent latent trait modeling, meets all four criteria as it combines the opposing strengths of the other two approaches.

Mathematical Modeling of Test Items

Mathematical modeling seeks to predict response patterns from an operationalized set of constructs. In the context of information-

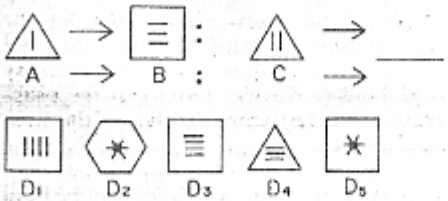

Figure 1. A geometric analogy that is similar to CAT items. (From "Information Structure on Geometric Analogies" by Susan E. Whitely and Lisa M. Schneider, *Applied Psychological Measurement*, 1981, *5*, 383–397. Copyright 1981 by Applied Psychological Measurement, Inc. Reprinted by permission.)

processing research, the constructs identify the processes, strategies, and structures that underlie task performance. Since the theoretical mechanisms are not directly observable, several methods have been developed to identify covert processing variables.

Task decomposition by mathematical modeling has been applied directly to the type of items that appear on popular psychological tests. In cognitive psychology, this line of research is known as cognitive component analysis (Carroll, 1976; Pellegrino & Glaser, 1979; Sternberg, 1977a, 1977b). Task decomposition has also been applied to personality and attitude tests (Cliff, 1977; Cliff, Bradley, & Girard, 1973), although few research studies have appeared in the literature.

The mathematical modeling approach can be applied directly to decomposing test items for two reasons. First, the data of the mathematical modeling approach is similar to the data from a psychological test in an important way. Like examinees who are administered the items of a test, the subjects in a mathematical modeling experiment respond to several independent tasks. In fact, even the task content has been similar. For example, the mental rotation items, geometric analogies, series completions, categorical syllogisms and verbal analogies that have been mathematically modeled in cognitive psychology are found on popular intelligence tests. Second, like item statistics on psychological tests, the mathematical modeling approach provides a quantitative index for each item or task.

The mathematical modeling methods that have been applied to test items thus far can be classified into two general categories: (a) the method of complexity factors and (b) the method of subtask responses. In the method of complexity factors, each item is scored on one or more factors that represent the item's position on the underlying theoretical variables. The variables can be either process or structure variables. For example, Mulholland, Pellegrino, and Glaser (1980) postulated that two processing factors were important in the solution of geometric analogies, such as those presented in Figure 1. These are (a) encoding complexity, which depends on the number of elements in Stimulus A in the analogy, and (b) transformational complexity, which depends on the number of transformations required to convert Stimulus A to Stimulus B. In Figure 1, Stimulus A contains two elements (the triangle and the line) and three transformations (a shape change for external element, an increase in number of internal elements, and a 90° rotation of the internal elements). Each item can be scored on the two complexity factors, encoding and transformations. A simple mathematical model of item difficulty is the following:[1]

$$p_i = \eta_1 q_{i1} + \eta_2 q_{i2} + a, \qquad (1)$$

where p_i is the difficulty of item i, q_{im} is the score of item i on complexity factor m, η_m is the weight of factor m in item difficulty, and a is a constant. Mulholland et al. (1980) found that mathematical models based on two complexity factors gave good prediction of response time and response accuracy on geometric analogy items.

An example of task decomposition on test items using structural variables is Cliff's (1977) study on the Comrey Personality Scale. Cliff's model stated that endorsement intensity on an item, such as "It is easy for me to talk with people," depends on both the internalized schema that the examinee uses to interpret the item and the self-image or persona that he or she wants to present. Cliff (1977) applied multidimensional scaling to

[1] The mathematical models in Mulholland et al. were actually exponential relationships, but for simplicity a simple linear model is presented here.

similarity judgments of the items from each scale on the test to determine the semantic space or schemas that are involved in the interpretation. Then, using these dimensions as predictors, he found substantial support for the following mathematical model of degree of item endorsement:

$$x_{ij} = \sum_m w_{mj} v_{im} + a_j, \qquad (2)$$

where x_{ij} is the endorsement of item i by person j, w_{mj} is the weight that person j assigns to dimension m, v_{im} is the coordinate of item i on dimension m, and a_j is the tendency for person j to endorse items.

In contrast to the method of complexity factors, the method of subtask responses requires the theoretical variables to be identified from responses to a series of subtasks that have been constructed from the items. For example, several studies (Whitely, 1980c, 1981; Parseghian, Goldman, Pellegrino, & Sallis, Note 1) have mathematically modeled item difficulty on verbal analogies. Table 1 presents the full item and two subtasks used by Whitely (1981) to study alternative models for processing verbal analogies. Each subject responded to the full item and then several weeks later to the subtasks. A mathematical model of the components predicts the probability of solving the full item, P_{iT}, from the probabilities of solving the subtasks. A simple multiplicative model for two processing strategies,[2] is the following:

$$P_{iT} = P_{i1} P_{i2}, \qquad (3)$$

where P_{iT} is the probability that the full item is correct, P_{i1} = the probability that the rule construction is correct, and P_{i2} = the probability that the response evaluation is correct. This model proposes a rule-oriented strategy that is represented by two components, rule construction and response evaluation. The probability of correctly executing the rule-oriented strategy is the product of the probabilities for rule construction and response evaluation. Moderately good prediction of item difficulty was obtained from subtask models like Equation 3 by Whitely (1981). Somewhat similar subtasks were attempted by Parseghian et al. (Note 1) and Whitely (1977), with comparable levels of prediction of item difficulty.

Table 1
Subtask Set for Verbal Analogy Components

Total item
Cat:Tiger::Dog:_____
a) Lion b) Wolf c) Bark d) Puppy e) Horse
Rule construction
Cat:Tiger::Dog:_____
Rule____?____
Response evaluation
Cat:Tiger::Dog:_____
a) Lion b) Wolf c) Bark d) Puppy e) Horse
Rule: A large or wild canine

Another example of subtask modeling is Sternberg and Turner's (1981) study of syllogistic reasoning. The following syllogism is an example:

All A are B; All B are C

(a) All A are C, (b) Some A are C, (c) No A are C, (d) Some A are not C, (e) None of the above.

Subtasks were used to separate accuracy of encoding the premises from accuracy in combining the information about the relationship of A to C. An encoding error occurs, for example, when the premise "All A are B" is thought to imply "All B are A." A combination error occurs in linking A to C through information about B. Some other components in Sternberg and Turner (1981) required scoring complexity factors for the premises (e.g., particular versus universal premises, affirmative versus negative premises). A mathematical model for a theory of syllogistic reasoning was identified from the subtask responses and the complexity factors.

The example above shows that the method of complexity factors and the method of subtasks are not mutually exclusive. A study may use a combination of the two methods.

[2] The models in Whitely (1981) contained a subscript for persons as well.

Evaluation. The mathematical modeling approach fails to meet the first of the four criteria because it does not specify any relation of item characteristics to person performance. Mathematical models can be fit to either averaged data (i.e., item means are regressed on predictors) or individual data (i.e., individual item responses are regressed on the predictors). In both cases, however, the target of the mathematical models is item variance, not person variance, and there is no way to anticipate what the relation between these sources of variance may be. That is, mathematical models do not predict the variances and covariances of person parameters.

To give an example, it is entirely possible to have a test of nearly uniform item difficulties, thus yielding little reliable item variance to predict from a mathematical model. However, such a test could possibly have a large person variance, particularly if the uniform item difficulty had a p value of .50. In this case, the mathematical models of the items obviously would not indicate the source of the person variance with respect to any postulated constructs in the items.

For the second criterion, it is clear that the mathematical modeling method provides for testing alternative theories of the constructs. This is a major advantage of the mathematical modeling methods.

The third criterion, the quantification of items, is only partially fulfilled by the mathematical modeling approach. Parameters for the items possibly can be obtained by the product of each complexity factor and the regression weight (i.e., $\eta_m q_{mi}$ in Equation 1). However, such parameters have limited utility in item banking for modern psychometric tests. The parameters are not generalizable over populations because the p values for item difficulty that are modeled depend on the ability level of the population that is studied. In extreme populations, the item p values will have restricted variance as compared with a more representative population. Thus, the complexity factor parameters, η_m, will not be the same as for a representative population. In addition to this limitation, the mathematical model parameters are not useful for selecting items to design a new test. Unlike modern item response theory (i.e., latent trait models), the reliability of the test for persons at various ability levels cannot be anticipated from these item parameters.

The fourth criterion, providing person parameters for individual differences measurement, is not satisfied in the mathematical modeling approach. Although it is possible to apply the models to the item responses of subjects and to obtain weights for each person that could represent individual differences in the constructs, unfortunately, these person parameters are not satisfactory for measuring individual differences. If persons differ in either response variance or predictability of responses from the mathematical model, the regression weights do not have a comparable metric across persons. For binary response data, as is typically obtained from test items, response variance depends directly on the performance level of the person.[3] Ironically, the more persons vary in performance levels, the more inappropriate the modeling parameters will be for individual differences measurement.

Psychometric Modeling: The Encounter of a Person With an Item

Modeling is also important in contemporary psychometric methods. Latent trait models have made a significant impact on testing methods (Lord, 1980; Lord & Novick, 1968). In latent trait models, however, the target of the modeling is the encounter of the individual person with a specified item. As in mathematical modeling, responses are assumed to result from more basic theoretical constructs. However, unlike mathematical modeling, the construct is a "latent trait," which is identified as the major dimension of response consistency (i.e., a common factor among items). Also unlike mathematical modeling, psychometric models contain parameters for both items and persons.

A latent trait model predicts the probability that person j solves item i, $P(x_{ij} = 1)$. In the simple logistic latent trait model (Rasch, 1961), this probability depends on the item's difficulty, b_i, and the person's ability, θ_j, as follows:

$$P(x_{ij} = 1) = \frac{e^{\theta_j - b_i}}{1 + e^{\theta_j - b_i}} \qquad (4)$$

The Rasch model often provides a good fit for typical test data because it predicts the well-known S-shaped ogive for the item characteristics curve. More complex latent trait models are available to reflect item characteristics curves with unequal slopes (i.e., item discriminations) or lower asymptotes (i.e., as arises from guessing on multiple choice items).

An inspection of Equation 4 reveals several important features of latent trait models that can be realized in test data that fit the model. First, the relationship between the response probability and the parameters is nonlinear. The parameters are contained in the exponents. If the response probability for a single item were regressed on ability, θ_j, the resulting item characteristic curve would have the S-shaped ogive that is typical for test data. Second, person ability, θ_j, and item difficulty, b_i, combine additively in the exponent. This implies that ability and item difficulty have been located on a common scale of measurement, namely position on the latent trait. Also implied by additivity, however, is that person ability and item difficulty have equal weight in determining the effective response potential, $P(x_{ij} = 1)$.

To give some examples, suppose that a good fit for the Rasch (1961) latent trait model is obtained for a 60-item verbal analogy test. The 60 parameters for items, b_i, and the parameters for persons, θ_j, represent their location on the latent trait that underlies performance. Most estimation procedures for the Rasch (1961) model anchor the item difficulties to a mean of zero (i.e., $\bar{b} = 0.00$). A person who is at the level of the item set would have a likelihood of .50 of passing an item in the set. If an average item ($b_i = 0.00$) is selected, a person who is at the level of that item has an ability of zero, as shown by the following equation:

$$P(x_{ij} = 1) = \frac{e^{0-0}}{1 + e^{0-0}} = .50 \qquad (5)$$

However, if a difficult item is selected, say b_{ij} equals 2.0, a person who is at the level of the item must have a high ability to compensate for the high difficulty of the item. In this case, θ_j must equal 2.0. That is,

$$P(x_{ij} = 1) = \frac{e^{2-2}}{1 + e^{2-2}} = .50 \qquad (6)$$

Further examples could be constructed to show the effects of easy items on response probabilities or to compare response probabilities between two persons with different abilities.

Latent trait models have been an attractive alternative to classical test theory methods in American testing for several reasons. Although this article cannot present these in detail (see Whitely, 1980a, or Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1978), the major practical advantage of latent trait models is the possibility of obtaining comparable ability estimates from any subset of calibrated items. Unlike classical test theory methods, which do not consider item difficulty in the calculation of test scores, latent trait models yield comparable estimates of ability even from item subsets that differ vastly in difficulty.

The technical advantages of latent trait models are deemed highly desirable by many test developers, so that items that fit the models are sought explicitly in item selection. If the Rasch model does not fit the data, more complex latent trait models may be implemented. If these do not fit, then the data fail to meet the two assumptions for latent trait models, local independence and unidimensionality. Local independence means that a response to a particular item is independent of responses to other items, and unidimensionality means that the test measures a single latent trait.

Unidimensionality among the items is the assumption that is most often violated in typical test data. However, items that are not good indicators of the underlying latent trait can be eliminated. It is often assumed that eliminating poorly fitting items increases the validity of the test because the remaining items provide more reliable information about the latent trait.

Unfortunately, obtaining unidimensionality in the item set does not necessarily imply

[3] For binary data, the mean is p, the proportion passing, and the variance is $p(1 - p)$. Obviously, the variance of binary data depends directly on performance level, p.

that a single psychological construct is being measured. As Lumsden (1957) and others have pointed out, multidimensional tests can appear unidimensional if the multiple factors have proportional contributions to item variance. Thus, scores on the test represent a weighted composite of the several underlying factors. Especially relevant to the current concerns, Reckase (1979) shows that if several small factors exist in the item set, the ability parameters of the Rasch model represent a composite of these influences. Even more significant, eliminating or weighting items according to how much information they provide about the latent trait may substantially alter the nature of what is measured by the test. Reckase (1979) shows that if the item contains one large factor and several small factors, the latent trait score will reflect predominantly the larger factor. Items that reflect the smaller factors will provide little information about this factor and thus may be eliminated. However, the resulting ability scores will no longer measure the smaller factors.

Evaluation. The psychometric modeling approach more than adequately satisfies the first criterion, specifying the relationship between item characteristics and person performance. The models specify exactly how person and item parameters combine to determine the likelihood of a correct response.

The psychometric method only partially fulfills the third and fourth criteria, providing item and person parameters on the constructs. The item parameters are quite adequate for item banking, since they are not population-dependent and can be used to anticipate information about person abilities on any selected item subset (see Lord & Novick, 1968). Similarly, the person parameters that result from latent trait models are typically regarded as an optimal scaling of ability. However, neither the item parameters nor the person parameters clearly represent theoretical constructs. That is, the trait that items and persons are scaled on may be a single construct or the completely confounded influence of several constructs.

Last, the psychometric method does not satisfy the second criterion, providing for comparisons of alternative theories. Data can be tested for unidimensionality by the psychometric models, but this test is useful only for those theories that postulate a single construct. Even within this limited class, a test of unidimensionality can be misleading. That is, a single dimension could be due to the completely confounded influence of several constructs.

Multicomponent Latent Trait Modeling

Multicomponent models combine a mathematical model of theoretical variables with a psychometric model of the encounter of the person with the item. In these models, the theoretical parameters are identified by task decomposition methods, just as in the mathematical modeling approach. Two families of models have been developed with these properties: (a) the linear logistic latent trait model (Fischer, 1973), which has been an influential model in European research but is virtually unknown in American testing, and (b) the multicomponent latent trait model (Whitely, 1980c), which is a relatively new development. The application of these models will be discussed separately, since they are applied with different methods for identifying the theoretical variables.

Linear logistic latent trait model. With the linear logistic latent trait model (LLTM), the theoretical variables are identified from item scores that represent the information structure of the items, as in the *method of complexity factors*. The LLTM also contains person parameters, θ_j, which can be used to measure individual differences. It should be noted, however, that although LLTM postulates multiple processing variables for items, it is a unidimensional model of individual differences. That is, LLTM contains ability parameters for only one dimension. The LLTM requires the following kinds of data: (a) responses to the intact item, when presented under standard test instructions, and (b) scores for each item on some variables that represent the theoretical complexity factors.

Consider the geometric analogy that was presented in Figure 1. As noted above, Mulholland et al. (1980) scored each item on two variables to represent information processing structure. These variables were the number of elements in the A term and the number of transformations required to convert A to B, which were expressed in a mathematical model as in Equation 1. A linear model of item difficulty could be generalized for any number of complexity factors, m, as follows:

$$b_i = \sum_m \eta_m q_{im} + a, \qquad (7)$$

where b_i is item difficulty, q_{im} is the number of operations of type m, η_m is the difficulty of processing operation m, and a is a normalization constant. Note that in Equation 7 item difficulty is represented as b_i, rather than p_i, which was a value that was calculated directly from the data. Equation 7 is the mathematical model in LLTM. It expresses the relationship of the complexity factors to item difficulty. Thus, item difficulty is given as a linear function of the information structure scores, q_{im}.

The linear logistic latent trait model also contains an individual differences model, the Rasch (1961) model, as given in Equation 4.

Combining Equation 7 with Equation 4 gives the complete linear logistic latent trait model as follows:

$$P(x_{ij} = 1) = \frac{e^{(\theta_j - (\sum_m q_{im}\eta_m + a))}}{1 + e^{(\theta_j - (\sum_m q_{im}\eta_m + a))}} \qquad (8)$$

An inspection of Equation 8 shows that in LLTM the relationship between persons and items is clearly specified, just as it is in other latent trait models. The main difference is that item difficulty is expressed in terms of underlying factors of stimulus complexity rather than individual parameters, b_i. The parameters of Equation 8 may be estimated by conditional maximum likelihood (Fischer & Formann, Note 2). An expanded version of this estimation procedure, along with likelihood ratio χ^2 tests is given by Whitely and Nieh (Note 3).

An example of an application of LLTM is Whitely and Schneider's (1981) study on information processing structure models for geometric analogies. Model I in their study was a replication of Mulholland et al.'s (1980) information structure variables, number of elements and number of transformations. More specific types of transformations were specified in two other models. Model II separated figure distortion transformations (e.g., shape change) from displacement transformations (e.g., rotation), whereas Model III included seven distinct types of transformations. It should be noted that all three models are rather simple accounts of processing difficulty because they consider only the first two elements of the analogy.

The goodness of fit of the models were compared by likelihood ratio χ^2 tests. The results showed that Model II fits the data significantly better than Model I, and Model III is significantly better than Model II. However, the amount of improvement by Model III over Model II was relatively trivial so that Model II is the best of the three models.

Model III was compared with the Rasch model to test the ability of these rather simple models of information processing to account for item difficulty. The Rasch model can be regarded as a special case of LLTM in which there is one parameter for each item. With respect to the family of LLTMs, the Rasch model can be considered as a saturated model, since it reproduces the item difficulty data. It can be seen that even the most complex information-processing model that was estimated for the data, Model III, is significantly different from the saturated model. This indicates other types of complexity factors should be identified for geometric analogies. Examples of several European studies with LLTM are given in Fischer (1978).

Parameters may be obtained from LLTM to indicate the theoretical properties of the items with respect to the complexity factors. For example, Whitely and Schneider (1981) obtained the following model of information processing structure on geometric analogies:

$$b_i = .01q_{i1} - .27q_{i2} + .96q_{i3} - .73, \qquad (9)$$

where q_{i1} is the number of elements, q_{i2} is the number of distortion transformations, and q_{i3} is the number of displacement transformations. As noted above, the item in Figure 1 has two elements and three transformations. Two transformations involve distortions (i.e., number and shape) and one transformation is a displacement (i.e., 90° rotation). The following component decomposition would be given:

$$b_i = (.01)(2) + (-.27)(2) + (.96)(1)$$
$$= .02 - .54 + .96 \qquad (10)$$

Thus, given the theoretical properties of an item, a component decomposition of difficulty can be obtained.

However, the overall difficulty of the item can also be indicated. That is, summing the separate component contributions gives

$$b_i = .44. \quad (11)$$

Since LLTM is a latent trait model, like those presented in the section on psychometric models, the b_is that are obtained from any selected subset of items can be used to anticipate the measurement properties of the test for a target population.

Multicomponent latent trait models. Like most information-processing models, multicomponent latent trait models (MLTM) assume that information from several component processing events is required to solve an aptitude test item. For MLTM, the outcomes of these processes are identified by the method of subtasks. Thus, the model requires two kinds of data: (a) responses to the intact item when presented under standard test instructions and (b) responses to a series of subtasks that represent an exhaustive set of information-processing components that are postulated to underlie item performance.

As does LLTM, MLTM combines a mathematical model of item accuracy, with an individual differences model of component process outcomes. The mathematical model specifies how process outcomes relate to the total item performance and the individual differences model specifies how item and person differences on the component process relate to the subtask outcomes.

As an example, consider subtasks that were presented in Table 1. If only one strategy is used to solve the item, a simple mathematical model of the response to the total item, P_{ijT}, could be given:

$$P(x_{ijT} = 1) = P(x_{ij1} = 1)P(x_{ij2} = 1), \quad (12)$$

where x_{ij1} and x_{ij2} are the responses to the rule construction and response evaluation subtasks, respectively. If the model is generalized to accommodate any number of component subtasks, K, then the following general mathematical model can be expressed thus:

$$P(x_{ijT} = 1) = \prod_k P(x_{ijk} = 1), \quad (13)$$

where $P(x_{ijT} = 1)$ is the probability that person j passes total item i, and $P(x_{ijk} = 1)$ is the probability that person j passes component k on item i. That is, the probability that person j passes total item i, $P(x_{ijT} = 1)$, is the product of his or her component response probabilities. Note that unlike in Equation 3, in Equation 13 the variables concern the response of a particular person to a specified item task or subtask. Equation 13 is the mathematical model of MLTM, since it relates the information outcomes on the subtasks to the outcome on the total item.

The MLTM also contains an individual differences model. In the case of MLTM, however, the individual differences model relates person and item parameters on each processing component to the outcome on the subtask. That is, the probability that person j successfully executes subtask k on item i is given as follows:

$$P(x_{ijk} = 1) = \frac{e^{(\theta_{jk} - b_{ik})}}{1 + e^{(\theta_{jk} - b_{ik})}}, \quad (14)$$

where θ_{jk} is the ability of person j on component k and b_{ik} is the difficulty of item j on component k. Combining Equation 13 with Equation 14 gives the full multicomponent latent trait model, as follows:

$$P(x_{ijT} = 1) = \prod_k \frac{e^{(\theta_{jk} - b_{ik})}}{1 + e^{(\theta_{jk} - b_{ik})}} \quad (15)$$

It should be noted that if only one processing component were involved in performance, then Equation 15 would be the Rasch (1961) model. Thus, the Rasch (1961) model is a special case of the MLTM. Maximum likelihood estimation for the model parameters has been derived (Whitely, Note 4) and is available in a computer program, MULTICOMP (Whitely, Note 5). It can be seen that this model expresses individual responses to an intact item as a joint function of person ability and item difficulty on a set of underlying processing components.

As in LLTM, MLTM precisely specifies the relationship between item characteristics and person ability. However, somewhat more detailed descriptions are given with the MLTM parameters. Both the process outcome and total item probabilities may be calculated on a given item for a person with

specified component abilities. Suppose that an item is selected that is difficult on image construction ($b_{i1} = 1.55$) and easy on response evaluation ($b_{i2} = -2.83$). Suppose further that Person 1 has high ability on image construction and low ability on response evaluation (i.e., $\theta_{11} = 3.00$, $\theta_{12} = -3.00$), whereas Person 2 has the opposite pattern ($\theta_{21} = -3.00$, $\theta_{22} = 3.00$). Thus, these two persons have opposite patterns of ability to perform the component operations. Quite different performance predictions will be given for the item because its component difficulties vary substantially. For Person 1,

$$P(x_{i1T} = 1) = \frac{e^{3-1.55}}{1 + e^{3-1.55}} \frac{e^{-3-(-2.83)}}{1 + e^{-3-(-2.83)}} = [.81][.46] = .37 \quad (16)$$

For Person 2, the parameter values give the following predictions:

$$P(x_{i2T} = 1) = [.01][.99] = .01.$$

Thus, drastically different component probabilities are given for these two people, and the total item performance is also quite different. The second person has an extremely low probability of passing the item because the item is difficult on the first component and the person has a low ability on the first component. Thus, the person has a very low probability of passing that component and, hence, the item. The first person has a much higher probability of passing the item, due to higher image construction ability. Thus, very detailed accounts of processing can be given by the multicomponent latent trait model parameters.

Different mathematical models, as shown above, may be specified for MLTM. Each mathematical model specifies the relationship of the processing components to the intact test item. Like other mathematical models, if alternative models represent distinct theoretical conceptualizations of the item-solving process, then comparisons between models will compare the theories. In MLTM, different models may be specified by either changing the relationships among the subtasks or adding subtasks to measure more components.

Like LLTM, the item parameters in MLTM are useful for designating the component contributions to item difficulty and to use in item banking and test development. In MLTM, each item receives a unique parameter on each component. Items can vary greatly in component difficulty patterns. Figure 2 presents a scatterplot of the MLTM model difficulties for 45 verbal classification items that were examined by Whitely (1981). The correlation between component difficulties is low ($r = .16$) so that items with substantial differences in relative component difficulties can be seen in Figure 2.

If MLTM model difficulties have been assessed for an item bank, tests with specified component processing demands can be selected. For example, on Figure 2, selecting items that are easy on response evaluation (e.g., items 6, 11, 32, 34, and 36) will lead to a test that measures subject ability on image construction only. That is, response evaluation is so easy that even a subject of average ability has an extremely high probability of passing the item. Any individual differences in performance on the item would arise from ability on image construction. Similarly, items with very low difficulties on image construction (e.g., items 20 and 24) could be selected to measure the response evaluation ability. Thus, it is possible to develop tests that depend on different processing components from the same set of items.

Person abilities are also given by MLTM. However, unlike LLTM, which scored persons only on one ability, MLTM gives an ability parameter, θ_{jk}, on each component. In MLTM, the measurement of multiple information outcomes for an item permits the estimation of multiple abilities. Thus, the ability of a person on each component is estimated in the model.

Evaluation. All four criteria for assessing construct representation are achieved in the multicomponent modeling approach. As in the mathematical modeling approach, alternative theories of the task may be tested. The multicomponent models contain mathematical models of the task that can be compared by statistical tests for goodness of fit. As in the psychometric modeling approach, the re-

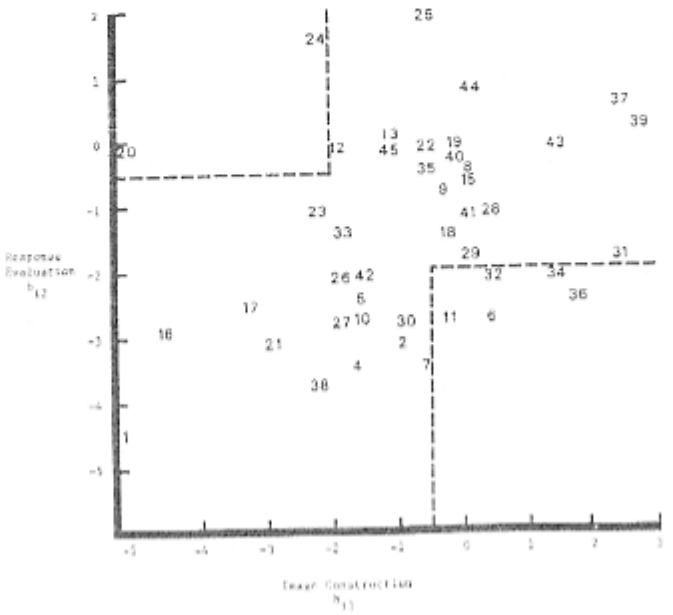

Figure 2. Scatterplot of item difficulties on two components for verbal classification items. (From "Measuring Aptitude Processes with Multicomponent Latent Trait Model" by Susan E. Whitely, *Journal of Educational Measurement*, 1981, *18*, 67–84. Copyright 1981 by National Council on Measurement in Education. Reprinted by permission.)

lationship between persons and items is clearly specified. The multicomponent models contain latent trait models that specify how person ability and item characteristics interact to produce response potential. Also like the psychometric approach, person parameters that are useful for measuring individual differences are an explicit outcome of the methodology.

However, unlike either approach alone, the multicomponent modeling approach gives adequate item parameters to represent the theoretical constructs. The item parameters are useful for item banking and test development because their relationship to person ability is known. But at the same time, the item parameters decompose item difficulty into contributions from underlying theoretical mechanisms.

It should be noted that latent trait models sometimes require large sample sizes to provide stable parameter estimates. For example, the two- or three-parameter logistic models require about 1,000 subjects to yield parameters that are sufficiently stable for item banking. However, if a latent trait model contains only one parameter, such as the Rasch model, reasonable parameter estimates for research purposes may be obtained from about 150 to 200 subjects. Both LLTM and MLTM may be regarded as extensions of the Rasch model. LLTM is actually a restricted version of the Rasch model in which less than one parameter is given per item. That is, the number of complexity factors is less than the number of items.

MLTM can be regarded as multivariate extension of the Rasch model that contains $K + 1$ times the number of items, where K is the number of components. Separate subtasks are used to estimate data for each component as well as the total item. Thus, although each item is associated with $K + 1$ parameters, these parameters are estimated from independent data. This obviously increases testing time substantially to present all the subtasks. However, if effective group methods of subtask presentation are used, such as special paper and pencil booklets (e.g., Embretson, Note 6) or computer-administered tests, the increased testing time becomes more feasible.

Assessing Nomothetic Span

Studies on nomothetic span are typically offered as the primary data to support construct validity for psychological and educational tests. Nomothetic span research is the stage in which the utility of a test for measuring individual difference is assessed. Individual differences on a test are correlated with other tests, group membership, or socially important criteria to support a theory about the construct or constructs that are measured on the test. A nomological network has a wide span if individual differences on the test have frequent and strong correlations with other variables that should correlate with the construct. Further, a specific pattern of correlations is usually required to support the test as a measure of a particular construct. For example, discriminant validity (Campbell & Fiske, 1959) with respect to other traits should be achieved as well as differential validity in predicting behaviors under different conditions. These aspects of the construct validation research are well established and need little elaboration here.

However, when the constructs that underlie item solving are identified in the construct representation phase, as proposed in the preceding section, several additional issues about nomothetic span should be studied. Further, since the constructs in the task are identified, it is necessary to use quantitative models for correlational data that permit a priori construct specification. Structural equation models (Joreskog, 1974, 1978; Bentler, 1980) have the desired characteristics. Although most issues about the nomothetic span of a test could be addressed by structural equation models, the current article will concentrate on the special issues in nomothetic span that arise from the construct representation phase.

Issues about nomothetic span. If multicomponent latent trait models are applied in the construct representation phase, parameters to represent individual differences on the constructs are available. For example, applications of MLTM yield person abilities on each construct. In the case of LLTM, it is possible to form item subsets that have different component difficulty patterns. If such item subsets require different person parameters to achieve goodness of fit, then multiple person measurements can be obtained. For both MLTM and LLTM these will be discussed here as component abilities. However, the following discussion could also be extended to attitude or personality measurements.

Four issues need to be examined about nomothetic span for the component ability measurements. First, the component abilities should represent meaningful dimensions; not only should they yield good prediction of individual differences on the whole test but also more than one component should have a significant prediction weight. If only one component [illegible] weight, little would be added by measuring the multiple component abilities. Components do not necessarily define meaningful dimensions for several reasons: (a) the component abilities simply may not vary systematically over subjects, (b) the component abilities may be too highly intercorrelated, or (c) the component abilities may not be reliably measured in the current item set due to extreme item difficulties. Thus, the pattern of component contributions to the original test score needs to be assessed.

Second, the component abilities should account for the external validity of the aptitude test. Carroll (1976) suggested that cognitive components should account for both the intercorrelations between separate aptitude tests and the correlations of aptitude with learning and achievement. Thus, component abilities should yield a new "structure of intellect," since they predict the correlations between measures.

Third, the component abilities should have differential validity in predicting learning under different instructional treatments or in separate content areas. If the component

abilities do not have differential validity, little practical utility is gained by subdividing aptitude into components. Failure to obtain differential validity implies that the original aptitude test score contains the optimal weighting of components for prediction.

Fourth, the component abilities should show across-task generalizability. If a component ability measures a basic theoretical construct, then that ability should be involved in more than one type of item. That is, if the same components are involved in two or more tasks, then the component abilities should show convergent and discriminant validity across the tasks from which they are measured.

Research examples. Whitely (1980b) tested all four issues about nomothetic span for verbal aptitude test items. Abilities on three components (image construction, response evaluation and event recovery) were estimated by MLTM from two different item types (verbal analogies and verbal classifications). Additionally, Whitely (1980b) obtained American College Test (ACT) scores to measure achievement.

Several structural equation models were compared for goodness of fit to the observed covariances between the variables to examine the four issues about nomothetic span for component abilities. Structural equation models consist of a hypothesized set of relationships among variables that attempt to reproduce their covariances. The covariances may be reproduced by either structural relationships (independent to dependent variable relationships, such as in multiple regression analysis) or measurement relationships (variables that measure the same common factor). Each hypothesized model may be evaluated for goodness of fit. Models that can reproduce the covariances from only a few hypothesized relationships are generally preferred over more complex models.

Figure 3 shows the final structural equation model that fit the verbal analogy data. The observed variables are indicated in the boxes, and the circles represent factors. The predictors are the person component parameters (obtained by MLTM) from the Image Construction, Responses Evaluation, and Event Recovery components from verbal analogies. The dependent variables are the original test scores for verbal analogies and the four area subtests of the ACT: English, Mathematics, Natural Science, and Social Science.

In Figure 3, the utility of measuring separate components is assessed by the significance of the component abilities in predicting the original verbal analogies test score. The weights for each component are located on the arrows from the three components to the dependent variable factors. It can be seen in Figure 3 that Image Construction and Response Evaluation contributed significantly to prediction of the analogy factors.[4] Thus, these data suggest that at least two meaningful component abilities can be separated on the task.

Figure 3 also indicated that the components have some potential to account for the external validity of the verbal analogy test. In a somewhat less complex model than Figure 3, the correlation of verbal analogies with the two ACT factors was modeled by their mutual dependence on the component abilities. No direct relationship between the analogy test and the ACT factors was postulated. That is, the arrows from the components to both the verbal analogy factor and the ACT factors indicate that the intercorrelations among the latter are due only to the common antecedents (namely, the components). This model did not completely reproduce the correlation between the verbal analogy factor and the ACT factors. Significant improvement was obtained by allowing correlated residuals between the verbal analogy test and the ACT factors. (These are shown on the curved lines in Figure 3.) Thus, the components did not completely explain the correlations of the analogy test with other variables.

Last, Figure 3 also provides data that is relevant to the possible differential validity of the component abilities. The prediction weights from the components to the two ACT

[4] The following two types of relationship in Figures 4 and 5 do not contain information about parameter significance: (a) the residual parameters, noted as ϵ or ζ, and (b) factor loadings that were fixed to specify the scale of measurement.

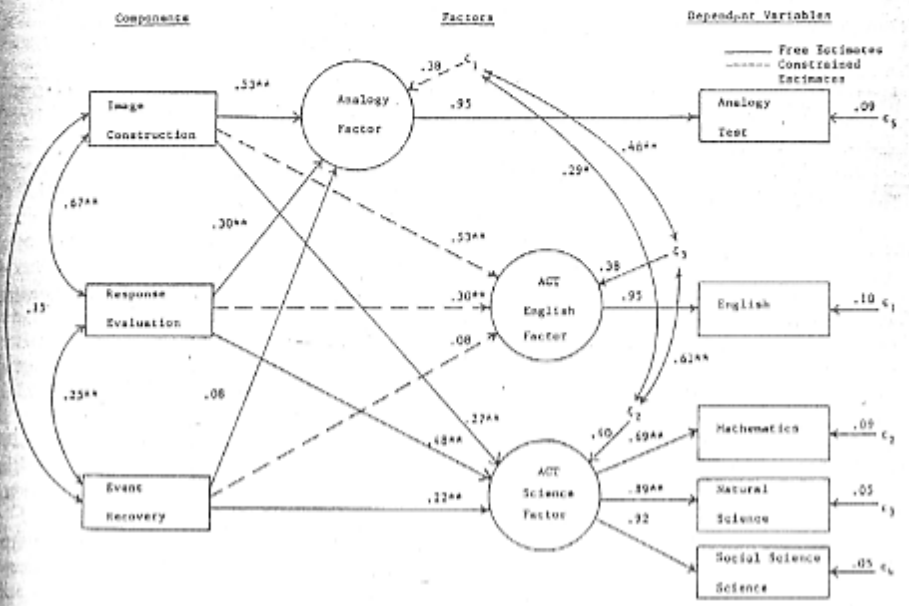

Figure 3. Differential external validity for separated criteria on Model 4: Verbal analogies. (From "Modeling aptitude test validity from cognitive components." by Susan E. Whitely, *Journal of Educational Psychology*, 1980, *6*, 750–769. Copyright 1980 by the American Psychological Association. Reprinted by permission.)

factors, which measure English and Science, show different patterns. A model in which the components had equal weights in predicting each ACT factor was rejected in favor of the differential validity model on Figure 3. It can be seen in Figure 3 that the image construction component is more important than response evaluation in predicting English, whereas the opposite pattern holds for predicting the Science factor.

Figure 4 shows the final structural equation model that Whitely (1980b) obtained to test hypotheses about the across-task generalizability of the three component abilities that were measured from two separate item types, verbal analogies and verbal classifications. The boxes in Figure 4 contain the three component ability measurements from each item type. If the components generalize across tasks, then the correlations among the six measures should be explained by three common factors that represent generalizable components. In Figure 4, the three circles in the center represent the common components factors across tasks. A less complex model than Figure 4 that contained only these common component factors in circles would have supported the across-task generalizability of the components. However, this model did not fit the data. Two additional circles shown in Figure 4 were added to represent specific influences on the components that arise within each item type. Here, this model is equivalent to specifying each component as a unique ability.

An inspection of the correlations between the common component factors shows why fit was not obtained on the model that contained only common components. Two of the common components factors are too highly correlated ($r = .99$) to provide useful information about individual differences. Thus, support for across-task generalizability was not obtained from this data. This suggests that the construct representation of one or both is inadequate.

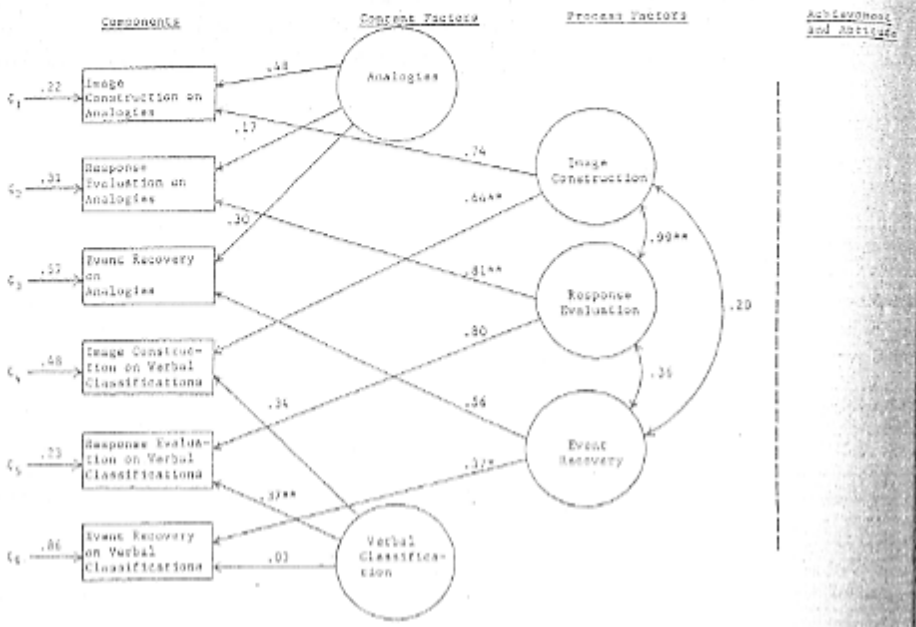

Figure 4. The separation of item content abilities from component processing abilities. (From "Modeling Aptitude Test Validity from Cognitive Components," by Susan E. Whitely, *Journal of Educational Psychology*, 1980, 6, 750–769. Copyright 1980 by the American Psychological Association. Reprinted by permission.)

Construct Validation Research Process

In the preceding sections, a distinction between two aspects of construct validation research was developed. A construct modeling approach, along with corresponding methods and models, was presented for both construct representation and nomothetic span. The purpose of this section is to compare the construct modeling approach with traditional test development procedures and to consider the possible relationship between construct representation research and nomothetic span research.

Perhaps the most salient difference between the construct modeling approach proposed here and traditional test development procedures is the construct representation phase. Typically in test development, construct representation is the responsibility of content or skill specialists who oversee the item writing process. Item specifications are formulated and item writers develop items to fulfill the specifications. However, the item specifications that are given usually do not reflect theoretical properties of the items. Instead these specifications reflect content areas for the items or a loose conceptualization about factors that may be important in item difficulty. In the sense that was discussed in this article, such item specifications can be considered a "theory" of the task. However, the test developer's specifications do not have the rigor of a theory, and further, the specifications are never tested for impact on item difficulty in the test development process.

The current article proposes a qualitatively different approach to construct representation. Test specifications are built from prior theory on the task, and further, the implications of these specifications for item difficulty are studied empirically. Thus, construct representation is a research phase of the construct validation process. The item specifications are truly considered a theory of a task, and the theory is tested by mathematical models that operationalize the constructs. Furthermore, by estimating the quantitative properties of the item with respect to theoretical constructs, the test developer can discard items that do not reflect the desired theoretical properties in the test to be developed.

The nomothetic span phase of construct validation research is familiar to test developers. However, it was noted that nomothetic span has some special aspects that can be studied when preceded by theoretically defined constructs in the construct representation phase. It was pointed out that a construct may be important in item performance, yet that construct may not provide a useful measure of individual differences. Thus, nomothetic span is essential to support a theoretical construct as a meaningful aspect of individual differences.

It is interesting to note that nomothetic span also provides data on the quality of the theoretical constructs that were identified in the construct representation phase. For example, a good theory should contain variables that are general across tasks. Thus, if across-task generalizability is not shown for individual differences measurements on the theoretical variables, the quality of the theory can be questioned. Similar implications for theory can be made from nomothetic span data on the frequency, strength, and pattern of the correlations of test scores with other variables. However, it is important to note that nomothetic span research is regarded as a converging operation to support theory in the construct modeling approach. It can indicate deficiencies in theory (as in the study on across-task generalizability that was presented above), but this is not the primary type of data about the quality of the theoretical constructs that is measured by the test.

Thus, construct representation and nomothetic span are interactive phases of the construct validation research process. Construct operationalization, as favored by Campbell (1960) and Bechtold (1959), is a major focus of the proposed approach to construct validation research. However, the nomological network of Cronbach and Meehl (1955) also has major emphasis in the research process. Nomothetic span research assesses the utility of a construct as a measurement of individual differences and also provides a converging operation to evaluate the theory that is represented by the constructs.

Conclusion

The success of the construct modeling approach, especially for construct representation research, will depend on the ability of researchers and test developers to develop quantitative indices that define the theoretical mechanisms that are involved in the tasks. In some cases, this may require rather innovative methods for studying a particular theoretical mechanism. It should be noted that task decomposition research on test items is quite an active area, so further methodological advances are quite likely.

Developments in the multicomponent latent trait modeling approach are also needed to reflect more complex quantitative models of theoretical mechanisms. Currently the models are able to estimate parameters for the performance of one or more strategies, but as yet no parameters are available to estimate individual differences in the propensity to apply a given strategy. If more than one strategy is important in solving test items, a single strategy model for LLTM or MLTM would fail to fit the data. Multiple strategy models are needed because it is thought that the propensity to apply an effective information-processing strategy to a problem may well be a very important aspect of aptitude. For example, research on expert behavior (Chi, Glaser, & Rees, 1982) has shown that different processing models are required for experienced versus inexperienced individuals.

To conclude this article, it is interesting to consider what it means for construct validation research to be concerned with determining the qualities that are reflected in the test score. This article suggests that the statement "reflected in" has two possible referents. It could refer to a construct that is directly involved in the successful performance of the test item, or it could refer to a construct that is correlated with constructs that are directly involved through some common antecedent development. Thus, a personality trait could be "reflected in" an aptitude test

score because environmental conditions or genetics fostered the development of both the personality trait and the aptitude components. In a more direct sense, the test does not measure the personality trait. Construct representation and nomothetic span were postulated as separate aspects of construct validity to distinguish these two meanings of "reflected in."

Reference Notes

1. Parseghian, P. E., Goldman, S. R., Pellegrino, J. W., & Sallis, R. Parallels between developmental and individual variability in analogical reasoning. Paper presented at the meeting of the *Society for Research in Child Development*, San Francisco, Calif., March, 1979.
2. Fischer, G., & Formann, A. K. An algorithm and a FORTRAN program for estimating the item parameters of the linear logistic test model. (Research Bulletin, No. 11). Vienna, Austria: Psychologisches Institut der Universität Wien, October, 1972.
3. Whitely, S. E., & Nieh, K., *Program LINLOG: An extension of Fischer and Formann's program for the linear logistic test model*. Unpublished manuscript, University of Kansas, 1981.
4. Whitely, S. E. *Maximum likelihood estimation of Multicomponent latent trait models for independent processes* (Tech. Rep. NIE-80-3). University of Kansas, Department of Psychology, Lawrence, Kansas, June 1980.
5. Whitely, S. E. *Program MULTICOMP*. Unpublished manuscript, University of Kansas, 1981.
6. Embretson, S. Contextualization effects in analogical reasoning (Tech. Rep. NIE-82-4). University of Kansas, Department of Psychology, Lawrence, Kansas, 1982.

References

Bechtold, H. Construct validity: A critique. *American Psychologist*, 1959, *14*, 619–629.

Bentler, P. M. Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, 1980, *31*, 419–456.

Campbell, D. T. Recommendations for APA test standards regarding construct, trait or discriminant validity. *American Psychologist*, 1960, *15*, 546–553.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, 1959, *56*, 81–105.

Carroll, J. B. Psychometric tests as cognitive tasks: A new "structure of intellect." In L. B. Resnick (Ed.), *The nature of intelligence*. New York: Erlbaum, 1976.

Carroll, J. B., & Maxwell, S. Individual differences in ability. *Annual Review of Psychology*, 1979, *30*, 603–640.

Cliff, N. Further study of cognitive processing models for inventory response. *Applied Psychological Measurement*, 1977, *1*, 41–49.

Cliff, N., Bradley, P., & Girrard, R. The investigation of cognitive models for inventory response. *Multivariate Behavioral Research*, 1973, *8*, 407–425.

Chi, M. T. H., Glaser, R., & Rees, E. Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 1). Hillsdale, N.J.: Erlbaum, 1982.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 1955, *52*, 281–302.

Fischer, G. H. The linear logistic model as an instrument in educational research. *Acta Psychologica*, 1973, *37*, 359–374.

Fischer, G. Probabilistic test models and their applications. *German Journal of Psychology*, 1978, *2*, 298–319.

Hambleton, R., Swaminathan, H., Cook, L., Eignor, D., & Gifford, J. Developments in latent trait theory: Models, technical issues and applicants. *Review of Educational Research*, 1978, *48*, 467–510.

Hunt, E. B., Lunneborg, C., & Lewis, J. What does it mean to be high verbal? *Cognitive Psychology*, 1975, *7*, 194–227.

Joreskog, K. G. Analyzing psychological data by structural analysis of covariance matrices. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology*. San Francisco: Freeman, 1974.

Joreskog, K. G. Structural analysis of covariance of correlation matrices. *Psychometrika*, 1978, *43*, 443–477.

Just, M. A., & Carpenter, P. A. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 1980, *87*, 329–354.

Lord, F. M. *Applications of item response theory to practical testing problems*. Hillsdale, N.J.: Erlbaum, 1980.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

Loevinger, J. Objective tests as instruments of psychological theory. *Psychological Reports*, 1957, *9*, 635–694.

Lumsden, J. A factorial approach to unidimensionality. *Australian Journal of Psychology*, 1957, *9*, 105–111.

Messick, S. Test validity and the ethics of assessment. *American Psychologist*, 1980, *35*, 1012–1027.

Mischel, W. On the future of personality measurement. *American Psychologist*, 1977, *32*, 246–254.

Mulholland, T., Pellegrino, J. W., & Glaser, R. Components of geometric analogy solution. *Cognitive Psychology*, 1980, *12*, 252–284.

Nisbett, R., & Ross, R. *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, N.J.: Prentice-Hall, 1980.

Pellegrino, J. W., & Glaser, R. Cognitive correlates and components in the analysis of individual differences. *Intelligence*, 1979, *3*, 187–214.

Posner, M. I. *Chronometric explorations of mind*. New York: Wiley, 1978.

Rasch, G. On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley Symposium on mathematical statistics*. Berkeley: University of California Press, 1961, *4*, 321–334.

Reckase, M. D. Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 1979, *4*, 207–230.

Segal, E. M., & Lachman, R. Complex behavior or higher mental process: Is there a paradigm shift? *American Psychologist*, 1972, *27*, 46–55.

Sternberg, R. J. Component processes in analogical reasoning. *Psychological Review*, 1977, *31*, 356–378. (a)

Sternberg, R. J. *Intelligence, information-processing and analogical reasoning: The componential analysis of human abilities*. Hillsdale, N.J.: Erlbaum, 1977. (b)

Sternberg, R. J., & Turner, M. E. Components of syllogistic reasoning. *Acta Psychologica*, 1981, *47*, 245–265.

Thurstone, L. L. *Vectors of the mind*. Chicago: University of Chicago Press, 1935.

Whitely, S. E. Some information-processing components of intelligence test items. *Applied Psychological Measurement*, 1977, *1*, 465–476.

Whitely, S. E. Latent trait models in the study of intelligence. *Intelligence*, 1980, *4*, 97–132. (a)

Whitely, S. E. Modeling aptitude test validity from cognitive components. *Journal of Educational Psychology*, 1980, *72*, 750–769. (b)

Whitely, S. E. Multicomponent latent trait models for ability tests. *Psychometrika*, 1980, *45*, 479–494. (c)

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

Whitely, S. E., & Barnes, G. M. The implications of processing event sequences for theories of analogical reasoning. *Memory & Cognition*, 1979, *1*, 323–331.

Whitely, S. E., & Schneider, L. M. Information structure for geometric analogies: A test theory approach. *Applied Psychological Measurement*, 1981, 383–397.

Received March 8, 1982
Revision received June 24, 1982 ■