TOWARD MORE SENSIBLE ACHIEVEMENT MEASURING: A VIEW AND REVIEW

Kenneth A. Sirotnik

CSE Report No. 217
1983

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

The research reported herein was supported in whole or in part by a grant to the Center for the Study of Evaluation from the National Institute of Education, U.S. Department of Education. However, the opinions and findings expressed here do not necessarily reflect the position or policy of NIE and no official endorsement should be inferred.

Table of Contents

Introduction

Much of what will follow here is a repeat of an unfamiliar--or at least unpopular--theme. The essence of this theme has been either implicit or explicit in writings dating as far back as the early 1930's and continuing up to the present. (See, for example, Walker, 1931; Guttman, 1944; Loevinger, 1947, 1948, 1954; Rasch, 1960; Lumsden, 1961; Bentler, 1971; and Wright & Stone, 1979.) Probabily the most entertaining and insightful review is a rarely quoted article by Lumsden (1976). These authors all propose different techniques (or variants of the same techniques) and analytic models for scaling the items on the ordinary test of achievement. But they all have two basic things in common: (1) they are critical of, and represent alternatives to, classical test theory and (2) they operate from fundamentally the same notion of what it means to measure. The essence of the common theme is, bluntly, that classical (and classical-like) test theories are not very useful when it comes to test construction and analysis.

Why has not the nearly exclusive practice of traditional[1] test theory methods abated during the last fifty years? Why does nearly every new issue of journals like Psychometrika or Educational and Psychological Measurement contain yet another theoretical exposition involving true and error score theory or some esoteric reformulation of the same old reliability coefficient? Were the above authors and others like them just on a flight of fancy proposing crazy ideas that happened to escape the eyes of critical reviewers? No! They merely challenged what to date[2] amounts to over 70

years' worth of archives of scholarly work on test theory models bearing little resemblance to how people ordinarily think about what it really means to measure. To be sure, each challenge did not offer a completely viable alternative to common practice. But it seems to be part of the human condition to hang on tenaciously to the familiar, to the security of a large investment, at least until the market crashes and/or the tide of opinion noticeably changes through the power of advertisement.

Such has been the case recently with the increased use of latent trait models, particularly the model proposed by Rasch (1960) and popularized in the U.S. by Wright (1968, 1969 [with Panchapakeson], 1977, and 1979 [with Stone]). The point of this report is not, however, to advertise any particular measurement model. Rather, I wish to continue advertising the self-evident notion that how one conceptualizes the act of measurement should have a lot to do with how one analyses the quality of the measurement act during its development, implementation and revision phases.

I will restrict this discussion to the measurement of achievement with items of the usual correct-incorrect (1-0) variety. (However, the basic notions are generalizable to ordered response scales more typical in the measurement of values, attitudes, beliefs, opinions, etc.) My point of view regarding how the measurement act is ordinarily conceptualized is not original nor very creative. It rests simply on analogy with measurement in the physical sciences where constructs are often experienced with the senses. The measurement of length, in particular, a person's height, is the usual example and will serve well here. Certainly most

constructs we attempt to measure in the behavioral sciences are not directly experienced and this, or course, constitutes the main source of difficulty. But it does not follow, necessarily, that the generic notions of measurement be any different. Nor does it follow that measurement models be deterministic, i.e., be developed in ideal terms from which deviations are unaccounted for. Probabilistic models are those wherein all deviations from the model have an expected probability of occurence. Both deterministic and probabilistic models exist in both the physical and behavioral sciences.

Implicit in this view of measurement is an assumption that the test items are all measuring the same thing (construct, trait, etc.). Extant psychometric literature is replete with confusion over what exactly is meant by this assumption and the two commonly used terms -- unidimensional and homogeneous -- referencing sometimes similar and sometimes dissimilar empirical interpretations of this assumption. The confusion, not surprisingly, reduces down to different views of the measurement act. Viewed in its original factor analytic sense, unidimensionality refers to one interpretable common factor explaining the item correlation matrix. This fits well with the notion of measurement as repeated single-item tests and the concept of reliability as internal consistency. But internal consistency is only a necessary and not a sufficient condition for a single common factor in an item set; yet, many traditional test theorists (e.g., Gulliksen, 1950; Ghiselli, 1964; Magnusson, 1967; and Allen & Yen, 1979) and practitioners have used both unidimensionality and homogeneity in reference to the internal consistency of a set of items.

To confuse the issue further, Guttman's (1944) "unidimensionality" and Loevinger's (1947) "homogeneity" both, in empirical consequence, refer to the cumulative ordering or scaling of a set of items -- a fundamentally different notion of the use of items to measure a single construct. The analogue of this notion for probabilistic models (e.g., latent class and latent trait models) is the concept of local indpendence, taken by many latent trait theorists (e.g., Lord & Novick, 1968; Hambleton & Cook, 1977; and Lord, 1980) as the equivaltent of the assumption of unidimensionality. (But see the discussion of Traub and Wolfe, 1981, p. 387.)

From my point of view, I assume that there exist sufficiently singular achievement constructs, represented by item sets, that are psychologically interpretable and that are of potential instructional use. A reasonably sucessful application of a measurement strategy is necessary but not sufficient evidence for a reasonably successful effort at measuring a singular construct. In other words, a singular construct is assumed at the outset; a priori verification of the assumption, is, in essence, an exercise in content validity; necessary a posteriori evidence lies, in essence, in the degree of success in developing the measurement device; sufficient evidence, however, is accumulated only through further construction validation studies.

In what follows, a common conceptual view of the act of measurement will be presented and contrasted, in general, with the act as implied by traditional test theories. This discussion will then be punctuated by a more specific overview of several traditional test theories to illustrate

the issue further. Finally, alternative models will be reviewed which are more in line with how the measurement act is ordinarily conceived.[3]

Precision and Accuracy: Disentangling the Concepts Measurement and Dependability[4]

It is important, first, to define measurement more explicity. Many definitions have been proposed resulting in disputes over what does and does not constitute measurement. My interest is not to debate the issue at a philosophical level, but rather to simply clarify how the term will be used here. It will serve my purposes well by following the lead of Torgerson (1958) who reserves the use of the term measurement as follows:

> The logic of measurement deals with the conditions necessary for the construction of a scale or measuring device. Measurement as used here refers to the process by which the yardstick is developed, and not to its use once it has been established, in, say, determining the length of a desk. It is essential that we keep this distinction in mind. The use of the established yardstick in "making a measurement" is a rather simple procedure involving merely the comparison of the quantity to be measured with standard series, or perhaps only reading the pointer or counter of an instrument designed for the purpose. We are here concerned with the more basic problem of establishing a suitable scale of measurement.
>
> ....measurement pertains to properties of objects, and not to the objects themselves. Thus, a stick is not measurable in our use of the term although its length, weight, diameter, and hardness might well be.
>
> Measurement of a property then involves the assignment of numbers to systems to represent that property. In order to represent the property, an isomorphism, i.e., a one-to-one relationship must obtain between certain characteristics of the number system involved and the relations between various quantities (instances) of the property to be measured.
>
> The essence of the procedure is the assignment of numbers in such a way as to reflect this one-to-one correspondence between these characteristics of the numbers and the corresponding relations between the quantities. (pp. 14-15)

Implicit in this usage is the preference not to use the term measurement in the broader sense of Stevens' classic definition: "Measurement is the assignment of numerals to objects or events according to rules" (Stevens, 1951, p. 22). Nominal scales, therefore, are not the result of measurement but of classification. Measurement presupposes, therefore, that the object has a property that exists in magnitudes that can be represented on either ordinal, interval or ratio scales. And again I align myself with Torgerson who finds it uninteresting to worry about what is or is not "permissable," in practice, with measurement scales of these several types:

> ....a major share of the results of the field of mental testing and of the quantitative assessment of personality traits has depended upon measurement by fiat. This is clear, for example, when curves are fitted by the process of least squares or when product-moment correlations, means, or standard deviations are computed. All of these presuppose that distance has meaning. Hence, either explicitly or implicitly, the experimenter is measuring the attribute on an interval scale whose order and distance characteristics have obtained meaning initially through definition alone.
>
> The discovery of stable relationships among variables so measured can be as important as among variables measured in other ways. Indeed, it really makes little difference whether [a] scale of length, for example, had been obtained originally through arbitrary definition through a relation with other established variables, or through a fundamental process. The concept is a good one. It has entered into an immense number of simple relations with other variables. And this is, after all, the major criterion of the value of a concept. (p. 24)

The "act" of measurement, then, refers generally to both the logic of measurement and the process of constructing a test, i.e., a rule or set of procedures operationalizing the construct in a manner consistent with the logic of measurement. What, then, is a test theory? I would prefer that the phrase "test theory" denote the complete act of not only constructing the measuring instrument, but also of assessing further the

validity of that instrument including its dependability[4] under specified conditions of use. In other words a theory of testing, to be complete, must include a measurement model, a dependability model and a validity theory. This last ingredient really includes (and goes beyond) the measurement and dependability models and is what justifies the usage of the term "theory." I know of no past or current "test theory" that deals explicitly with all three aspects. Traditional test theories are theories of dependability (some more restricted than others) with some validity theory. The newer latent trait models are just that, models for measuring a presumed construct. The focus of this paper is clearly on measurement, but by way of contrasting the act of measurement with the dependability of obtained measures.

Now suppose we had before us a small collection of the usual multiple-choice (or true-false, completion, etc.) items of the type commonly found on a test designed to measure a specific achievement outcome. On their face, all such tests "look alike." However, depending upon the conceptual model of measurement underlying the analytical process for selecting these items, this innocent looking collection could be quite different in terms of item composition and empirical characteristics. It is the contention here that classical theory is conspicuously lacking in explicit regard for the potential value of the individual item. By this I mean that there is no explicit recognition of the measurement function served by items. Classical true and error models characterize the consequence of applying a measurement rule--they do not characterize the essence of the rule itself.

Let's consider the "essence of a measurement rule" by continuing the analogy with measuring a person's height. In measuring height, a tape measure and its properties operationalize the rule. Instead of "tape measure," let's use the simpler term "ruler." Suppose we use a ruler (of sufficient length) to measure peoples' heights. Traditional test theories have a lot to say about what to do with the obtained measurement; they have little to say, however, about how the ruler is constructed in order to obtain the measure, i.e., how the ruler is calibrated and how a numerical result eventually becomes associated with each person as a quantitative indicant of the height of the person. In other words, rather than the question of precision with which any given measurement is obtained, traditional test theories take the measurements as given and pursue the question of accuracy, i.e., how consistent the measurement rule is over repeated applications.

Precision and accuracy are cornerstone concepts of any theory of approximate numbers. They reflect fundamentally different ideas in the measurement process. Yet they are used inter-changeably in the behavioral sciences as a synonym for reliability. Two examples out of many are the following quotes:

> The physical scientist generally has expressed the accuracy of his observations in terms of the variation of repeated observations of the same event. The mean of the squared deviations of these observations about the obtained mean is the "error variance." This is a measure of precision or reliability....We regard reliability as the consistency of repeated measurements of the same event by the same process.... (Cronbach, 1947, p. 1.)
>
> Reliability of measurement, then, pertains to the precision with which some trait is measured by means of specified operations....Such indices will be useful

> for comparing different tests so we can ascertain which gives us the most precise or stable scores, and will permit us to ascertain whether the reliability with which a test measures is sufficient for our purposes....Casting reliability in terms of the coefficient of correlation between parallel tests provides another way of describing the precision of measurement (Ghiselli, 1964, pp. 215-218).

In the physical sciences, the concepts of precision and accuracy are clearly distinguished although not always in the same way. In the absence of empirical error, a measurement m precise to the nearest u^{th} unit has an inherent absolute error equal to $\pm u/2$. In this case, accuracy becomes relative error due to imprecision, i.e., $(u/2)/m$. But when empirical error exists--that is, error due to the measurer, the measuree, and/or the measurement circumstances--accuracy (not precision) is usually defined as in the first sentence of Cronbach's (1947) quote above. The dictionary is of little help in sorting out any systematic distinctions. For example, Webster's New World Dictionary (College Edition) gives us this definition: "Precision, the quality of being precise; exactness; accuracy." And in the same dictionary, is this definition: "Accuracy, the quality of being accurate or exact; precision."

At the risk of confusing the issues further, I will elect the versions of these two concepts that serve to keep two fundamental properties of the measurement act separable. Suppose in measuring the height of a person, the ruler is marked off in feet; we can then measure anybody's height to the nearest foot. This is a statement of precision. Included in this notion of precision is the overall length of the ruler. If it is only 5 feet long, the measurement of people over 5 feet tall would necessarily be much

less precise. Precision is intrinsic in the construction of the measuring instrument; it can be increased by conceptualizing and adding more hash marks to the ruler. Half feet can be added to the ruler enabling the measurement of height to be precise to the nearest half foot. It is not really necessary that the hash marks be at equal intervals, or that the addition of hash marks be midpoints of each interval.

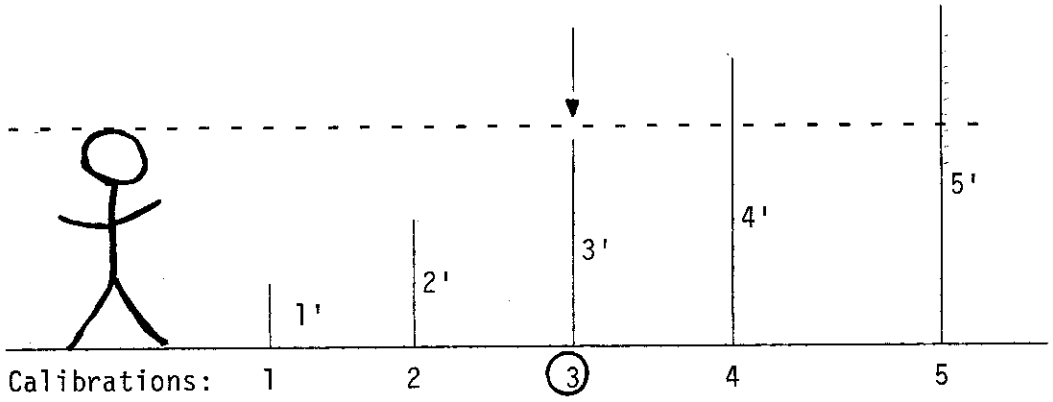
Possibly a better conceptualization of precision is gained by defining it as the number of measurement decisions an instrument can potentially make. The ruler calibrated in half feet can potentially make twice the number of relative height decisions as can the ruler calibrated in feet.

To facilitate the analogy with test items, the ruler can be reconceptualized as a collection of straight sticks consisting of a 1-foot stick, a 2-foot stick, a 3-foot stick, and so on. The more precise ruler is reconceptualized as a set consisting of a 1-foot stick, a 1½-foot stick, a 2-foot stick, a 2½-foot stick, etc. Measurement of height, then, is the process of isolating two adjacent (ordinality being assumed) sticks within which lies the height in question and judging which of these sticks is closest, i.e., to within u/2 units where u is the unit of precision. Alternatively, the measure of a person's height is the number of sticks surpassed by the person's height (plus u/2). If the person is judged to be shorter (by u/2 or more) than the stick, he/she is scored zero; if taller, he/she is scored one. The person's height is then the total score after being tested on the set of sticks. Figure 1 lays out the process schematically. Whether sticks are ordered as calibration marks on a ruler or unordered and used summatively, the result is the same: The person's

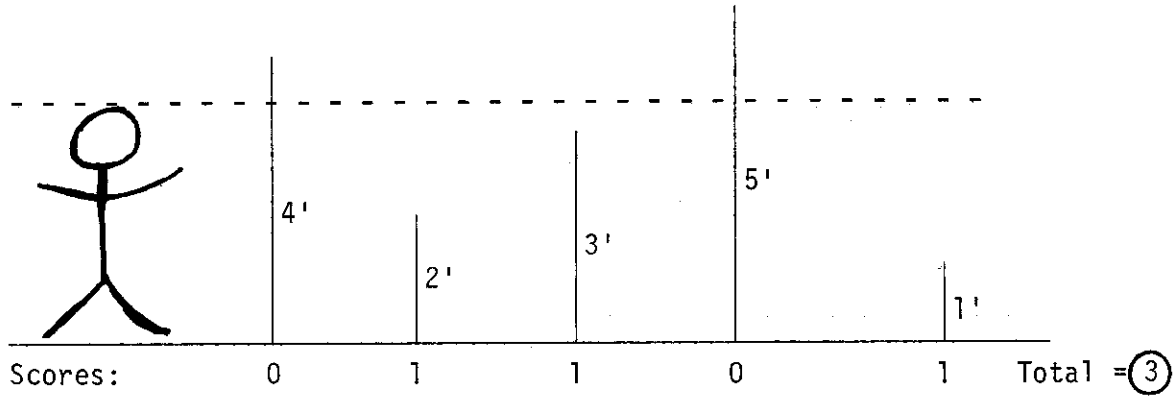
Figure 1

Schematic Representation of the
Act of Measurement
(Height as an Example)

Ordered Sticks:



Unordered Sticks:



height is judged to be 3 feet to the nearest foot. That is, the person's height is somewhere in the theoretical interval of 2½ to 3½ feet. Precision is inherent in the way in which the measuring instrument is calibrated and made operational.

Accuracy is reserved here as a term for describing the degree to which the use of the measuring instrument is error-free. Accuracy is an empirical concept given an already calibrated instrument. Indexing the level of accuracy involves repeated measurements under the circumstances in which accuracy is required. In the above example, to the extent that we can consistently arrive at (or close to) the same measurement of height (to the nearest foot or half-foot depending upon which ruler we use), we have an accurate measuring procedure. The more accurate the procedure the less variability in obtained measurements over repeated measurement trials.

The complete independence of the concepts of precision and accuracy should be clear: A highly precise instrument can be grossly inaccurate (a rubber measuring stick calibrated to the 32nd of an inch) compared to the accuracy of a less precise instrument (a steel measuring stick calibrated in yards). Moreover, accuracy is a function not only of instrument "decay," but also of the circumstances under which it is used. Technically, therefore, we assess the accuracy of the measurement procedure which includes error due to the instrument itself, the person doing the measurement, the person being measured, and the environment in which the measurement process takes place.

Given this distinction, reliability (or, more generally, dependability) as defined by classical (and classical-like) test theory models is clearly

a synonym for the accuracy of a test. Empirically and theoretically, the concepts of reliability and dependability have been concepts of repeated measurements. In this sense, it matters little whether the repeated measurements are replicates (strictly parallel) or samples from a domain (randomly parallel); that is, the generic concept of accuracy remains intact regardless of the conceptual changes in meaning of "true score" implied by the several classical models. So long as we envision only the composite result of the testing process, the classical models are quite analogous to the physical model of measurement. The test score is analogous to the "ruler score," i.e., the obtained height measurement. If we are interested in assessing the accuracy of a single ruler, then we could use the original classical test theory model of strictly parallel repeated measurements. If, instead, we are more interested in the accuracy of a variety of rulers (wood, steel, cloth, etc.) from different manufacturers, then the item sampling models of randomly parallel repeated measurements would be useful. The domain of generalizability changes, but the notion of accuracy does not--empirical estimates obtained through repeated measurements, either with the same ruler (strict parallelism) or with a sample of rulers (random parallelism).

However, the physical model and traditional test theory models part company when it comes to the notion of internal consistency. Inquiry into the internal consistency of a ruler would be directed at the verification of the calibrations vis-à-vis the construct in question and the selected measurement unit standard--an investigation of the precision of measurement. In test theory, the inquiry is directed, as it should be, toward the items. But in traditional theories, the inquiry proceeds by

simply recasting items into the same role as the test, viz., repeated measurements--an investigation of the accuracy of measurement.

Where in the traditional test theory models is the concept of precision? Conceptually speaking, the answer is, "Nowhere." Now of course precision is manifested in the test item, in particular, the difficulty[5] of the test item. A student passing a more difficult test item evidences more ability than does a student who can pass only a less difficult item. The analogy with Figure 1 should be clear. The collection of items is the ruler, conceptualized as an ordered bundle of sticks. The item difficulties are analogous to the lengths of the sticks. Measuring the ability of a student involves locating that pair of adjacent items B and A such that the student correctly answers B (and all other items easier than B) but not A (nor all other items more difficult than A). Traditionally, the student's measure is the ordinal position of item B, or, equivalently, the total number of items answered correctly by the student.

Certainly this analogy is lacking in some non-trivial respects. In particular, the determinacy in the ordering of sticks is hardly (if ever) realized in the ordering of items. If stick C is shorter than stick B, and a student's height surpasses the length of stick B, then it will surely pass that of stick C. Such is the beauty of measuring constructs we can understand with our senses. But if item C is easier than item B, and a student correctly answers item B, then it is not always a sure bet that he/she will correctly answer item C as well.[6] Such is the legacy of the attempt to measure abstract behavioral constructs. Moreover, the procedure for assigning an invariant metric to the measurement of height is straightforward; it is much less so when using items to measure ability.

But I believe these to be minor details compared to the conceptual identity between sticks and items and their role as calibrations on the "ruler." The point to be made here is that this is not the role cast for items by classical (or classical-like) test theories. Lest I may have begun to lose some readers who are rusty on classical (and what I am referring to as classical-like) test theory, I will turn to an overview of several such theories with the expressed intent of further illustrating the argument thus far presented. (Readers already familiar with these models may skip to the Discussion in the next section with little or no loss in continuity.)

Traditional Test Theories

Some would probably argue (and justifiably so) that the sampling of alternative approaches to follow should not be lumped into a single class of test theories, especially one including classical test theory. I do this here only because, in terms of their fundamental conceptualization of the measurement process and important empirical consequences, they are more similar to each other than to the models to be discussed next.

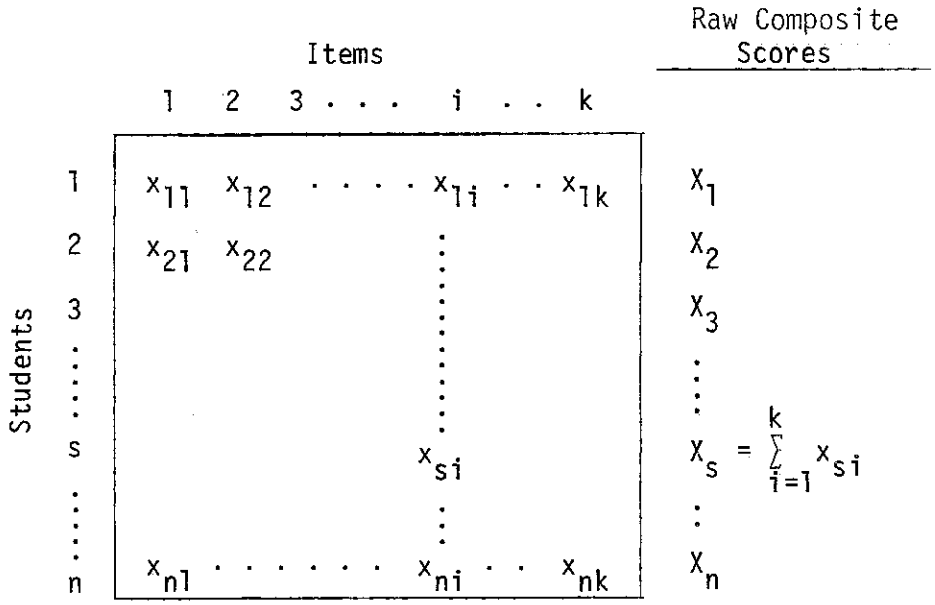
Classical Test Theory

The basic postulate of classical test theory defines a belief regarding the composition of the raw score obtained by a student, namely, that this observed score is simply the student's true score plus what's left over, commonly designated as the error score.

Using some fairly standard notation and the usual matrix layout of the scores of n students on k items, we obtain the schematic in Figure 2.

Figure 2

Student-by-Item Raw Score Matrix and Notation
(x_{si} = 1 or 0 if student s answers item i correctly or incorrectly.)



Item Difficulties $p_1 . . . . p_i = \frac{1}{n} \sum_{s=1} x_{si} . . p_n$

Using T and E for true and error scores, the classical test theory model posits for any student s that:

$$X_s = T_s + E_s \tag{1}$$

A number of relationships obtain from this model when several additional assumptions are made about the true and error score components of repeated measurements on any students.[7] Specifically, these assumptions are (a) errors are totally random and cancel each other out; therefore, the mean error is zero ($\bar{E} = 0$); (b) the correlation between true and error score components is zero ($\rho_{TE} = 0$); and (c) the correlation between errors over repeated measurements is zero ($\rho_{EE'} = 0$).

Assumption (b) leads directly to the variance composition of the linear model above, viz., observed score variability is the sum of variability in true and error scores:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \tag{2}$$

Assumption (c) leads further to the fundamental theorem that the covariance between observed scores on any two repeated measurements is equal to that between the true scores on these measurements:

$$\rho_{XX'}\ \sigma_X \sigma_{X'} = \rho_{TT'}\ \sigma_T\ \sigma_{T'} \tag{3}$$

Finally, if a fourth assumption is added--(d) the repeated measurements are <u>parallel</u> measurements where parallel measurements are defined as having equal true scores ($T = T$) and equal error variances ($\sigma_E^2 = \sigma_E^2$)--then <u>reliability</u> (defined as the correlation between parallel measures, $\rho_{XX'} = \rho_{XX}$) is the equivalent of the ratio of true score to observed score variance:

$$\rho_{XX} = \frac{\sigma_T^2}{\sigma_X^2} \tag{4}$$

But this is also the coefficient of determination in predicting observed scores from true scores (or vice versa), i.e., the correlation between parallel measurements is equivalent to the square of that between observed and true score components:

$$\rho_{XX} = \rho_{XT}^2 \tag{5}$$

A little bit of algebraic manipulation of equations (2) and (4) gives us an equation for the error variance in terms of reliability and observed score variance. In standard deviation terms, this equation is

$$\sigma_E = \sigma_X\sqrt{1-\rho_{XX}} \tag{6}$$

and is commonly referred to as the standard error of measurement. Noting again the relationship in (5), this equation also represents the standard error of estimate in predicting X from T:

$$\sigma_E = \sigma_X\sqrt{1-\rho_{XT}^2} \tag{7}$$

So much for theory. In practice we have only what we observe--raw scores X and the variance of these scores s_X^2 which we use as an estimate of σ_X^2. In view of the above theoretical relationships, if we can also estimate ρ_{XX}, then estimates for the remaining parameters can be automatically computed. The estimate of reliability (denoted r_{XX}) is usually obtained in one or more of three fundamentally different ways with attendant differences in empirical interpretation.

Reliability as Stability. This is the test-retest formulation of reliability as the correlation between two administrations of the same test

over a specified interval of time. If the time interval is too long and allows for true individual changes in the construct being measured, then the test-retest correlation has little to do with reliability. But if the time interval is well-defined in relation to the expected consistency in individual true scores over that period of time, then the test-retest correlation estimates the stability form of test reliability.

Reliability as Equivalence. This is the test-retest formulation of reliability as the correlation between two administrations of parallel tests at the same (or nearly so) point in time. This procedure most closely approximates the classical reliability definition but relies heavily upon the extent of true equivalence between tests. (The same test could, of course, be used twice, but then practice effects might lead to inflated test-retest correlation.) This procedure most closely approximates the empirical assessment of accuracy as discussed in the previous section.

Reliability as Internal Consistency. This is the test-retest paradigm taken to its logical conclusion. For example, split-half reliability is one form of internal consistency equal to the correlation between two random halves of the test when adjusted upwards by the Spearman-Brown (Spearman, 1910 and Brown, 1910) equation to correspond to the full length test. But then we could compute a "split-fourths" coefficient by averaging all possible correlations between four random quarters of the test and adjusting this average accordingly. Eventually, we get down to the item level, treating each item as a parallel replicate "test." The intraclass correlation (average inter-item correlation) stepped-up by a factor of k

(the number of items on the total test) by the Spearman-Brown formula turns out to be equivalent to the mean of all possible split-half coefficients (computed using the Rulon-Guttman formula [Rulon, 1939 and Gutman, 1945]) and was originally derived by Kuder and Richardson (1937) as their formula number 20:

$$KR20 = \frac{k}{k - 1} \left[1 - \frac{\Sigma p_i (1 - p_i)}{s_x^2} \right] \tag{8}$$

Since $p_i(1 - p_i)$ is the variance (s_i^2) of a binary item, this formula is often written more generally as

$$KR20 = \frac{k}{k - 1} \left[1 - \frac{\Sigma s_i^2}{s_x^2} \right] \tag{9}$$

Moreover, since the total variance s_x^2 can be decomposed into an additive sum of all item variances and twice the sum of all possible inter-item covariances, this formula can also be written as

$$KR20 = \frac{\overline{r_{ij} s_i s_j}}{\frac{1}{k} \overline{s_i^2} + \frac{k - 1}{k} \overline{r_{ij} s_i s_j}} \tag{10}$$

$$= \frac{\text{average inter-item covariance}}{\frac{1}{k} \left(\begin{matrix} \text{average} \\ \text{item variance} \end{matrix} \right) + \frac{k - 1}{k} \left(\begin{matrix} \text{average inter-item} \\ \text{covariance} \end{matrix} \right)}$$

From equation (10) it is evident that this estimate of reliability (a) approaches 1 as the number of items increases (so long as additional items are positively correlated with the total test score) and (b) is a measure of the extent to which items are intercorrelated--with each other

or, equivalently, with the total test score. Hence, the use of the term "internal consistency." It becomes clear, then, that this is not only an index of reliability, but also an index (necessary but not sufficient) of the extent to which the set of items comprising the test are measuring the same construct (ability). In the sense of internal consistency, therefore, reliability has a direct bearing upon the construct validity of the test. As noted above, it is for this reason that many traditional test theorists and practitioners have used the terms "homogeneous" and "unidimensional" to refer to this property of a test.

In a nutshell, these are the tenets and consequences of classical test theory. I have ignored a few other important consequences, primarily those having to do with the conceptualization of validity (effects of test length, correction for attenuation, and so forth). For purposes of comparison, however, the concepts so far developed are sufficient to illustrate what I believe to be profound differences between classical test theory and other, perhaps more realistic, measurement models.

Item Sampling Theory

One of the more difficult assumptions to accept (and empirically realize) is that requiring strictly parallel tests (or items). But with a slight shift in perspective, this assumption can be avoided. Consider again the layout in Figure 2. Suppose the k items are a random sample from a conceptually infinite population (universe, domain, pool, bank, etc.) of items over which a student's score would be meaningful. This score would theoretically be the student's true score. Likewise, the n

students can be conceptualized as a random sample from an infinite population of students. And an item's true "score" (difficulty) is the theoretical average score on that item for the population of students.

In essence, what we have is the well-known random effects and analysis of variance design, i.e., an n-by-k, students-by-items, random matrix sample from an infinite students-by-items matrix population. Once again, a linear, additive model is assumed; adopting the convention of using Greek letters for the population parameters, any student's (s) observed score on any item (i) is decomposed as follows:

$$X_{si} = \mu + \tau_s + \pi_i + \varepsilon_{si} \qquad (11)$$

where

μ = the overall mean reflecting the general level of response relative to no response zero;

τ_s = true score for students s;

π_i = true score (difficulty) for item i;

ε_{si} = residual or error effect which could also be regarded as the student-by-item interaction effect ($\tau\pi_{si}$) for a design with one random observation per cell.

With the addition of one more critical assumption--the statistical independence of student-item responses--the components of variance mean square expectations shown in Table 1 can be derived (Cornfield & Tukey, 1956).

Table 1

Components of Variance Mean Square Expectations
For the n x k Random ANOVA Model

Source	df	Mean Square	Expected Mean Square
Students	n - 1	MS_S	$\sigma_\varepsilon^2 + k\sigma_\tau^2$
Items	k - 1	MS_I	$\sigma_\varepsilon^2 + n\sigma_\pi^2$
Error	(n - 1)(k - 1)	MS_E	σ_ε^2

Now an internal consistency form of reliability can be derived without resorting to a definition based upon strict parallelism. Already, in accordance with the model, items can be characterized as <u>randomly</u> "parallel." We can proceed directly by defining reliability (ρ_{xx}) as the proportion of total score variance (σ_x^2) that is the true score variance (σ_τ^2). Since the model implies that

$$\sigma_x^2 = \sigma_\tau^2 + \frac{1}{k}\sigma_\varepsilon^2 , \tag{12}$$

reliability can be expressed as

$$\rho_{xx} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \frac{1}{k}\sigma_\varepsilon^2} \tag{13}$$

Using mean squares as estimates of their corresponding expected values, reliability can be estimated as

$$r_{xx} = \frac{MS_S - MS_E}{MS_S} \tag{14}$$

which, with a bit of algebraic manipulation, can be shown to be identical to equations (8), (9) and (10) above. (This form of KR20 was first derived by Hoyt, 1941.) $\sqrt{MS_E}$, of course, is the corresponding estimated standard error of measurement equivalent to equation (7).

In terms of at least two important applied consequences (and there are more), then, both classical test and item sampling theories lead to the same result. Perhaps they are more similar than one might think. Indeed, with the exception of the strict versus randomly parallel test distinctions, both theories are formally equivalent. It can be shown that the Cornfield and Tukey (1956) assumptions of the random components model imply assumptions (a), (b) and (c) above for the classical test theory model, and vice-versa. (See Lord & Novick, 1968, section 2.7.)

Nonetheless, the ANOVA framework implied by the item sampling model provides a convenient conceptual and analytic rubric that "liberates" (Cronbach et al., 1963) the several classical reliability notions--that is, the sampling model emphasizes the multiplicity of possible reliability coefficients depending upon practical measurement consequences. Cronbach and his associates (Cronbach et al., 1972) have formalized these concepts under the label "generalizability theory." In the simplest design, namely that represented in Figure 2, the "generalizability" coefficient is, of course, given by equation (14), designated previously by Cronbach (1951) as coefficient alpha (α). But other more complicated designs are also relevant and are obtained by adding more factors (facets)--and, therefore, more than one kind of true score parameter each with its corresponding reliability coefficient--to the ANOVA design. Suppose, for example, n classes are observed k times by r raters on o occasions. We can now talk about (and compute) reliability coefficients not only for the main effects due to observations, raters and occasions, but for

the possible interaction effects as well. Using generalized Spearman-Brown procedures, data from one study can then be used to estimate the k, r and o necessary to reach desired reliability levels in a future study. Moreover, some facets might be considered fixed and others, random; and some populations finite, others infinite--all depending upon the practical applications intended.

However, notwithstanding the considerable conceptual and applied benefits accrued through liberating classical test theory of its strict assumption of parallel measurements, both theories conceive the fundamental dynamic of an achievement test identically: Items play roles as replicate measurement rules rather than calibrations on a single measurement rule. Hence, they are first and foremost theories of accuracy--not of precision--as these concepts have been defined above.

Binomial Error Model

An interesting twist on the item sampling model occurs if we restrict our attention to the single student s and conceptualize his/her responses to a random sample of k items as k independent binary events, each with the probablity ζ_s of a correct answer where ζ_s is the hypothetically true proportion correct score for student s in the population of items from whence the sample was drawn. This is the simple "loaded coin-flipping" model, i.e., a binomial model, where the probability for success (say, "heads") is p. Over repeated trials of n coin flips each, the standard deviation of the sampling distribution (i.e., the standard error) of the observed proportions of "heads" is well known to be $\sqrt{p(1-p)/n}$.

Translated to the notation and purpose here, the standard error (of measurement) for student s is the standard deviation of his/her sampling

distribution of observed proportion correct scores ($\overline{X}_s$) on repeated random samples of k as described in the paragraph above. This standard error (denoted σ_{ε_s}) is given, therefore, as

$$\sigma_{\varepsilon_s} = \sqrt{\frac{\zeta_s(1-\zeta_s)}{k}} \tag{15}$$

This standard error of measurement is estimated for each student by correcting (15) for sampling bias and substituting observed scores for true scores:

$$s_{\varepsilon_s} = \sqrt{\frac{\overline{X}_s\ (1-\overline{X}_s)}{k-1}} \tag{16}$$

It should be clear from equation (15) that for item sampled tests of fixed length k, <u>different</u> standard errors of measurement obtain for different true scores. Students obtaining a score of 50 percent will have the largest estimated standard error, i.e., $.5/\sqrt{k-1}$; s_{ε_s} decreases symetrically as scores either go up towards 100 percent or go down towards 0 percent.

This outcome, of course, is completely contrary to the assumption of independence of true and error scores in the classical test theory and item sampling models. In both of these models, the standard error of measurement (equation [7]) is a <u>constant</u> for all students regardless of their observed scores.

We can, however, derive a single standard error of measurement for the binomial model by simply computing the mean of the individual s_{ε_s}. To do this requires generalizing the binomial error model for an individual's score to that for a distribution of scores. (See Lord and Novick,

1968, Chapter 23.) And in so doing, a couple of interesting results emerge. Assuming a linear relationship between true and observed scores, the usual formulation of reliability as the ratio of true score to observed score variance leads to the following estimate for internal consistency:

$$KR21 = \frac{k}{k-1}\left[1 - \frac{\bar{x}(k-\bar{x})}{k\sigma_x^2}\right] \qquad (17)$$

This, of course, is Kuder and Richardson's formula 21 developed originally as an approximation to KR20. Clearly, it is a function only of the observed score mean (or mean item difficulty since $n\bar{p} = \bar{x}$) and observed score variance. KR21 will always be less than KR20 unless there is no variation in item difficulties. When all items are of equal difficulty, they are, of course, equal to their average and formula (17) becomes identical to formula (8).

Analogous comparisons hold for the standard error of measurement. For the binary model, it follows that the estimated correlation between true and observed scores is $\sqrt{KR21}$ and the estimated standard error of measurement is:

$$s'_\varepsilon = s_x\sqrt{1 - (KR21)} \qquad (18)$$

It can be easily shown that s'_ε is the mean of the individual student standard errors of measurement s_{ε_s}. This quantity will always be greater than its analogue in classical and item sampling models (equation [7] with sample estimates) unless, again, item difficulties are equal.

Discussion

Thus, excepting the test construction consequences of strict versus

randomly parallel items, all three "traditional" models appear, for all practical intents and purposes, to be equivalent when item difficulties are equal (or nearly so). This makes a lot of sense when one teases out the subtle differences in the conceptions of true score inherent in each model. In the general binary error model, the true score is a parameter of the item population, but each student receives a different randomly sampled set of items. Ordinarily, a student will have different true scores on each of those item samples, but these are not the true scores of interest. Rather, it is the mean of these true scores (the item population true score) that is to be estimated for each student. A similar conception of true score holds for the item sampling model except that each student responds to the same randomly sampled set of items. The classical model is a degenerative form of the item sampling model where all π_i are equal. But in the event that items are all of equal difficulties, true scores will be identical, in each item sample, and, of course, these are identical to the true score in the population. However, if this is not the case, and students respond to different item samples, more variation can be expected to enter into any summary statistics designed to reflect measurement error.

So where in these "traditional" test theories is the concept of precision as I have defined it? Where do the theories speak to the construction and calibration of the measurement device? Again, the answer is nowhere. I am not, of course, suggesting that items go unrecognized in traditional test theories. However, I am suggesting that the item parameters, for example, in the model specified by (11), are there mostly by

default. Moreover, I'm suggesting that precision, which is indeed gained in the composite test score, is serendipitous--items are invariably non-parallel and tests are usually long enough with sufficient variation in item difficulties so that total scores are at least positively and monotonically related to the underlying ability continuum. Put slightly differently, I am suggesting that the wrong theoretical framework for conceptualizating the act of measurement has been used to evaluate what turns out to be a fairly common and intuitively sensible approach to the measurement of ability.

Consider this ironic outcome in terms of classical test theory: differences in item difficulties (desirable building blocks for measurement) are evidence for violating the fundamental assumption of parallelism for the internal consistency form of reliability. Moreover, such differences automatically put a ceiling on the maximum level of KR20 (or alpha) due to the ceiling on phi coefficients when marginal proportions are not identical. For these reasons, we all learned that the "best" possible test was one with items of near equal difficulty and, preferably, all at the .5 level to maximize the potential for total score variance--all nice ingredients for norm-referenced applications. Not surprisingly, it is under the "ideal" condition of equal item difficulties that all three traditional test theory models are, for practical intents and purposes, identical.

This "ideal" student-item response pattern highlights the folly of treating items as merely short (the shortest) repeated tests. As implied above, maximum KR20 obtain when items are at the .5 difficulty level and all students either get all items right or wrong. For a k-item test, then,

half the students have a score of k and half have a score of 0. Clearly little information is obtained when only two decisions can be made. (Latent trait models, which attack the issue of calibrating test items directly, can not even utilize "perfect" response vectors since they have no utility in pinpointing locations on the latent continuum.) Equally ironic implications of this "ideal" score matrix occur for validity coefficients. (See Loevinger, 1954.) It is a rather sad commentary that "something fishy" about classical test theory was smelled early on by scholars who continued to propagate the methods:

> It may be, if items of graded difficulty levels are used, that counting one point for each item correct is not a proper scoring method. The score assigned should rather be a best estimate of the difficulty level reached, analogous to that used in the Binet test.... Another limitation in the theory here developed should be pointed out. The criterion of maximizing test variance cannot be pushed to extremes. Test variance is a maximum if half of the population makes zero scores, and the other half makes perfect scores. Such a score distribution is not desirable for obvious reasons, yet current test theory provides no rationale for rejecting such a score distribution. Obviously the "best" test score distribution is one which accurately reflects the "true" ability distribution in the group, but there is perhaps little hope of obtaining such a distribution by the current procedure of assigning a score based upon sheer number of correct answers. At present the only solution to such difficulties seems to lie in some type of absolute scaling theory... (Gulliksen, 1945, pp. 90-91.)

As a final example of the ironies inherent in classical models consider the classical test theory notion of a <u>constant</u> standard error of measurement for every possible score. Does it make sense that particular high (or low) scoring students would have the same random error distributions around their true scores as would intermediate scoring students?

At a purely intuitive level this doesn't make much sense at all. The binomial error model makes it clear that errors are smaller at the ends of the score distribution and larger towards the center. This makes perfect sense if we think of sampling items as analogous to sampling balls from an urn to achieve accuracy of estimation--blue balls are items answered correctly, red ones are incorrect items, and a student's estimated true score is the proportion of blue balls obtained when selecting k balls at random from the urn.

But it makes no sense if items are conceived as fundamental building blocks of the measurement process. In this case, "error" ought to become much more associated with the precision of measurement. In fact, the error pattern should be the complete reverse of that predicted by the binomial model. Errors would be larger toward the extremes of the score distribution and smaller towards the center. At the extremes, we know nothing about the ability level of persons scoring 0 or k on a k-item test. The analogy to physical measurement is again instructive. It is equivalent to selecting that bundle of sticks of appropriate length such that they can center on the person's height. If the smallest stick is too long (a 0-scorer) or the longest stick too short (a 1-scorer), we have failed to measure the person's height to within the given units of precision.

In sum, it can be said that classical (and classical-like) test theories are good models for assessing the dependability of measurements whose internal measurement properties are already well understood or at least accepted as given. (Generalizability theory becomes particularly useful in these circumstances as noted previously.) But they are poor

models for directing and assessing the development of item-based measures which, as suggested by the physical measurement analogy, rely upon item difficulties as proxies for calibrations on the "ruler." Again, many achievement tests produce useful results serendipitously for the obvious reason that practitioners of classical testing methods sense the necessity for including items of varying difficulty. But the reasons for the eventual presence or absence of items on their tests are the wrong ones, being rooted in a "theory" of dependability rather than measurement. I will now turn to an illustrative survey of some measurement models which are theoretically oriented in the latter direction.

Cumulative Test Models

For lack of a better one, I am using the term cumulative to refer to a rather heterogeneous class of measurement models which explicitly acknowledge the measurement function of items as heretofore discussed. If not already obvious, the descriptive value of this term will be apparent shortly. A potpourri of these models will be presented in just enough detail to highlight how they radically differ from classical (and classical-like) test theories in their conceptual approach to the measurement act. All these cumulative models approach the measurement act directly (using the items-as-sticks notion) relying on item difficulty variance for precision and calibration and the total score (or a function of the total score) as an indicant of the ability being measured.[8]

Before beginning this survey, I wish to note a side benefit to using the "items-as-sticks" notion in developing a measurement rule (i.e., test).

In 1963, a seminal article by Glaser stimulated the so-called criterion-referenced testing movement. Soon thereafter, an important article by Popham and Husek (1969) rightly noted the inappropriateness of norm-oriented classical test theory methods for handling the development and analysis of criterion-referenced tests. The literature virtually exploded with attempts to adapt classical test theory to fit the requirements of criterion-referenced tests. The focus of these efforts was quite misdirected. The fundamental issue was not testing or even purpose of testing; rather, it was an issue of measurement. The proper role of items in a test forces (or should force) the test constructor to match item content with the cognitive processes to be assessed. Assuming a singular construct and a scalable set of k items having different difficulties $k + 1$ "mastery" levels can be assessed. "Criterion-referenced testing," therefore, is simply sensible measurement.[9] Of course, following sensible measurement, one can always (a) select a particular mastery level for criterion-referenced decisions or (b) compile group statistics for comparative purposes, thereby developing norm-referenced test interpretations.

Guttman's Scalogram Analysis

David Walker (1931, 1936, 1940), perhaps the first person to recognize the value of the doubly ordered raw score matrix, began a series of investigations on the relationship between response patterns and the resultant shape of score distributions. In the course of this inquiry, Walker conceptualized the ideal response pattern and attempted to index departures from this pattern, a condition he nicknamed "hig" after the term "higgledy-piggledy" to describe the apparent haphazardness in non-ideal response

patterns. But his interest centered on implications for test score scatter rather than the more profound implications for measurement itself.

Guttman (1944) reversed this focus and formalized a scaling procedure for assessing the degree to which items conformed to the ideal response pattern. Figure 3a presents an example of an ideal cumulative response pattern for 20 students responding to five items. However, that this is an ideal pattern is not immediately obvious until the score matrix is arranged in rank order on both student scores and item difficulties. One such convient "double sorting" of the score matrix orders students from highest to lowest scores and items from easiest to most difficult. In Figure 3b we see the cumulative nature of the scoring pattern inherent in the unsorted data as presented in Figure 3a. Figure 4 presents the same score distribution, but this time there are some "errors," i.e., student-item responses which do not fit the ideal pattern. For example, student 8 should have answered item 1 correctly and item 5 incorrectly, thereby contributing two student-item response errors to the total 20x5 (i.e., nk) possible student-item responses. Finally, Figure 5 depicts yet again the same score distribution but with many errors resulting in a very poor cumulative pattern.

To index the degree of cumulativeness present in the pattern, Guttman used a deterministic approach. All deviations (e) from the ideal pattern are errors, i.e., the approach makes no allowance for probable deviations. An obvious index then is the proportion of non-errors in the entire response matrix (1-e/nk). Guttman named this index the coefficient of <u>reproducibility</u> (REP) insofar as it reflected the extent to which the response pattern

Figure 3a

Unsorted Cumulative Response Pattern
for a Hypothetical Ideal Score Matrix

STUDENTS	ITEMS 3	2	5	1	4	X
12	0	1	0	1	0	2
19	0	0	0	0	0	0
16	0	0	0	1	0	1
11	0	1	0	1	0	2
5	1	1	0	1	1	4
15	0	1	0	1	0	2
2	1	1	1	1	1	5
13	0	1	0	1	0	2
3	1	1	0	1	1	4
9	1	1	0	1	0	3
1	1	1	1	1	1	5
6	1	1	0	1	0	3
20	0	0	0	0	0	0
14	0	1	0	1	0	2
10	1	1	0	1	0	3
17	0	0	0	1	0	1
4	1	1	0	1	1	4
8	1	1	0	1	0	3
18	0	0	0	1	0	1
7	1	1	0	1	0	3
	10	15	2	18	5	
p_i =	.50	.75	.10	.90	.25	

Figure 3b

Sorted Cumulative Response Pattern
for a Hypothethical Ideal Score Matrix
(Rep = 1.00; CS = 1.00; α = .76)

STUDENTS	ITEMS 1	2	3	4	5	X
1	1	1	1	1	1	5
2	1	1	1	1	1	5
3	1	1	1	1	0	4
4	1	1	1	1	0	4
5	1	1	1	1	0	4
6	1	1	1	0	0	3
7	1	1	1	0	0	3
8	1	1	1	0	0	3
9	1	1	1	0	0	3
10	1	1	1	0	0	3
11	1	1	0	0	0	2
12	1	1	0	0	0	2
13	1	1	0	0	0	2
14	1	1	0	0	0	2
15	1	1	0	0	0	2
16	1	0	0	0	0	1
17	1	0	0	0	0	1
18	1	0	0	0	0	1
19	0	0	0	0	0	0
20	0	0	0	0	0	0
	18	15	10	5	2	
p_i =	.90	.75	.50	.25	.10	

Figure 4

Moderately Cumulative Response Pattern
(Rep = .86; CS = .63; α = .57)

STUDENTS	ITEMS 1	2	3	4	5	X
1	1	1	1	1	1	5
2	1	1	1	1	1	5
3	1	1	1	1	0	4
4	0	1	1	1	1	4
5	1	0	1	1	1	4
6	1	1	1	0	0	3
7	0	1	1	1	0	3
8	0	1	1	0	1	3
9	1	1	1	0	0	3
10	1	0	1	1	0	3
11	1	0	0	1	0	2
12	1	1	0	0	0	2
13	1	1	0	0	0	2
14	1	0	0	1	0	2
15	1	1	0	0	0	2
16	1	0	0	0	0	1
17	1	0	0	0	0	1
18	1	0	0	0	0	1
19	0	0	0	0	0	0
20	0	0	0	0	0	0
	15	11	10	9	5	
p_i =	.75	.55	.50	.45	.25	

Figure 5

Poor Cumulative Response Pattern
(Rep = .74; CS = .46; α = .49)

STUDENTS	ITEMS 1	2	3	4	5	X
1	1	1	1	1	1	5
2	1	1	1	1	1	5
3	1	1	1	1	0	4
4	1	1	0	1	1	4
5	0	1	1	1	1	4
6	1	0	1	1	0	3
7	0	0	1	1	1	3
8	1	0	1	0	1	3
9	1	1	1	0	0	3
10	0	1	1	1	0	3
11	1	0	0	1	0	2
12	1	0	0	0	1	2
13	0	1	0	0	1	2
14	0	1	0	1	0	2
15	1	1	0	0	0	2
16	0	0	0	0	1	1
17	0	0	1	0	0	1
18	1	0	0	0	0	1
19	0	0	0	0	0	0
20	0	0	0	0	0	0
	11	10	10	10	9	
p_i =	.55	.50	.50	.50	.45	

could be perfectly reproduced from the student scores or item difficulties. Thus,

$$REP = 1 - \frac{e}{nk} \tag{19}$$

But REP can never be smaller than the average of the observed item difficulties (p_i) or easinesses ($q_i = 1 - p_i$), whichever are greatest. That is:

$$Min(REP) = \frac{\Sigma Max(p_i, q_i)}{k} \tag{20}$$

The degree of improvement (IMP) over minimum reproducibility is, therefore,

$$IMP = REP - Min(REP) \tag{21}$$

Moreover, the maximum possible improvement is

$$Max(IMP) = 1 - Min(REP) \tag{22}$$

Thus, a more realistic appraisal of the degree to which items scale, above that expected by the marginal results alone, can be seen in the ratio of IMP to Max(IMP). Denoted the coefficient of scalability (CS) by Menzel (1953), this index can be written as follows:

$$CS = \frac{REP - Min(REP)}{1 - Min(REP)} \tag{23}$$

It has usually been recommended that reasonable scalability requires REP $\geq .9$ and $CS \geq .6$. The score matrices in Figures 3a, 4 and 5 depict what are ideally, moderately and weakly cumulative response patterns. These descriptors are clearly reflected in the values of REP and CS accompanying each score matrix.

There are probably three basic reasons why Guttman scaling received little favor in the achievement testing arena. First, for reasonably homo-

geneous objective domains, it is difficult to write achievement items which scale well. In fact, Guttman devised the scalogram procedure for attitude measurement, where it is often easier to write items with distinctly different affective magnitudes (item "difficulties") covering the same essential domain. Second, Guttman made unrealistic claims regarding the power of scalogram analysis to test unidimensionality, thereby opening up the procedure to a barrage of criticism. (See, for example, Festinger, 1947 and Loevenger, 1948.) In line with the discussion of unidimensionality earlier in this report, Guttman would have treaded firmer ground were he to have simply suggested that a scalable set of items is necessary but not sufficient evidence that a set of items measures the same thing to within reasonable evidence of content (and/or construct) validity. Third, and probably most critical, the model was deterministic and offered no statistical (i.e., probabilistic) tests of fit. (See Torgerson, 1958.)

But no criticism was ever directed at the most important notion behind Guttman's approach, namely, the measurement role of items as, in essence, calibrations on a "yardstick." The approximation to the ideal pattern (Figure 3b) would most likely be the acknowledged goal of most achievement test constructors. Yet, instead of expending considerable effort in mapping the cognitive consequences of instructional units and writing, testing, modifying and rewriting relevant items that do begin to show nice cumulative properties, test constructors have been content to build tests on the classical test theory principle of redundancy, i.e., repeated measurements to realize reliability (as internal consistency).

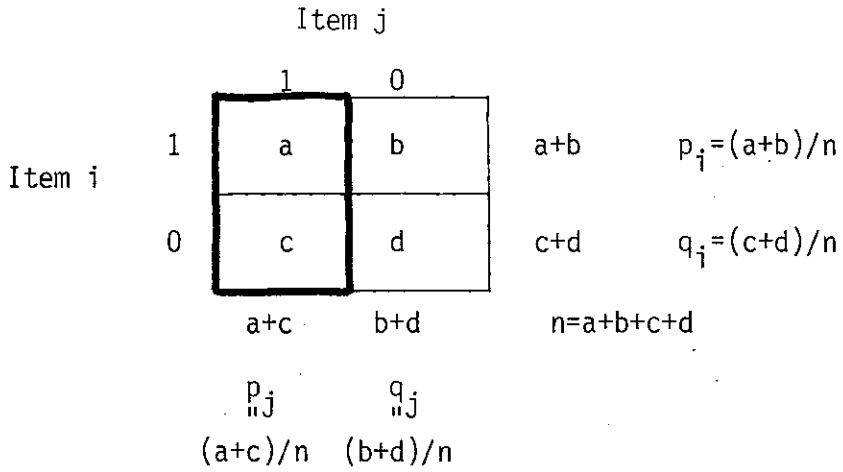
As an interesting aside note, even the deterministic nature of Guttman scaling was rendered a non-issue by a number of writers. Perhaps

the most ingenious approach was based upon Cox's (1954) analysis of covariance model for cumulative repeated measurements (see Maxwell, 1959 and Ten Houten, 1969). Other techniques were investigated by Goodman (1959), Sagi (1959) and Schuessler (1961). The point of this note is simply that attention needs to be redirected towards the underlying principles of measurement and away from the worry of more or less sensitive statistical indicators--not that the latter are unimportant, but that the former are much more so.

Loevinger's Homogeneity Analysis

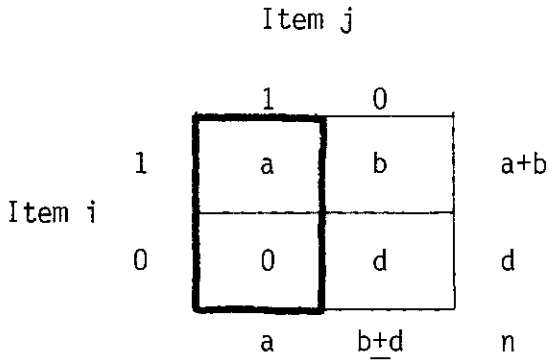
In her 1947 monograph, Jane Loevinger delivered what I believe to be among the best and most provocative critiques of classical test theory; and she followed up with an equally provocative critique of item sampling theory in 1965. To be sure, some of Loevinger's criticisms were a bit overstated, particularly her judgment that the axioms of classical test theory were circular (see Novick, 1966). But generally, her view regarding the inappropriateness of treating items as repeated measurements and her switch in focus from reliability to constructing cumulative scales represents the fundamental contribution.

Like Guttman, Loevinger's approach is based upon deviations from the ideal response pattern. Unlike REP (and its derivatives), however, her homogeneity index (H) reflects these discrepancies in terms of maximum expectations given the difficulty level of the items. Assuming items are arranged in ascending order of difficulty, then for any two items i and j the usual four-fold classification table obtains:



a, b, c, and d are the number of students in each of the respective possible score patterns. Since we have arranged the data assuming item i is easier than j, a+b must be greater than a+c; in proportion terms, $p_i > p_j$.

Ideally, no one ansering the more difficult item correctly would answer the easier item incorrectly. The ideal four-fold classification table would then look like this:



But in the actual testing process, "errors" do occur and c, the number of students getting the more difficult item right but the easier item wrong, is often not zero. These are the deviations from the ideal scale types in Figure 4 and 5.

Loevinger's index of "homogeneity" focuses just on the outcomes a and c, that is on the easier item's scoring pattern for those students answering the more difficult item correctly (heavily outlined column in above schematics.) In other words, the index is based upon the conditional probability $p_{i|j}$ of answering item i correctly given that item j is answered correctly. In the general case, this probability is given by the number of students a who answered both items correctly divided by the total number of students a+c who answered item j correctly:

$$p_{i|j} = \frac{a}{a+c} = \frac{p_{ij}}{p_j} \tag{24}$$

where p_{ij} is simply the proportional equivalent of a, viz., a/n, which is the probability of answering both items i and j correctly. In the ideal case, perfectly homogeneous items (like in Figure 3b), c=0 and $p_{i|j}=1$. In the perfectly heterogeneous case, we would expect items to function completely independently, i.e., $p_{ij}=p_i p_j$, in which case $p_{i|j}=p_i$ by (24) above. An index of homogeneity between the two items i and j can then be formed as follows:

$$H_{ij} = \frac{\text{observed improvement in } p_{i|j} \text{ over that expected under perfect heterogeneity}}{\text{maximum possible such improvement if items were perfectly homogeneous}}$$

$$= \frac{p_{i|j} - p_i}{1 - p_i} \tag{25}$$

In form and intent, this coefficient is analogous to the coefficient of scalability (23) proposed for Guttman scaling. But H_{ij} has a number of further properties. Among the more interesting is the following:

$$H_{ij} = \frac{\phi_{ij}}{Max(\phi_{ij})} \tag{26}$$

where ϕ_{ij} is the ordinary Pearson product-moment correlation between two items which, since the items are binary, is also the fourfold point correlation computed as:

$$\phi_{ij} = \frac{p_{ij} - p_i p_j}{p_i q_i p_j q_j} \tag{27}$$

But ϕ_{ij} cannot reach unity unless the marginals p_i and p_j are equal, i.e., unless the item difficulties are equal. This is exactly the circumstance under which the two items are useless for purposes of precision, i.e., they replicate the same calibration information rather than add decision points to the scale. And of course this is exactly the condition most suited for classical test theory, a theory of accuracy.

However, we can "correct" ϕ_{ij} by dividing it by the maximum possible value it can assume in the case of <u>unequal</u> p_i and p_j. That is

$$Max(\phi_{ij}) = \frac{p_j - p_i p_j}{p_i q_i p_j q_j} \tag{28}$$

and thus

$$\frac{\phi_{ij}}{Max(\phi_{ij})} = \frac{p_{ij} - p_i p_j}{p_j - p_i p_j} \tag{29}$$

Upon dividing both numerator and denominator of (29) by p_j, the equivalency given by (26) is verified.

But the result is more than algebraic. The maximum ϕ_{ij} is obtained when all the students answering item j correctly also answer item i correctly, i.e., when $p_{ij}=p_j$. This, of course, is the ideal cumulative response pattern shown in the above schematic. Thus, $\phi_{ij}/Max(\phi_{ij})$ is really

measuring the extent to which this ideal is obtained and ranges from 0 to 1 accordingly. Unfortunately, this index suffers a bit from the fact that it can also be 1 in value for items of equal difficulties when the b cell is also zero. Even in the extreme case of Figure 6, the overall index (H_t) of homogeneity (see below) is unity. Guttman indices suffer from the same problem. In effect, the scaling indices being presented here are necessary but not sufficient indicators of the cumulative nature of the test items. (See footnote 8.) We must also, therefore, have some indication of item difficulty spread over the ability range of interest.

To complete the discussion of Loevinger's approach, we note that a weighted average of H_{ij} can be formed for all item pairs i and j (such that $p_i > p_j$) yielding an overall index of test homogeneity (H_t). The most straightforward approach to constructing H_t is to reconsider equation (29) which was formed as a ratio of equations (27) and (28). Since the item variances in the denominators of (27) and (28) cancelled out, (29) is, in effect, the ratio of the observed <u>covariance</u> of items i and j to the maximum possible covariance given the p_i and p_j. An overall index can then be formed as a ratio of the sum of the k(k-1)/2 unique observed covariances to the sum of the corresponding k(k-1)/2 maximum covariances:

$$H_t = \frac{\sum_{i \neq j} \sum_{j} (p_{ij} - p_i p_j)}{\sum_{i \neq j} \sum_{j} (p_j - p_i p_j)} \qquad (30)$$

$$= \frac{\overline{Cov_{ij}}}{\overline{Max(Cov_{ij})}}$$

Figure 6

A Degenerate Case:
The Perfect Classical Test Response Pattern
(Rep = 1; CS = 1; α = 1)

STUDENTS	ITEMS 1	2	3	4	5	X
1	1	1	1	1	1	5
2	1	1	1	1	1	5
3	1	1	1	1	1	5
4	1	1	1	1	1	5
5	1	1	1	1	1	5
6	1	1	1	1	1	5
7	1	1	1	1	1	5
8	1	1	1	1	1	5
9	1	1	1	1	1	5
10	1	1	1	1	1	5
11	0	0	0	0	0	0
12	0	0	0	0	0	0
13	0	0	0	0	0	0
14	0	0	0	0	0	0
15	0	0	0	0	0	0
16	0	0	0	0	0	0
17	0	0	0	0	0	0
18	0	0	0	0	0	0
19	0	0	0	0	0	0
20	0	0	0	0	0	0
	10	10	10	10	10	
p_i =	.50	.50	.50	.50	.50	

where Cov_{ij} denotes the covariance between items i and j. Some algebraic manipulation of (30) will verify that it can also be written as

$$H_t = \frac{\sum\limits_{i \neq j} \sum p_j q_i \, H_{ij}}{\sum\limits_{i \neq j} \sum p_j q_i} \tag{31}$$

i.e., H_t is a weighted (by $p_j q_i$) average of $H_{ij} = \phi_{ij}/\text{Max}(\phi_{ij})$. This makes intuitive sense since $p_j q_i$ is the expected proportion of errors in the completely heterogeneous (non-cumulative) case.

It should be clear that H_t is an average inter-item statistic assessing the degree to which all possible ordered item pairs are homogeneous (in the cumulative sense) on the average. Thus, it does not increase merely as a function of increased number of items as does the internal consistency coefficient α in traditional test theory. This is as it should be since H_t is intended to index the cumulative structure of items while α is aimed at assessing the reliability of repeated item measurements.

Ironically, Horst (1953), capitalizing on the seductively simple relationship between H_t and the intraclass reliability coefficient of classical test theory, has proposed "blowing up" H_t by a factor of k using the Spearman-Brown prophecy formula to correct the ceiling effect problem of unequal item difficulties in classical test theory. To his credit, Horst is among the few test theorists who has recognized conceptual differences between reliability and homogeneity and devoted ample space to Loevinger's work in his book on measurement theory (Horst, 1966). But although I can relate to the intended use of the modification offered by Horst, the modification once again confuses fundamental measurement issues by commingling

the concepts of precision and accuracy.

Consider, first, the specifics of the modification. The intraclass reliability (r_{ii}) in classical test theory is the reliability of the average single-item test. It can be shown that by adjusting r_{ii} upwards by a factor of k using the classical Spearman-Brown formula, we end up with the KR20 (or α) formula for reliability at the total test level. Noting that r_{ii} can be defined as the ratio of the average inter-item covariance to the average item variance, i.e.,

$$r_{ii} = \frac{\overline{r_{ij}s_i s_j}}{\overline{s_i^2}} = \frac{\overline{Cov_{ij}}}{\overline{Var_i}} \tag{32}$$

the relationship given in equation (10) leads directly to the Spearman-Brown "correction" as follows:

$$KR20 = \frac{k\, r_{ii}}{1 + (k-1)r_{ii}} \tag{33}$$

Now the maximum possible r_{ii} given the disparities in item difficulties is

$$Max(r_{ii}) = \frac{\overline{Max(Cov_{ij})}}{\overline{Var_i}} \tag{34}$$

If we correct r_{ii} in the usual manner, it is obvious that

$$\frac{r_{ii}}{Max(r_{ii})} = \frac{\overline{Cov_{ij}}}{\overline{Max(Cov_{ij})}} = H_t \tag{35}$$

The suggested modification by Horst, therefore, is to substitute the corrected r_{ii}, i.e., H_t, in equation (33), thereby making it possible for KR20 to reach unity even when item difficulties are unequal.

$$\text{Corrected} \quad KR20 = \frac{k\ H_t}{1 + (k-1)H_t} \tag{36}$$

Consider, second, the implication of this formula. A test can be perfectly homogeneous by adding an infinite number of mostly heterogeneous items so long as they are positively correlated. Now this seems reasonable for achieving increasingly accurate measurements; but it does not necessarily lead to increased precision and a more scalable set of items. Suppose, for example, the test is doubled in length by adding k parallel items, i.e., items that are equal in difficulty, one-for-one, to those in the original test and that scale identically to those in the original test. We now have twice the test information at each ability level but still the same number of ability levels represented in the test. Suppose, again, that the new items are equally scalable but have difficulty levels between those of the original items. We now have the same information at each ability level but twice the number of ability levels that can be assessed. Formulas such as (36) "blow-up" the index indiscriminately thereby conflating the issues of accuracy and precision.

Horst (1966) makes an effort to distinguish reliability and homogeneity by noting that reliable items are a necessary but not sufficient condition for high H_t. Thus, high H_t is, in part a function of reliability. Now this is true for reliability at the item level. But it is not true for reliability (as internal consistency) at the test level. Again, I am trying here to clearly separate the precision obtained through calibrating a homogeneous or unidimensional test from the accuracy of test.

Bentler's Monotonicity Analysis

I include a discussion of Bentler's (1971) approach here primarily to emphasize that multidimensionality is not an intractable issue when measurement is conceived and operationalized as a cumulative scaling process. Thus far I have avoided the issue of empirical dimensionality suggesting, instead, that a scalable or homogeneous set of items plus reasonable evidence of content validity is a necessary but not sufficient condition for unidimensionality. Although I (and others) often use the terms unidimensional and homogeneous synonymously, it should be understood that the former is not an automatic consequence of the latter.

Preferring the term monotonic (instead of cumulative), Bentler quite cleverly recognized that Yule's Y coefficient (a simple function of the more familiar Yule's Q coefficient) for association in a four-fold table (see Yule, 1912) possessed none of the drawbacks of ϕ or ϕ/ϕ_{max} when subjected to an ordinary principal components factor analysis. For any two items i and j, this index, renamed the monotonicity coefficient by Bentler since he developed it in a more general form, is given as follows:

$$m = \frac{bc - ad}{bc + ad + 2\sqrt{abcd}} \qquad (37)$$

where a, b, c and d are as given in the four-fold table layout in the previous section. The nice thing about Yule's association measure is that it becomes 1 (or -1) only when one (or more) cells are empty. These include exactly those four-fold response patterns of cumulative scales; and a principal components factor analysis of the inter-item m-matrix will recover two or more cumulative scales embedded in a set of items.

As an index of homogenity, m is very similar to H_{ij}. And, like Loevinger, Bentler proposes the average of all k(k - 1)/2 inter-item monotonicity coefficients, $\bar{m}$, as an overall measure of inter-item homogeneity. But then, like Horst, Bentler becomes concerned with the length of the test not being represented in the index. Thus, he proposed the same Spearman-Brown transformation of $\bar{m}$ for a final, overall measure of the test's homogeneity (h),

$$h = \frac{k\,\bar{m}}{1 + (k - 1)\bar{m}} \tag{38}$$

and, in my view, falls into the same trap of mixing up fundamentally distinct measurement issues.

Sato's Student-Problem (S-P) Matrix Analysis

Sato (1980) developed yet another means for indexing departures from the perfect Guttman or cumulative scale. But this time the notion seems to have caught on. It is difficult to tell at this time whether it is the novelty of the procedure (and its more sophisticated mathematical basis) or whether more methodologists have begun to internalize the need to reconceptualize the proper measurement role of items. In any case, Sato's contribution reiterates the appropriate focus for understanding the measurement act, viz., the doubly ordered student-by-item (problem) matrix of raw responses (e.g., Figures 3b-5).

Interestingly, Sato's approach, unlike those discussed previously, utilizes a mathematical model of the ideal non-cumulative response pattern. An index of fit, then, is based on the extent of observed response pattern

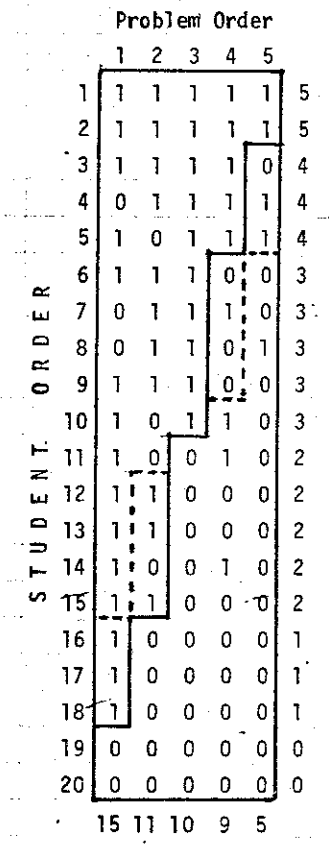
departure from the perfectly heterogeneous model. Specifically, any ordered student-by-problem (item) matrix can be partitioned into sections corresponding to the expected ideal cumulative patterns based on either the student scores, the S-curve, or problem scores (item difficulties), the P-curve.

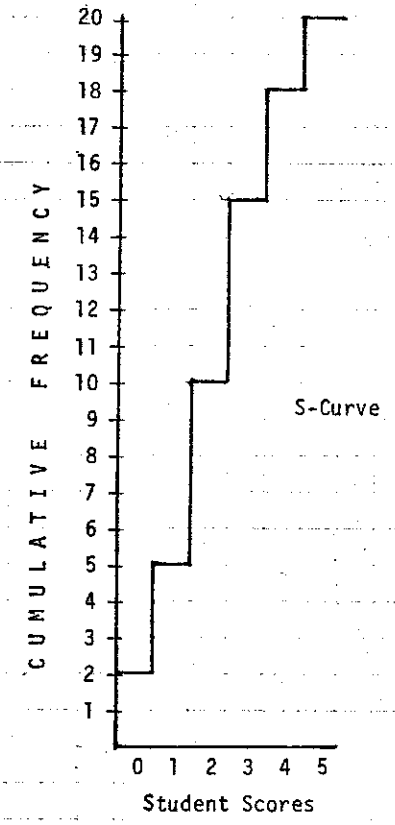
Figure 7 depicts the process of analyzing the student-problem matrix in this manner. Figure 7 is simply Figure 4 again, but this time the cumulative student and problem score distributions are presented, separately, and superimposed, on the S-P matrix itself. As an exercise, superimpose the S-curves and P-curves appropriate for the matrices in Figures 3b and 5. You will discover that in the ideal case (Figure 3b) the S- and P-curves are coincident; and in the case of poor cumulative response pattern (Figure 5), the curves are quite far apart and much more so than they are for the moderately cumulative pattern exhibited here (and in Figure 4).

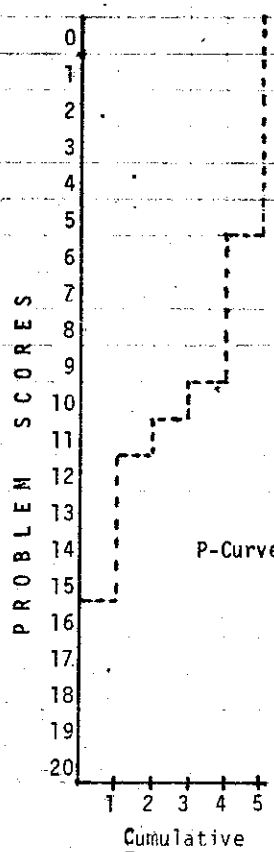
Thus, the area between the S- and P-curves--proportional to the number of student-item responses between the curves--reflects the degree of departure from the ideal cumulative response pattern. (In general, the number of student-item responses between the S- and P-curves is close to, but is not functionally related to, the total number of Guttman errors, viz., twice the number of 0's above, or 1's below, the S-curve.) To construct an index similar to the coefficient of scalability for Guttman scales, the maximum possible area between the S- and P-curves must be calculated for the perfectly heterogeneous student-problem response matrix of the same dimensions and mean performance. Sato models the ideal heterogeneous matrix by assuming simple binomial sampling for problems and students. Thus, the cumulative

Figure 7

S-P Matrix and Cumulative Distributions
for Student Scores (S-Curve)
for Problem Scores (P-Curve)







binomial distributions with parameters k and $\bar{p}$ and parameters n and $\bar{p}$ model the S- and P-curves respectively. Denoting the areas between the observed and binomial S- and P-curves as $A(n,k,\bar{p})$ and $A_B(n,k,\bar{p})$ respectively, Sato's disparity coefficient is given as follows:

$$D = \frac{A(n,k,\bar{p})}{A_B(n,k,\bar{p})} \tag{39}$$

(A more computationally tractable estimate of D is given by Sato, 1980.)

This index reaches 1 in the case of perfect heterogeneity and 0 in the case of a perfect cumulative (homogeneous) response pattern. It therefore varies inversely (and I expect quite highly) with the other indices of homogeneity discussed in this section. Moreover, Sato (1980) defines analogous coefficients at the individual student and problem levels (called caution indices) which serve to highlight those students and items which depart considerably from ideal expectations. Loevinger (1947) developed a similar index for items whereas Guttman relied exclusively on visual inspection of the response matrix. In the final analysis, the increasing popularity of Sato's approach is most likely due to the emphasis placed on the raw score matrix, with handy indices (for spotting aberrant cases) of great practical utility for the ordinary classroom teacher. For recent developments in the U.S., see Tatsuoka (1978), McArthur (1982), Harnisch and Linn (1981), and Miller (1981).

Rasch Measurement: A Latent Trait Model

Latent trait theory, or item response theory (Lord, 1980), refers to a whole class of statistical measurement models based on the same fundamen-

tal conception of the measurement act guiding the cumulative models surveyed thus far. However, latent trait models make important allowances for those "minor" points we glossed over while drawing the analogy to the physical sciences. Specifically, these were the points relating to the variability of both the item difficulty positions as "hash marks" on the "ruler" and the underlying ability continuum itself, as one moves from one "ruler" to the next. Four our purposes here, we will review only the simplest of the latent trait models, viz., the 1-parameter model, developed three decades ago by Georg Rasch. A number of good presentations and/or reviews of latent trait models generally, and the Rasch model in particular, currently exist. Some examples are: Rasch (1980 reprint of 1960 edition); Wright and Stone (1979); Hambleton and Cook (1977; see that entire issue of the Journal of Educational Measurement); Lord, 1980; and Traub and Wolfe (1981).

The Rasch model (and latent trait models generally) assumes a single invariant ability parameter and specifies a probability function over the entire 0-1 range that any item will be answered correctly by students of a given ability. Specifically, Rasch first approached the problem by imagining independent person and item parameters reflecting, respectively, ability and difficulty (or, its reciprocal, easiness). Second, he envisioned the same cumulative response pattern as the ideal outcome when persons with varying abilities encounter items of varying difficulties. But he modeled the process probabilistically, not only to avoid the determinism of previous approaches, but to establish an invariant measurement scale -- so long as the model fits the empirical reality of the test data in question.

The model he selected is a simple odds ratio, i.e., the odds (θ'_{si}) of student s with ability A_s correctly ansering item i with difficulty D_i are given as

$$\theta'_{si} = \frac{A_s}{D_i} \tag{40}$$

Instead of odds, we can use the more convenient 0-1 scale of probability. If P_{si} is the probability of student s answering item i correctly, then, by definition, $P_{si} = \theta'_{si}/(1+\theta'_{si})$. Thus equation (40) can be rewritten as

$$P_{si} = \frac{A_s}{D_i+A_s} \tag{41}$$

It should be clear that, as hypothesized, the model predicts a lower chance of success for a student with lower ability encountering a relatively more difficult item, a higher chance of success for a student of higher ability encountering a relatively less difficult item, and a 50-50 chance of success when the ability of the student and the difficulty of the item are identical. These are invariant properties of the person and the item and are presumed to be independent of each other as well as of the other abilities of the persons being measured and the other difficulties of items doing the measuring. Again, this <u>specific objectivity</u> (as Rasch calls it) is operational only to the extent that these presumptions fit the reality of the data.

Equation (40) becomes computationally more tractable as a simple linear function by taking the logarithm of both sides, i.e.,

$$\log (\theta'_{si}) = \log (A_s) - \log (D_i) \tag{42}$$

Likewise, equation (41) can be so converted; but it is usually expressed in exponential form using the natural base e and the substituted parameters

$\alpha_s = \log\ (A_x)$ and $\delta_i = \log_e (D_i)$. In other words, $e^{\alpha_s} = A_s$ and $e^{\delta_i} = D_i$ and equation (41) becomes the so-called logistic function

$$P_{si} = \frac{e^{\alpha_s - \delta_i}}{1 + e^{\alpha_s - \delta_i}} \qquad (43)$$

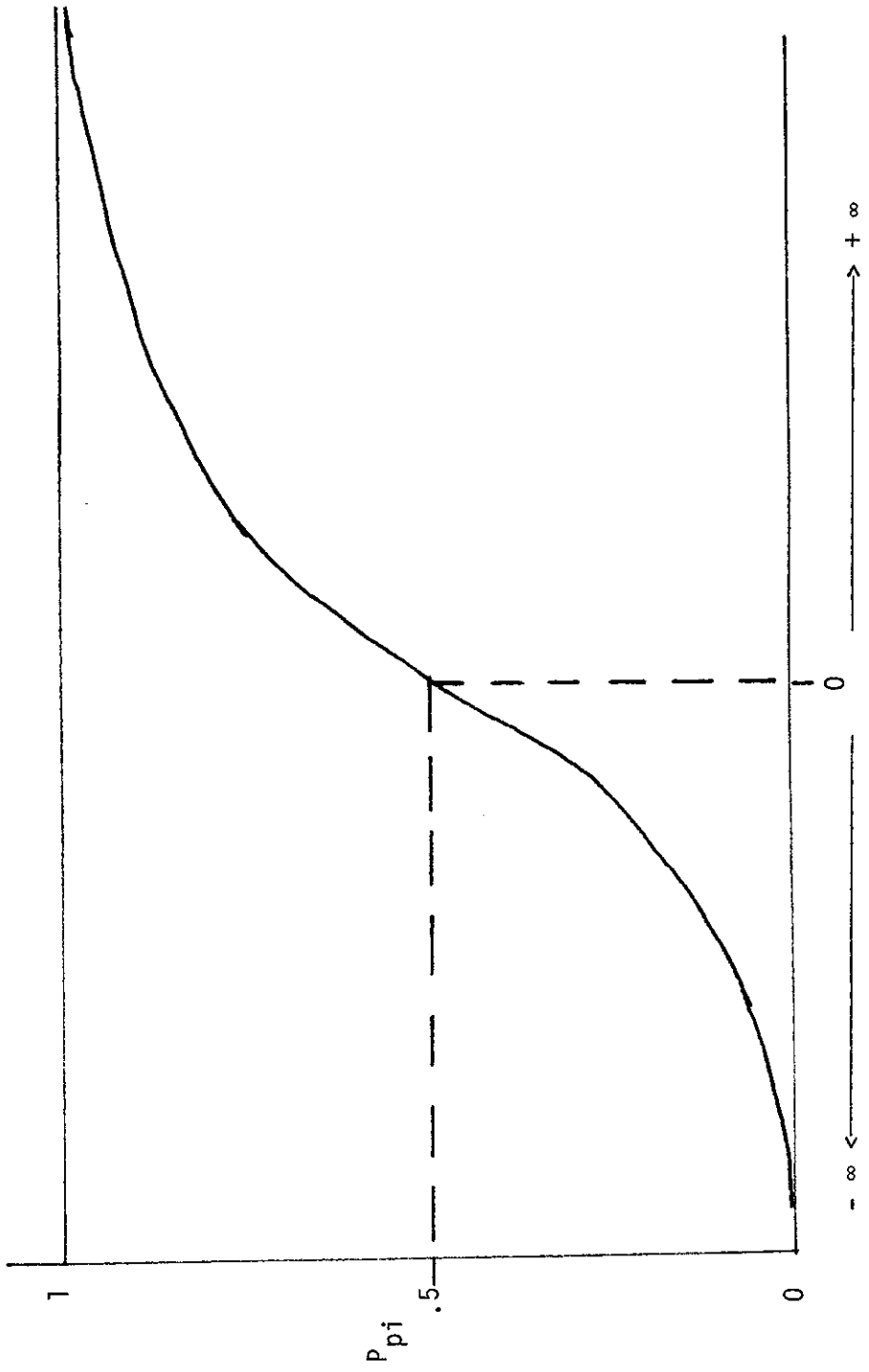
Of course, the same logic is embedded in (43) as was in (41), except now the interplay of person encountering item is reflected in the difference between the transformed ability parameter α_s and difficulty parameter δ_i. When equation (43) is graphed for all possible values of this difference, i.e., for $\xi_{si} = \alpha_s - \delta_i$ where $-\infty \leq \xi_{si} \leq +\infty$, the so-called response characteristic curve results (see Figure 8). This represents the simplest logistic model, often called the 1-parameter model, since P_{si} is really only dependent upon the single discrepancy ξ_{si}. Alternatively, for fixed difficulties δ_i or abilities α_s, the ogive in Figure 8 represents equally well the item characteristic or person characteristic curves respectively.

The rather elegant simplicity of the Rasch technique for scaling is realized through this important property of the model: the student raw scores (r_s) and observed item difficulties (p_i) are sufficient data from which to derive the best estimates of α_s and δ_i respectively. In effect, the double ordering of the student-by-item raw score matrix best estimates the ordering that would occur were we to know the actual α_s and δ_i. Thus, persons with the same raw score r from the same set of items will receive the same ability estimate α_r.

To estimate an α and δ, therefore, the n x k raw score matrix is merely collapsed row-wise such that rows now constitute the k+1 possible raw scores

Figure 8

Item/Person Characteristic Curve



and cell entries are the proportions of persons in the rth raw score group correctly answering the ith item. If the index r is substituted for the index s in equation (43), it should be clear from the above property that these cell proportions ($\hat{P}_{ri}$) are all estimates of their corresponding P_{ri}. In general, then, there are k(k+1) equations of the form

$$\hat{P}_{ri} = \frac{e^{\alpha_r - \delta_i}}{1 + e^{\alpha_r - \delta_i}}$$

with only 2k+1 unknown values of the α and δ.[10] (In practice, no information is provided by raw scores classes r = 0 or k or by observed item difficulties p = 0 or 1 and these rows and/or columns, should they occur, are eliminated for purposes of analysis.)

There are several approaches to the solution of these equations and testing the fit of the results to what the model predicts. (See references noted previously.) The important point for our argument here, however, is that this model again conforms to the measurement of a property as we ordinarily conceive of it. Moreover, when this particular model fits the data reasonably well, the parameter estimates of α and δ are reasonably independent of the particular ability and difficulty levels of specific student and item samples, thereby providing viable approaches to normally thorny testing problems such as test equating, item banking, tailored testing, and so forth.

Finally, it is interesting to note that for each person's ability estimate, there exists a so-called standard error estimate. But the only thing this estimate has in common with the standard error in traditional

test theories is its name. The latent trait standard error is really based upon an information function that reflects the level of precision at the various ability calibrations. It bears no relationship whatsoever to any notion of item/test replication, i.e., accuracy (or dependability). Thus, the latent trait standard error is an index of precision and behaves accordingly, i.e., it is larger for ability estimates towards the extremes and lower for ability estimates towards the center of the item difficulty range.

Summary

To summarize the foregoing view and review, test theoreticians and practitioners must carefully distinguish their model of measurement from their model of the dependability of measurements. The former refers to the concept of precision that is applied in the construction of tests. The latter refers to the concept of accuracy that is applied to the result of testing under specified conditions of use. Items play a central role in measurement models; in models for dependability, they are of incidental importance insofar as the accuracy of estimated ability measurements is of primary importance. Clearly, truly useful test theories necessarily require both measurement and dependability models.

Classical (and classical-like) test theories are really models for the dependability of measurements. They are good for assessing the accuracy of the results of a testing process when the process is conceived as one (or several) of a great many (often infinite) measurement attempts. When each of the repeated measurements is conceived as a replicate (per-

fectly parallel) measure, we have classical test theory as originally developed. When the measurements are conceived as a random sample from a domain of interest (i.e., randomly parallel measures), we have the item sampling versions of classical test theory. At the core of all of these theories, however, is the concept of repeated measurements. Whenever the results of behavioral assessments can be so conceived, classical test theories, in particular generalizability theory, enjoy a wide range of application. (See the recent review by Shavelson and Webb, 1981.)

But these test theories "dig their own grave" when they attempt to translate repeated measurements concepts to the internal structure of the test itself. Recasting items into the role of strictly parallel (or randomly parallel) measurements can't help but give rise to "test construction" procedures based on maximizing inter-item relationships. This procedure automatically eliminates items reflecting ability at the upper and lower ends of the "ruler." Thus, empirical evidence for internal consistency (in the reliability sense) or homogeneity/unidimensionality (in the construct validity sense) is based upon the wrong covariance structure.

In contrast, measurement models attack the issue of test construction directly. They assume a singular construct from the start (relying primarily upon content validation) and proceed to develop items of varying difficulties analogous to hash marks on a ruler. To the extent that the set of items fits the cumulative response pattern expectation, we have evidence (necessary, but not sufficient) that our measurement goal has been achieved. Once satisfactorily constructed, it is quite appropriate that the instrument be subject to all relevant forms of dependability and

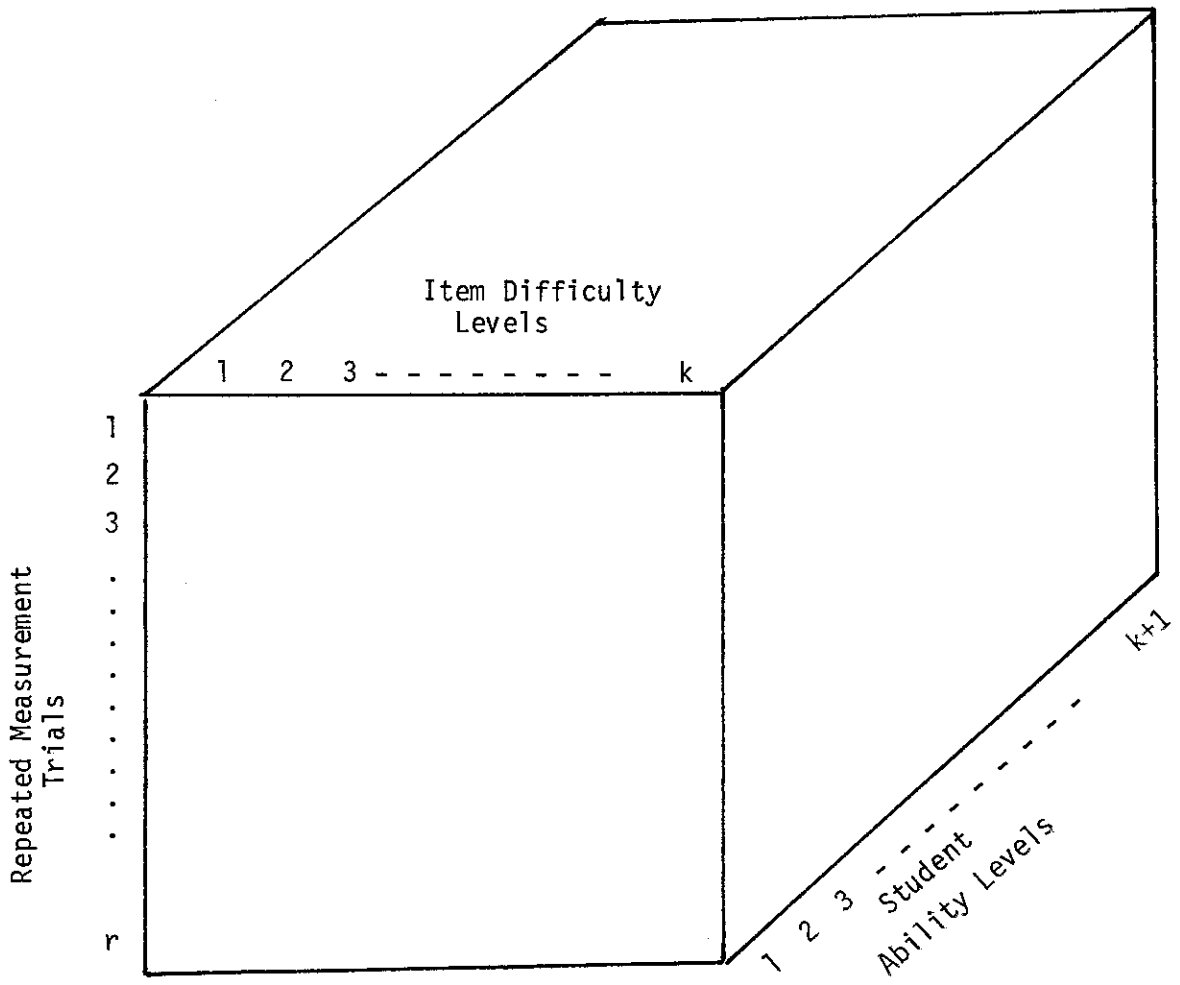
validity procedures under the conditions for use in actual practice. These several ingredients comprise a complete test theory.

Moreover, it should be possible to incorporate dependability at the item level as well. The schematic in Figure 9 portrays the data box necessary to sort out -- at least in theory -- the contrasts between test precision and both item and test accuracy. Vertical slices of the data box contain the data necessary to assess the accuracy of items at each difficulty level for all ability levels. Horizontal slices contain the data necessary to test the scalability of items representing the difficulty levels for each replication. Cross slices could be used to assess the accuracy of items at the various difficulty levels holding ability constant. Collapsing the data box along the difficulty dimension produces the data matrix necessary for assessing accuracy at the test level. Of course, generalizability facets could be crossed or nested with the repeated measurement trials to assess accuracy (dependability) under different conditions. The complete empirical suggestion of Figure 9 may be quite intractable from an operational viewpoint, although, for some highly specifiable items domains (e.g., arithmetic fundamentals) on which ability varies systematically with other measureable examinee characteristics (e.g., age), it may not be too far-fetched.

In conclusion, classical test theory has probably enjoyed a long life not only because of psychological well-being through cognitive dissonance reduction, but because tests have never really been developed without variation in item difficulties. It is time now that we construct tests with varying item difficulties by design--not by happenstance--and use item analysis techniques that correspond to an appropriate theory of measurement.

Figure 9

A Model for Contrasting Accuracy
with Precision and Calibrating a Test
of a Singular Achievement Construct



Moreover, it is fitting that this view forces upon us an issue of perhaps even greater importance, namely, the correspondence of item structure with the cognitive process to be assessed. (See, for example, the arguments recently advanced by Glaser, 1981.) It may well be that the simplistic notions of dichotomous responses (right-wrong) to multiple choice or true-false items are unrealistic indicators of the cognitive processes underlying the abilities we try to measure. Different measurement models from those outlined here may offer more realistic solutions. (For example, see the recent latent class approaches such as Wilcox's (1981) answer-until-correct scheme.)

Footnotes

1. I will use the term "traditional" to refer to classical and classical-like test theories, a distinction that will be clearer in the sequel.

2. I have chosen Spearman's (1910) work, apparently inspired in 1908 by G. Udny Yule (see Yule, 1922), to mark the beginning date for classical test theory.

3. It is important to note at the outset that I do not intend to extol any one notion of what it means to measure achievement. Rather, I wish to explicate a popular intuitive notion of measurement and the extent to which it is compatible with existing measurement theories.

4. In general, I prefer the term "dependability" to the older term "reliability." As used in generalizability theory (Cronbach et al., 1972), dependability denotes reliability under specified conditions of use. At times throughout this report, however, I will use the term "reliability" to facilitate the discussion of traditional test theory concepts.

5. I am using the term "difficulty" here more in a parametric sense than as a synonym for observed p-values.

6. The analogy could be improved upon in this regard by imagining the sticks to be subject to increases or decreases in length as a function of various and sundry effects (some random and some systematic) due to all aspects of the measurement context. This is a less sadistic equivalent of Lumsden's (1976) flogging wall test.

7. Two classical test theory frameworks are in general use. One arises out of the definition of error as proposed originally by Spearman (1910). The other arises out of a definition of true scores as proposed originally by Brown (1910) and elaborated by Kelley (1924). The former approach is presented here since it's simpler. All derivations end up being the same so that it is a purely academic matter which approach is "better." See Gulliksen's (1950) seminal volume on classical test theory and the good hisorical overview by Tryon (1957).

8. An important caveat should be stated here: Except for the latent trait models, the illustrations I have selected do not in and of themselves provide sufficient information for calibrating items and estimating precision. Nevertheless, they are useful both historically and heuristically for underscoring the point of this discussion, viz., the contrast between dependability and measurement.

Footnotes (continued)

9. I am using the phrase "criterion-referenced testing" in the more profound sense rather than simply as a procedure for assessing a criterion level of performance. The criterion is, rather, the content and the attempted isomorphism between the content and the measurement rule. To quote Glaser (1963): "Criterion-referenced measures indicate the content of the behavioral repertory, and the correspondence between what an individual does and the underlying continuum of achievement." (p. 520)

10. Although useful for expository purposes here, this is not really the best procedure for estimating α and δ. (See Choppin, 1983.)

References

Allen, M. J., & Yen, W. M. Introduction to measurement theory. Monterey, CA: Brooks/Cole, 1979.

Bentler, P. M. Monotonicity analysis: An alternative to linear factor and test analysis. In D. R. Green, M. P. Ford, & G. B. Flamer (Eds.), Measurement and Piaget. New York: McGraw Hill, 1971.

Brown, W. Some experimental results in the correlation of mental abilities. British Journal of Psychology, 1910, 3, 296-322.

Choppin, B. The Rasch model for item analysis. CSE Report No. 219. Los Angeles: UCLA Center for the Study of Evaluation, 1983.

Cornfield, J., & Tukey, J. W. Average values of mean squares in factorials. Annals of Mathematical Statistics, 1956, 27, 907-949.

Cox, D. R. The design of an experiment in which certain treatment arrangements are inadmissible. Biometrika, 1954, 40, 287-295.

Cronbach, L. F. Test "reliability": Its meaning and determination. Psychometrika, 1947, 12, 1-16.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. Theory of generalizability: A liberation of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements. New York: John Wiley & Sons, 1972.

Festinger, L. The treatment of qualitative data by "scale analysis." Psychological Bulletin, 1947, 44, 149-161.

Ghiselli, E. E. Theory of psychological measurement. New York: McGraw Hill, 1964.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Glaser, R. The future of testing: A research agenda for cognitive psychology and psychometrics. American Psychologist, 1981, 36, 923-936.

Goodman, L. A. Simple statistical methods for scalogram analysis. Psychometrika, 1959, 24, 29-43.

Gulliksen, H. The relation of item difficulty and inter-item correlation to test variance and reliability. Psychometrika, 1945, 10, 79-91.

Gulliksen, H. Theory of mental tests. New York: John Wiley & Sons, 1950.

Guttman, L. A basis for scaling qualitative data. American Sociological Review, 1944, 9, 139-150.

Guttman, L. A basis for analyzing test-retest reliability. Psychometrika, 1945, 10, 255-282.

Hambleton, R. K., & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Harnisch, D. L., & Linn, R. L. Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 1981, 18, 133-146.

Horst, P. Correcting the Kuder-Richardson reliability for dispersion of item difficulties. Psychological Bulletin, 1953, 50, 371-374.

Horst, P. Psychological measurement and prediction. Belmont, CA: Wadsworth, 1966.

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Kelley, T. L. Statistical methods. New York: Macmillan, 1924.

Kuder, G. F., & Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. Psychological Monographs, 1947, 61(4), Whole No. 285.

Loevinger, J. The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. Psychological Bulletin, 1948, 45, 507-529.

Loevinger, J. The attenuation paradox in test theory. Psychological Bulletin, 1954, 51, 493-504.

Loevinger, J. Persons and populations as psychometric concepts. Psychological Review, 1965, 72, 143-155.

Lord, E. M. Applications of item response theory to practical testing problems. Hillsdale, N. J.: Lawrence Erlbaum Associates, 1980.

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

Lumsden, J. The construction of unidimensional tests. Psychological Bulletin, 1961, 58, 122-131.

Lumsden, J. Test theory. In M. R. Rosenzweig & L. W. Porter (Eds.), Annual Review of Psychology (Volume 27). Palo Alto, CA: Annual Reviews, Inc., 1976.

Magnusson, D. Test theory. Reading, Mass.: Addison-Wesley, 1967.

Maxwell, A. E. A statistical approach to scalogram analysis. Educational and Psychological Measurement, 1959, 19, 337-349.

McArthur, D. L. Detection of item bias using analyses of response patterns. Paper presented to the Annual Meeting of the American Educational Research Association, New York, 1982.

Menzel, H. A new coefficient for scalogram analysis. Public Opinion Quarterly, 1953, 17, 268-280.

Miller, M. D. Measuring between-group differences in instruction. Unpublished doctoral dissertation, University of California, Los Angeles, 1981.

Novick, M. R. The axioms and principal results of classical test theory. Journal of Mathematical Psychology, 1966, 3, 1-18.

Popham, W. J., & Husek, R. T. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press, 1980. (Originally published in 1960 by the Danish Institute for Educational Research.)

Rulon, P. J. A simplified procedure for determining the reliability of a test by split-halves. Harvard Educational Review, 1939, 9, 99-103.

Sagi, P. C. A statistical test for the significance of a coefficient of reproducibility. Psychometrika, 1959, 24, 19-27.

Sato, T. The S-P chart and the caution index. NEC (Nippon Electric Company) Educational Information Bulletin. Japan: Computer and Communication Systems Research Laboratories, 1980.

Schuessler, K. F. A note on statistical significance of scalogram. Sociometry, 1961, 24, 312-318.

Shavelson, R. J., & Webb, N. M. Generalizability theory: 1973-1980. British Journal of Mathematical and Statistical Psychology, 1981, 34, 133-166.

Spearman, C. Correlation calculated with faulty data. British Journal of Psychology, 1910, 3, 271-295.

Stevens, S. S. Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), Handbook of Experimental Psychology. New York: Wiley, 1951.

Tatsuoka, M. M. Recent psychometric developments in Japan: Engineers grapple with educational measurement problems. Paper presented at the Office of Naval Research Contractor's Meeting on Individualized Measurement, Columbus, Missouri, 1978.

Ten Houten, W. D. Scale gradient analysis: A statistical method for constructing and evaluating Guttman scales. Sociometry, 1969, 32, 80-98.

Torgerson. W. S. Theory and methods of scaling. New York: John Wiley and Sons, 1958.

Traub, R. E., & Wolf, R. G. Latent trait theories and the assessment of educational achievement. In D. C. Berliner (Ed.), Review of Research in Education (Volume 9). American Educational Research Association, 1981.

Tryon, R. C. Reliability and behavior domain validity: Reformulation and historical critique. Psychological Bulletin, 1957, 54, 229-249.

Walker, D. A. Answer-pattern and score-scatter in tests and examinations. British Journal of Psychology, 1931, 20, 73-86.

Walker, D. A. Answer-pattern and score-scatter in tests and examinations. British Journal of Psychology, 1936, 26, 301-308.

Walker, D. A. Answer-pattern and score-scatter in tests and examinations. British Journal of Psychology, 1940, 30, 248-260.

Wilcox, R. R. Solving measurement problems with an answer-until-correct scoring procedure. Applied Psychological Measurement, 1981, 5, 399-414.

Wright, B. D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1968.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. D., & Panchapakeson, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

Wright, B. D., & Stone, M. H. Best test design. Chicago: Mesa Press, 1979.

Yule, G. U. On the methods of measuring association between two attributes. Journal of the Royal Statistical Society, 1912, 75, 579-642.

Yule, G. U. An introduction to the theory of statistics. London: Charles Griffin and Co., 1922.