EOLSS
Encyclopedia of Life Support Systems

http://www.eolss.net (entry, *6.27.3.4)*

Reprint of:

THE CONSTRUCTION AND USE OF PSYCHOLOGICAL TESTS AND MEASURES

Bruno D. Zumbo, Michaela N. Gelin, & Anita M. Hubley
The University of British Columbia, Canada

Zumbo, B. D., Gelin, M. N., & Hubley, A. M. (2002). The Construction And Use Of Psychological Tests And Measures. In the Psychology theme of the *Encyclopedia of Life Support Systems (EOLSS),* Eolss Publishers, Oxford, UK.

THE CONSTRUCTION AND USE OF PSYCHOLOGICAL TESTS AND MEASURES

Bruno D. Zumbo, Michaela N. Gelin, & Anita M. Hubley
The University of British Columbia, Canada

Keywords: Tests, measures, reliability, validity, item analysis, item response modeling

Contents

Summary

The successful development and appropriate and meaningful use of psychological tests and measures rests on the validation of inferences made from test scores obtained from a given sample in a given context. The modern, expanded, view of validity as an on-going process argues that researchers need to gather evidence to support the inferences made from the scores obtained on their measures. A general review is presented of a select number of psychometric analyses that can contribute to this evidential basis. The classical test approaches to reliability and item analyses are presented as well as approaches that take into account the latent continuum of variation. This analysis is appropriate after having determined by factor analysis that the items, as a whole, measure one latent variable. These techniques are presented using the Center for Epidemiologic Studies Depression scale (CES-D) as an example. The CES-D is useful as a demonstration because it is commonly used in the life and social sciences for both obtaining scores and for classifying individual respondents. The latter purpose necessitates methods that help one determine the cut-off score for classification (e.g., sensitivity, specificity, and ROC curves).

1. Introduction

The topic of this chapter "The Construction and Use of Psychological Tests and Measures" could nearly fill an encyclopedia of its own. In fact, many books have been written on the historical, mathematical, philosophical, and applied matters in psychological testing and measurement. Given the limited space of this piece, it is by necessity that coverage is selective and that there is a focus on some issues while others are mentioned only in passing or not at all. Furthermore, some of the subtleties that consume psychometricians and measurement specialists will be glossed over. Given that the readers of this volume are life and social scientists who will be selecting, developing, adapting, or using their own tests and measures, the motivation for selection of topics is governed by two goals. The first goal is to provide a bird's eye view of the issues and objectives in test construction by focusing on the matter of selecting items to arrive at tests and measures from which one can make valid inferences. The second goal is to provide a practical presentation of some contemporary approaches to assessing the statistical (psychometric) properties of tests and measures. This latter goal focuses on some of the new technology of measurement and how it may reasonably develop in the future. In this presentation, it is assumed that the reader has an understanding of basic statistics including correlation and regression.

With the above goals in mind, technical matters will be discussed in the context of real data involving a commonly used measure in the life and social sciences: the Center for Epidemiologic Studies Depression Scale (CES-D). The data presented herein is a sub-sample of a larger data set collected in northern British Columbia, Canada. As part of a larger survey, responses were obtained from 600 adults (290 females with an average age of 42 years and 310 males with an average age of 46 years).

2. Psychological Tests and Measures

2.1 What Is A Psychological Test Or Measure?

A psychological test or measure may be viewed as a set of self-report questions (also called "items") whose responses are then scored and aggregated in some way to obtain a composite score. The terms "test" and "measures" are used interchangeably in this context even though "tests" are, in common language, used to imply some educational achievement or knowledge test with correct or incorrect

responses. In many psychological measures (e.g., attitudinal measures), there are not “correct” or “incorrect” responses, per se. Furthermore, the term “scale” is also often used in the life and social sciences interchangeably with the term “questionnaire” to refer to the set of questions whose responses are aggregated into a composite score. The essential features therefore are: (a) a series of questions to which an individual responds, and (b) a composite score that arises from scoring the responses to these questions. The resultant set of questions together is referred to as a “scale”, “test”, or “measure”.

Two types of scores can be obtained from items, but it is important to note that it is not the question format that is important here but the scoring format. Binary scores, which are also referred to as dichotomous item responses, are obtained from either: (a) items (e.g., multiple choice) that are scored correct/incorrect in aptitude or achievement tests, or (b) items (e.g., true/false, agree/disagree) that are dichotomously scored according to a scoring key in an attitude, opinion, or personality scale. Ordinal item responses, which are also referred to as graded response, Likert, Likert-type, or polytomous items, involve more than two scoring options such as a 5-point strongly agree to strongly disagree scale on a personality or attitude measure. Note that, in this context, the word polytomous is used to imply ordered responses and not simply multi-category nominal responses. For simplicity and consistency with the life and social sciences literature, the various terms denoting ordered multicategory scores will be referred to as "Likert-type" throughout this piece although this deviates from the original and very strict definition of a Likert format. An interesting feature of ordinal or Likert-type scores is that, for some research purposes, they can also be re-scored in a meaningful binary fashion.

The items in a test or measure are considered to be indicators or markers of the phenomenon under study (also called a construct or latent variable) and therefore their composite is also an indicator of the phenomenon and not the phenomenon itself. For example, the CES-D is a 20-item scale introduced originally by Lenore S. Radloff to measure depressive symptoms in the general population. It has also been shown to be useful in clinical and psychiatric settings although it is not intended for diagnostic purposes, but rather as an index of current feelings of general depression. The CES-D has been translated into many different languages and is widely used in both large-scale and small-scale epidemiologic studies. The key point here is that the composite (i.e., scale) score is not depression itself but rather an observable indicator of depression -- or more accurately, the score is an indicator of depressive symptoms.

The CES-D prompts the respondent to reflect upon his/her last week and respond to questions such as “My sleep was restless” using an ordered or Likert-type response format of “not even one day”, “1-2 days”, “3-4 days”, “5-7 days” during the last week. The items typically are scored from zero (not even one day) to three (5-7 days). Composite scores therefore range from 0 to 60, with higher scores indicating higher levels of depressive symptoms. It was noted above that Likert-type items are sometimes re-scored into a binary format. Several such re-scoring options can be found with the CES-D. A very common binary re-scoring of the CES-D is used when researchers are only interested in the presence or absence of depressive symptomology rather than a degree of symptomology so they score all responses other than “not even one day” as “1” so that the resulting scale is “not even one day” equals 0 and all other responses equal 1. This binary scoring format is sometimes called the “presence method” of scoring. Note that as the example shows, re-scoring may result in the instrument measuring a subtly different construct. Throughout this chapter, the original Likert-type response format, which conveys not only presence or absence of symptoms but also the degree, will be used.

As a note to the general social and policy researcher, although our example focuses on a psychological dysfunction, the methods in this piece also apply to scales of opinions and attitudes (e.g., a measure of

one's feelings of personal safety, life satisfaction, or spending preferences; attitudes toward social policies, gun control, or abortion).

2.2 For What Are Tests And Measures Used?

There are two main purposes of measurement in applications in the life and social sciences:

- Descriptive: Assigning numbers to the results of observations for the purpose of obtaining a scale score in scientific or policy research.
- Decision-making: Using the scale scores to categorize individuals or groups of individuals based on their responses to the test or measure.

The latter purpose subsumes the former but is also concerned with setting cut-off scores used to meaningfully categorize individuals or groups of individuals. For example, a cut-off score of 16+ is commonly used with the CES-D in epidemiologic studies to yield an estimate of the proportion of individuals in the population likely to have a disorder severe enough to require professional intervention.

2.3 Organization Of This Article

In summary, it should become evident as one progresses in understanding measurement technology that the field distinguishes items, scales, and the phenomenon of interest. Individuals respond to statements or questions, the responses are then combined into a composite score, and the composite score is related to the phenomenon of interest. The phenomenon itself is often unobservable and hence is referred to as a latent variable. In the most commonly used statistical measurement techniques, the phenomenon of interest is assumed to be a quantity (as opposed to some sort of typology); thus, the latent variable is assumed to be a continuous latent variable.

In the case of the CES-D, individuals respond to 20 statements describing depressive symptoms occurring within the last 7 days. These responses are then combined into a composite scale score. The composite scale score is not the phenomenon of depression, per se, but rather is related to depression such that a higher composite scale score reflects higher levels of the latent variable depression. In describing measurement in this way, it seems obvious that a primary concern should be the selection of questions or items that adequately reflect symptoms of depression. Cast in this way, a central question of evaluating, developing, and adapting tests and measures is how the items come together to reflect the phenomenon of interest. This question is addressed through item analysis. The item analysis technology of tests and measures was developed to help answer the following practical questions faced by researchers and policy-makers alike: (a) Given that the items are combined to created one scale score, do they measure just one latent variable? (b) How much of the observed variation is true variation and therefore how precisely do the items measure? and (c) How does this precision change across the levels of the continuous latent variable? Due to space limitations, a description of methods (differential item functioning) to investigate whether the items measure differently for different groups (e.g., males and females) is not included. The reader should note that one should not confuse precision and accuracy. The former term implies little measurement error whereas the latter term implies that one is tapping the dimension of interest (rather than some other dimension).

Sections 2 – 3 of this chapter present methods to answer each of these three questions, respectively, and focus on the descriptive purpose of measurement described above. Section 4 will concern itself with the techniques of the decision-making purpose of measurement. As a whole, this chapter concerns itself, in

its essence, with validation; therefore, the chapter ends with a review of current thinking in validation as it applies to test and measures. This last section brings together all of the previous sections with the purpose of providing evidence for the validity of the inferences one makes from the scale scores.

3. Do The Items Measure Just One Latent Variable?

An interesting and provocative historical point is that, in the early 1900s, Professor Charles Spearman presented two separate papers analyzing the same data two different ways. In one paper he introduced the foundations of the methods for answering the question of whether items measure just one latent variable (i.e., factor analysis). In the other paper, he introduced the fundamental ideas to answer the question of how much of the observed variation is true variation (i.e., reliability and classical test theory). It has been argued that these two papers represent one underlying mathematical model, often called factor analysis, that describes the relation between observed and latent variables (i.e., unobservable variables). It has been further argued that over the course of the next century of research factor analysis and reliability theory have been treated as essentially different models when, in fact, Spearman, their developer, may have viewed them as interrelated models, if not the same mathematical model.

To answer the question of whether the items on a test measure one or more latent variables, measurement specialists historically (due to Spearman's work in the early 1900s) have focused on the covariation among the items comprising a scale. The statistical theory is based on the assumption that items covary among themselves because they have some unobservable (latent) variable in common. The latent variable, of course, is the construct or phenomenon of interest measured by the set of items. In other words, the latent variable accounts for the covariance among the items and represents the attribute that the item responses share in common – hence this is sometimes called "common factor analysis".

In the framework of modern statistical theory, the previous paragraph describes the analysis of covariance matrices using covariance structure models. In the context of this section, these covariance structure models are called confirmatory factor analysis (CFA) models. In the typical CFA model, the score obtained on each item is considered to be a linear function of a latent variable and a stochastic error term. Assuming p items and one latent variable, the linear relationship may be represented in matrix notation as

$$\mathbf{X} = \Lambda\xi + \delta, \tag{1}$$

where $\mathbf{X}$ is a (p x 1) column vector of scores for person i on the p items, Λ is a (p x 1) column vector of loadings (i.e., regression coefficients) of the p items on the latent variable, ξ is the latent variable score for person i, and δ is (p x 1) column vector of measurement residuals. It is then straightforward to show that for items that measure one latent variable, Equation 1 implies the following equation:

$$\Sigma = \Lambda\Lambda' + \Psi, \tag{2}$$

where Σ is the (p x p) population covariance matrix among the items and Ψ is a (p x p) matrix of covariances among the measurement residuals or unique factors, Λ' is the transpose of Λ, and Λ is as

defined above. In words, Equation 2 tells us that the goal of CFA is to account for the covariation among the items by some latent variables.

More generally, CFA models are members of a larger class of general linear structural models for a *p*-variate vector of variables in which the empirical data to be modeled consist of the *p x p* unstructured estimator, the sample covariance matrix, *S*, of the population covariance matrix, Σ. A confirmatory factor model is specified by a vector of *q* unknown parameters, θ, which in turn may generate a covariance matrix, $\Sigma(\theta)$, for the model. Accordingly, there are various estimation methods such as generalized least-squares or maximum likelihood with their own criterion to yield an estimator $\hat{\theta}$ for the parameters, and a legion of test statistics that indicate the similarity between the estimated model and the population covariance matrix from which a sample has been drawn (i.e., $\Sigma = \Sigma(\theta)$). That is, formally, one is trying to ascertain whether the covariance matrix implied by the measurement model is the same as the observed covariance matrix,

$$S \cong \hat{\Lambda}\hat{\Lambda}'+\hat{\Psi} = \Sigma(\hat{\theta}) = \hat{\Sigma}, \tag{3}$$

where the symbols above the Greek letters are meant to imply sample estimates of these population quantities.

As in regression, the goal of CFA is to minimize the error (in this case, the off-diagonal elements of the residual covariance matrix) and maximize the fit between the model and the data. Most current indices of model fit assess how well the model reproduces the observed covariance matrix.

In the example with the CES-D, a CFA model with one latent variable and some specified error covariances reflecting the test format was specified and tested using a recent version of the software LISREL. Suffice it to say that an examination of the fit indices such as the Chi-square test and the root mean-squared error of approximation (RMSEA), a measure of model fit, showed that the one latent variable model was considered adequate for the purpose of demonstrating the item analysis techniques that follow in the sections to come.

4. How Much Of The Observed Variation Is True Variation And How Precisely Do The Items Measure?

4.1 Classical Test Theory And Omnibus Measures

A key element of classical test theory that was introduced by Charles Spearman is the idea that an observed score is a linear combination of a true score and an error score (although hints of this idea were evident in a less formal way in the earlier work of F. Y. Edgeworth). In this framework, the true score is that attribute which the items share in common, and the error score is the difference between the observed and true scores. More formally, let the observed score, X, for an individual be the algebraic sum of two components, the true score, T, and the error score, E, $X = T + E,$ and with the fairly standard assumptions that: (a) the covariation among T and E is zero, (b) if one takes into account the true score, the average error is zero, and (c) the errors are not correlated across items. With these assumptions in hand, it is well known that the theoretical reliability of a scale can be defined as

$$reliab = \frac{\text{var}(T)}{\text{var}(T) + \text{var}(E)} = \frac{\text{var}(T)}{\text{var}(X)}, \tag{4}$$

which equals the squared correlation of T and X. This formulation means that the theoretical reliability of a measure is the proportion of observed variance that is true variance or likewise the variance in observed scale scores that is due to differences between individuals.

Of course, one does not have at hand the true variance so many rather clever methods have been developed over the course of nearly a century of work in test theory to estimate the theoretical reliability from observed data. Four such methods are: (a) to create two parallel or even alternate forms of a test and compute the correlation coefficient between the two halves, (b) to simply retest the individuals and compute a test-retest correlation, (c) to divide the items into two separate halves and compute a split-half correlation between the two halves, and (d) to compute Cronbach's coefficient alpha. The most widely used approach is that of Cronbach's coefficient alpha. Interestingly, Cronbach's coefficient can also be seen from Hoyt's earlier framework, wherein he introduced the fact that classical test theory can be formulated as a random effect analysis of variance, or variance components. Hoyt's results are noteworthy because they laid the foundation for many of the developments in generalizability theory, which allows us to partition and take into account the sources of error variation.

Coefficient alpha has evolved into two separate, but interrelated uses: (1) as an estimate of the theoretical reliability in Equation 4, and (2) as an estimate of the internal consistency of responses to the scale items. The reason for the second use is that coefficient alpha can be conceptualized as an average of all possible adjusted split-half correlations. With this conceptualization in mind, a coefficient alpha of large magnitude would mean that the items are highly interrelated and hence "hang together well". However, this internal consistency approach has led to an inappropriate use of coefficient alpha wherein it has been said by some text authors and researchers that a large coefficient alpha means that the test is measuring a univocal (i.e., unidimensional) construct. In fact, it has been shown in the psychometric literature that this is not true (e.g., coefficient alpha can be large even though the item responses are accounted for by more than one latent variable) and that matters of the number of latent variables or dimensions tapped by a set of items is best considered through the use of factor analysis. In an important sense, the internal consistency conceptualization has misled researchers and has since been supplanted by developments in factor analysis. Therefore, it is recommended that one only use coefficient alpha as an estimate of the theoretical reliability as described in Equation 4, and that discussions of how the items "hang together" or whether they tap one latent variable be addressed using factor analytic techniques.

At the beginning of this section measurement error was introduced as the discrepancy between the observed score and the true score. Measurement error can be of two types or causes: (1) unsystematic errors, and (2) systematic errors. Unsystematic errors include item selection, test administration factors such as anxiety or environmental noise, and test scoring that change from one test session to another. Systematic errors, on the other hand, are error that occur consistently but are often unknown to the tester. For example, a test may claim to measure intelligence, but because of the way it is set up or the questions it asks, it is also actually measuring something else like anxiety or speed of responding. The error term in Equation 4 refers to random error.

The standard error of measurement is basically an estimate of the amount of random error surrounding scores on a particular measure. If the standard error of measurement was zero, then that would indicate that there was no random measurement error at all. The larger the magnitude of the standard error, the more random measurement error that is present. More formally, the standard error of measurement describes an estimate of the standard deviation of the distribution of test scores that would presumably be obtained if a person took the test an infinite number of times. Of course, the average of these infinite independent test scores obtained from a person is the true score. The standard error of measurement is computed as $sd \times \sqrt{1 - reliability\ estimate}$, where sd denotes the sample standard deviation of the composite score.

The widely available software package SPSS was used to compute Cronbach's coefficient alpha for the CES-D scale for the 600 respondents described earlier. The resultant coefficient alpha is 0.906. What this means is that if the responses to the 20 items can be accounted for by one latent variable (as was shown above to be the case), then roughly 90% of the observed variance is true score variance. The assumption of one latent variable is sometimes stated (using a different terminology and psychometric framework) as essential tau-equivalence. Of course, it is also well known that if essential tau-equivalence does not hold (i.e., more than one latent variable is involved) then coefficient alpha is a lower bound estimate of the theoretical reliability. That is, the theoretical reliability will be greater than or equal to the coefficient alpha. For our sample, the standard deviation of the composite score is 9.46 therefore the standard error of measurement is 2.90 (i.e., $2.90 = 9.46 \times \sqrt{1 - 0.906}$). Note that the standard error, in this case, is measured in the zero to sixty scale of the composite. If, however, one were using the sample z-scores the standard error of measurement would be 0.307. This is a reminder that the magnitude of the standard error of measurement reflects the metric that one is using for interpretations, in this case either the raw composite or the raw composite transformed to z-scores.

The reliability coefficient has one inherent limitation in its interpretation that needs to be kept in mind. That is, it is not known which, of two general sources, is the actual cause of a low reliability coefficient. As can be seen from Equation 4 above, low reliability can result from two possible sources. The first source is poor measurement precision or said another way, high error variance. That is, if, for illustrative purposes, one holds constant the true score variance, then one could see that a low reliability can result from an increase in the error variance. The second source is a sample with low true score variation. That is, from Equation 4, if one held constant the error variance, one would see that a low reliability can result from decreased true score variance. This may also be explained as the effect that a restricted range of true scores would have on the squared correlation between true scores and observed scores. These sources of unreliability are important to note because they play an essential role in understanding and interpreting the reliability of scores. It is also worth noting that although one can determine the proportion of observed score variance that is due to either true score variance or error variance, one does not know the amount or magnitude of these variances.

Because of the limitations of classical test theory estimates of reliability described above, some measurement specialists have recommended that classical reliability estimates (and hence standard error of measurement) no longer be used in measurement practice. Others do not take this view and instead warn practitioners that they need to keep in mind this inherent limitation when interpreting classical reliability estimates. The rationale for not abolishing the traditional reliability conceptualization is two-fold. First, the reliability coefficient shares its limitation with all correlation-based statistical methods and hence all of those methods would have to be abolished as well (e.g., regression analysis, analysis of variance, and other general linear statistical model methods). Second,

practitioners simply need to be clear about precisely what the reliability coefficient, as an estimate of the theoretical reliability, is providing; that is, the proportion of the observed score variation that is true variation. Put another way, one can think of variation as noise and reliability as an indicator of what proportion of that noise is signal, as in a signal to noise ratio.

Two additional points are worth noting. First, reliability is a property of the test scores (and not the test itself) and thus it is dependent upon both the sample of respondents from which, and the context in which, the scores were obtained. Second, there has been increased recognition that reliability estimates might also vary for scores representing different levels of the latent variable. Therefore, it is necessary that researchers report: (a) a classical reliability coefficient such as coefficient alpha based on their own sample of data (rather than simply reporting favorable reliability coefficients from the literature or a test manual), and (b) a reliability estimate that takes into account the varying levels of the latent variable – such as the test statistics based on nonparametric item response modeling. The latter point is particularly important for test developers as well. These two steps together will give the researcher a sense of how much of the observed variation is true variation and then how that quantity varies across the levels of the latent variable (e.g., for high, moderate, and low scorers). The next section is devoted to presenting an item response modeling methodology.

4.2 Nonparametric Item Response Modeling: Considering The Latent Variable In Psychometric Analyses

As described above, commonly used classical test indices provide an overall index of reliability irrespective of the level of the latent variable whereas the item response modeling approach provides information over the range of the latent variable.

Most of the item analysis techniques used in scale construction and scale evaluation are based on *classical* test theory. The common indices are:

- coefficient alpha,
- item-total correlations (which measure each item's capacity to discriminate among individuals).

As a brief refresher, scale or questionnaire construction (i.e., the forming of aggregate scores) involves designing a set of items that are intended to be indicators of where someone is placed on a continuum of a quantitative latent variable (or *a continuum of variation*). Item analysis uses the item response data to assess the extent to which each item indicates the continuum of variation; in this early empirical stage, the existence of the latent variable is inferred from the relations among the item responses. Although one begins with a theoretically defined latent variable, that latent variable later becomes an empirical inference by virtue of evidence from an exploratory or confirmatory factor analysis.

It should be noted, however, that the item analyses presented herein are founded on the notion that the item responses are primarily determined by the amount, or level, of some *single* continuum of variation (i.e., a latent variable). Alternatively, however, a researcher can also be interested in investigating the extent to which a single latent variable determines performance -- i.e., how do the items perform if the scale or aggregate measure is treated *as if* it were univocal. Both exercises above are relevant when one is investigating the validity of the inferences made from the aggregate (or total) scale score.

Techniques in item response modeling (either parametric or nonparametric) allow researchers to model item responses as a function of a continuum of variation. Importantly, the item response modeling approach provides information over the range of the latent variable. In a sense, then, a researcher using

item response modeling can obtain more information about how the item performs in relation to the scale score (or aggregate) than they would using the classical indices. This is an important advantage of item response modeling.

There are two main forms of item response modeling: parametric and nonparametric. Parametric item response modeling conceives of the item response function with a parametric form. Nonparametric item response modeling lacks such strict assumptions about the form of the relationship and more closely models the functional relationship with the data at hand. We will focus on nonparametric item analysis because we envision readers using it on small data sets (e.g., less than 300 respondents) and prefer graphical depictions of the item responses. A particularly useful approach to nonparametric item response modeling has been developed by Jim Ramsay. The presentation herein closely follows Ramsay's development.

In describing item response modeling, it is a reasonable assumption that each individual responding to an item of the CES-D scale possesses some amount, θ, of depression. Item response modeling (often called item response theory) has the basic, but fundamental, goal of developing statistical models that account for the likelihood of endorsing an item as a function of some characteristics of the item itself (i.e., item parameters; an example of which might be the ability of items to discriminate among individuals) and of the amount of the latent variable, θ.

The cornerstone of item response modeling is the item response function, which is the relationship between the likely item response and the various levels of the continuum of variation. A useful way to introduce nonparametric item response modeling is by starting with the commonly-found regression model in social research. Figure 1 depicts a scatter diagram of a linear regression of

$$E(y|\theta) = \alpha + \beta\theta + \varepsilon , \tag{5}$$

where the expected value of y -- $E(y|\theta)$; more precisely the conditional distribution of y -- is a linear function of θ, α is an intercept term, β is the slope, and ε denotes error, as is commonly found in simple linear regression.

The model in Equation 5 underlies traditional item analysis procedures described above such as those for calculating item discrimination (i.e., the item-total correlation) and estimates of reliability.

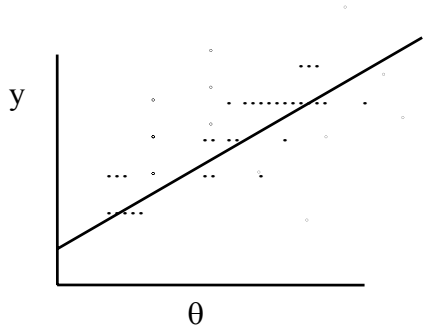

Figure 1. Generic Scatter Diagram With Plotted Regression Line

In essence, what Equation 5 states is that at the various values of θ there is a sub-distribution (also called a conditional distribution) of *y* scores, and the center of each of these sub-distributions is lined up in a straight line. The simplest way to conceptualize the conditional distributions in regression is to think of a population scatter diagram as in Figure 2.

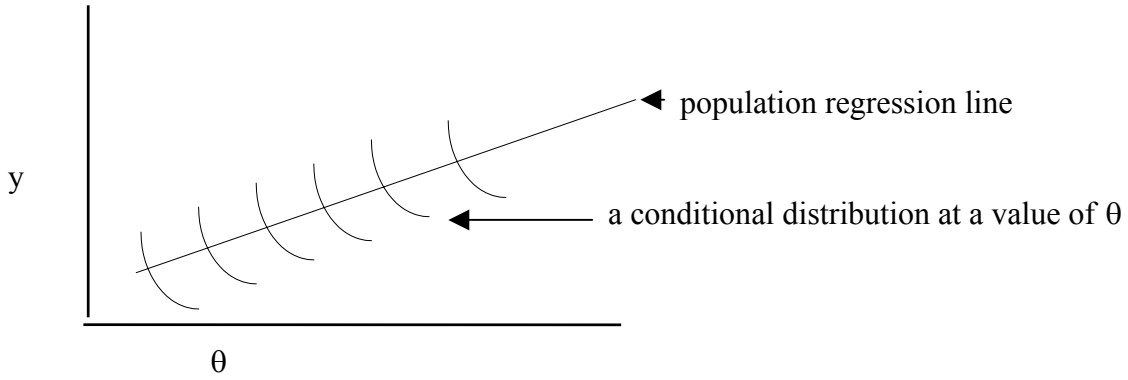

Figure 2. A Scatter Diagram Depicting Conditional Distributions in Regression

If, instead of Equation 5, one sets out to model the conditional distributions of *y* at the various values of θ in a naïve nonparametric regression one would conceptually compute the conditional distribution at each and every value of θ . However, one would have to assume that there are many replications of y at each value of θ , and one would indeed need a very large sample size to accommodate such an approach. Nonetheless, if one assumed such a large and diverse sample composition, one could then see how the mean (or if the data is skewed, the median) of these conditional distributions line up and it would not have to be a straight-line pattern but rather one would let the data speak for themselves. In this case, the relationship between *y* and θ would not necessarily have to be linear as in Equation 5. Instead, one could depict this relationship in the following way:

$$E(y|\theta) = f(\theta) + \varepsilon \,. \tag{6}$$

The resulting regression line would follow the data closely. In essence, one would dissect the θ into a large number of narrow intervals and plot a conditional distribution at each one of those intervals as in Figure 3.

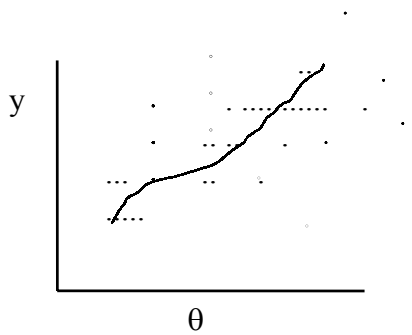

Figure 3. Generic Scatter Diagram With Plotted Nonparametric Regression Line

Of course, in practice it is more typically the case that one has a relatively small sample size at hand and one would have to work with fewer intervals along θ, each interval possibly being larger and containing relatively few data points. The class of nonparametric regression methods used in nonparametric item response modeling is a more sophisticated approach of interval widths and estimation within those intervals. Ramsay's approach, as implemented in the freely distributed software program TESTGRAF, relies on Kernel smoothing procedures, i.e., a Gaussian Kernel. In fact, in using this technique, the form of the function relating y and θ is determined by the data at hand and less by any explicit pre-determined form, such as a linear function. The values of *y* may, for example, increase or decrease as a function of θ.

In summary, in this nonparametric variety of item response modeling, rather than conceive of the item response function as a well-behaved smooth increasing function with a parametric form, one can instead use nonparametric regression to model the functional relation between the likely value of *y* to the value of a latent variable. The technique is nonparametric in the sense that, in comparison to parametric forms, it lacks assumptions about the form of the relationship between the response and explanatory variables.

Figure 4 is an example of an item response function computed through nonparametric item response modeling, using Ramsay's TESTGRAF computer program. Figure 4 depicts the nonparametric item response function for the CES-D item "My sleep was restless" on a response format of (0) "not even one day", (1) "1-2 days", (2) "3-4 days", and (3) "5-7 days" during the last week.

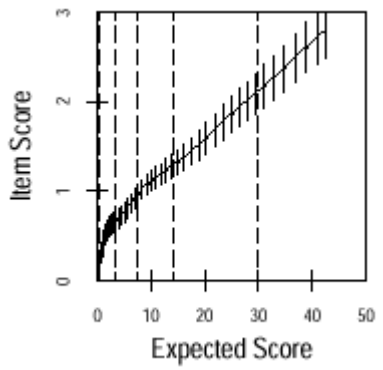

Figure 4. A Nonparametric Item Response Function For CES-D Item 11 Using TestGraf.

The X-axis represents the latent variable, θ, score and the Y-axis is the likely response at the corresponding θ score. Note that the function in Figure 4 is not as well-behaved as the function given in Figure 1. Again, the form of the function in Figure 4 is determined in part by the data at hand and not by a pre-specified form, like that in Equation 5 and depicted in Figure 1. It is important to note that the purpose of this type of nonparametric item response modeling, as developed by Ramsey, is to replace the equation and parameters in Equation 5 and Figure 1 with graphical displays.

The nonparametric item response function, commonly called the item characteristic curve, ICC, is depicted as a solid line in Figures 4 and 5. Along this function there are also small vertical solid lines which indicate the 95% pointwise confidence limits (these are not 95% confidence limits for the entire curve but rather confidence limits at particular points on the continuum of variation).

Also shown on the plot in Figures 4 and 5 are vertical dashed lines across the entire plot indicating various quantiles of the standard normal distribution. The third dashed vertical line, from the right, indicates the mean while 50% of the scores lie between the second and fourth lines, and 5% beyond the first and fifth vertical dashed lines. Going from left to right, the dashed vertical lines indicate the 5th, 25th, 50th, 75th, and 95th percentiles, respectively. In fact, the far left dashed line is nearly on the Y-axis. The fact that the dashed lines are all tending toward the left of the X-axis, tell us that the latent scores are bunched up near the lower scores.

Figure 5 is a density plot (i.e., a type of probability or frequency plot) of the expected (i.e., latent) score on the CES-D. The plot gives us a sense of the distribution of the observed variable and suggests that most of the scores are at the low end of the scale (i.e., indicating little to no depressive symptomatology). This finding of skewed scores is, of course, expected in a general population of respondents.

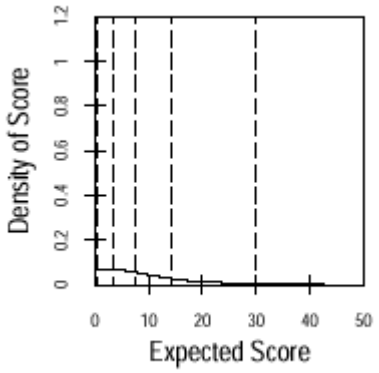

Figure 5. Density Plot of the CES-D Expected (Latent) Scores.

In all of the graphical displays that follow, it is important to note that nonparametric item response modeling, as implemented in TESTGRAF, does not use the numerical values of the observed total score (or aggregate score) in its displays. Rather, because the goal of the item analysis is to work at the latent variable level of analysis, TESTGRAF first replaces the observed aggregate score by their ranks, and then replaces the ranks by the corresponding standard normal quantiles (z-scores) prior to applying the Gaussian Kernel smoothing for the nonparametric regression. One is able to perform the computations because one cannot measure a latent variable in the usual sense. As Ramsay discusses, one can only measure the latent variable to within any transformation that preserves the rank order among the observed total scores -- called a monotone transformation.

Although the class of monotone transformations is limited, it allows for several possible options for the abscissa (X-axis) on the graphical displays, e.g., standard normal quantiles or formula scoring. In some applications, researchers have chosen to display the standard normal quantiles (z-scores) because they are familiar to most researchers and they are the quantities used in computation of the smoothing regression. In this presentation, the expected scores are displayed in terms of the original scale, 0-60, as described above.

As a final note, because of both the method of estimation and the reliance on graphical displays, the nonparametric item response modeling presented here does not have the excessive demands on sample

size found in the parametric item response models. The parametric models (particularly the 3-parameter model) require a thousand or more subjects to get adequate parameter estimates and thus are used exclusively by large testing companies like ETS in the United States, which have access to very large datasets. The nonparametric item response modeling approach developed by Ramsay, however, has made it viable for many researchers to use item response modeling.

One can obtain an intuitive understanding of the item response function and the information it portrays by noting, for example, that for the function shown in Figure 4, when one has an expected scale score of 0.0 on the latent variable, i.e., they display no depressive symptomatology, the most likely item response to the question "My sleep was restless" is 0 ("not even one day"). Likewise, individuals who are expected to score approximately 16 (the cut-off typically used to help compute the prevalence of depression in a population) on the total scale will likely provide a response of 1 ("1-2 days in the last week") to that question. Note that the "likely responses" for the prior sentences were obtained by drawing a line parallel to the abscissa for the lower and upper pointwise confidence limits at the various expected scores. Finally, as expected, the item response function starts at the bottom left of Figure 4 and increases steadily to the top right corner of that plot. The slope of that line at any given interval of the expected score gives one the sense of how well the item discriminates among respondents. Clearly a flat item response function would mean that everyone, irrespective of their expected scale score, would give the same likely response to the item. This would not be a particularly useful item. Furthermore, one may be able to imagine that an item for which the item response function starts at the top left and then declines (while all the other items are the opposite slope) is either incorrectly scored or, performing terribly.

Figure 6 shows the item response functions for four of the 20 items of the CES-D with a sample of 600 respondents. Figure 7 lists all of the corresponding questions so that the reader can get a sense of what a psychological test looks like and it will aid in interpreting how the items in Figure 7 are performing. For example, it is instructive to compare item #2, "I did not feel like eating; my appetite was poor" with items #14, #18, or #20. Clearly when looking at item #2 it can be seen that one needs to be exhibiting a very high level of depressive symptomotology before one will endorse the appetite question much beyond 1-2 days per week.

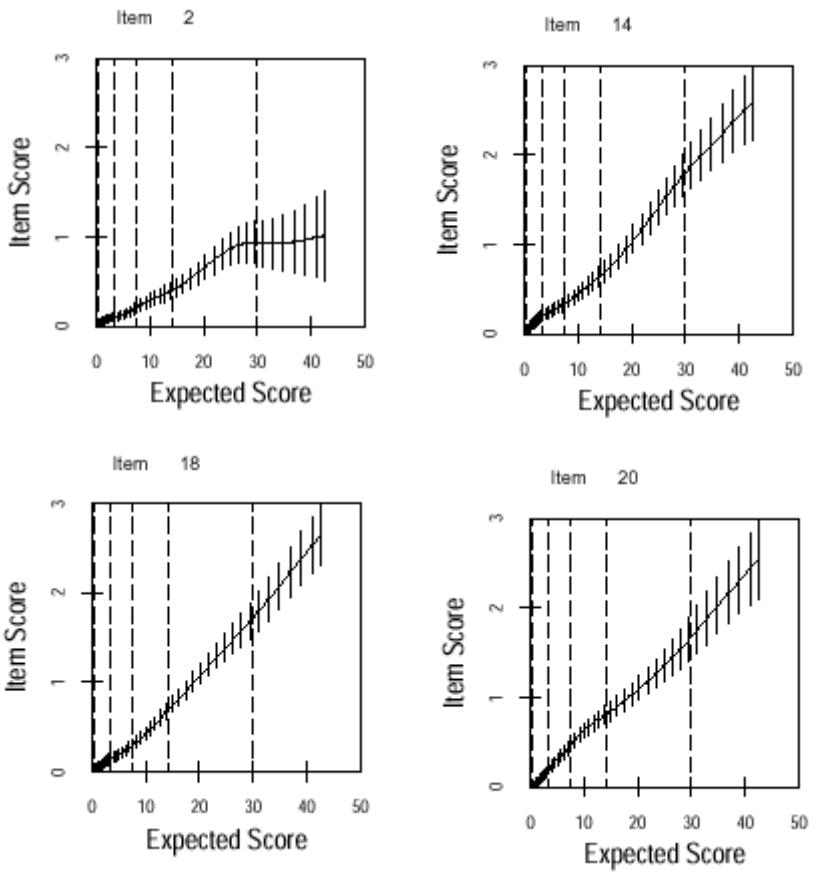

Figure 6. Item Response Functions for Four of the CES-D Items.

For each statement, circle the number (see the guide below) to indicate how often you felt or behaved this way **during the past week**.

0 = rarely or none of the time (less than 1 day)
1 = some or a little of the time (1-7 days)
2 = occasionally or a moderate amount of time (3-4 days)
3 = most or all of the time (5-7 days)

	not even 1 day	1-2 days	3-4 days	5-7 days
1. I was bothered by things that usually don't bother me.	0	1	2	3
2. I did not feel like eating; my appetite was poor.	0	1	2	3
3. I felt that I could not shake off the blues even with help from my family or friends.	0	1	2	3
4. I felt that I was just as good as other people.	0	1	2	3
5. I had trouble keeping my mind on what I was doing.	0	1	2	3
6. I felt depressed.	0	1	2	3
7. I felt that everything I did was an effort.	0	1	2	3
8. I felt hopeful about the future.	0	1	2	3
9. I thought my life had been a failure.	0	1	2	3
10. I felt fearful.	0	1	2	3
11. My sleep was restless.	0	1	2	3
12. I was happy.	0	1	2	3
13. I talked less than usual.	0	1	2	3
14. I felt lonely.	0	1	2	3
15. People were unfriendly.	0	1	2	3
16. I enjoyed life.	0	1	2	3
17. I had crying spells.	0	1	2	3
18. I felt sad.	0	1	2	3
19. I felt that people dislike me.	0	1	2	3
20. I could not get "going".	0	1	2	3

Note: Items 4, 8, 12, and 16 were reverse coded.

Figure 7. The CES-D Items.

Finally, Figure 8 contains a display of the reliability of the scale inferences at various levels of the latent variable. Several points are noteworthy. First, the reliability ranges from 0.80 to 0.91 whereas the traditional coefficient alpha for these same items and participants was 0.906. In a sense, the traditional coefficient alpha is a marginal (or average) reliability across the continuum of variation. Secondly, if

one focuses on the range from the 25th to the 95th percentiles, with expected scores between approximately 5 and 15, the reliability is always greater than 0.90. In fact, it is outside of this range that the reliability begins to decline. This last display underscores the importance of taking into account the continuum of variation when conducting item analyses and furthermore highlights that the scale operates optimally in the 25th to 75th percentile range for this sample. Interestingly, the reliability the cut-score of 16 is quite good and in the range of 0.90.

Figure 8 also contains the conditional standard error of measurement. As expected from the formula for the standard error of measurement, the higher the reliability, the smaller the standard error of measurement. The standard error of measurement varies from approximately two to four (out of 60 possible) points across the continuum of variation. Recall that the classical standard error of measurement for this sample was 2.90, like the classical reliability it is akin to an average across the continuum of variation, θ , and this one score would have been reported irrespective of the score obtained by the respondent. Note that, coincidentally, the conditional standard error of measurement at the cut-score of 16 is approximately 2.90, but the standard error of measurement would have been much larger (i.e., 4) if the cut-score had been set at 29.

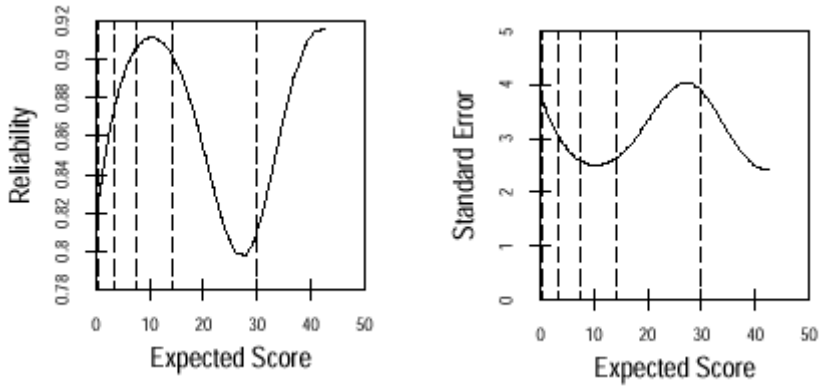

Figure 8. The Conditional Reliability and standard error of measurement (SEM) for the CES-D Plotted Against the Expected Score.

Three further points are noteworthy. First, it is worth reminding the reader that the conditional reliabilities and the conditional standard errors of measurement obtained are dependent on the sample and context in which they are obtained. Second, if one uses the expected score on the abscissa, one may find sharp increases in reliability (and hence decreases in the conditional standard error of measurement) at the extremes of the latent variable but these increases (and decreases) should be ignored because they are an artifact of transforming from the standard normal quantiles to the expected number correct. Figure 9 presents the conditional reliability plot for the same data as in Figure 8 except that it is plotted against the standard normal (i.e., z scores) scores rather than the expected score. Note that the term "Proficiency" in the plot is a legacy of the fact that these techniques were initially developed for achievement tests. The term "proficiency" is more accurately replaced by "the latent variable or latent continuum of variation". Third, the conditional standard error of measurement is reported in the metric of z-scores; therefore, it varies from 0.20 to 0.50 standard deviation units across θ . Recall that the classical standard error of measurement, in z-score units, was 0.307.

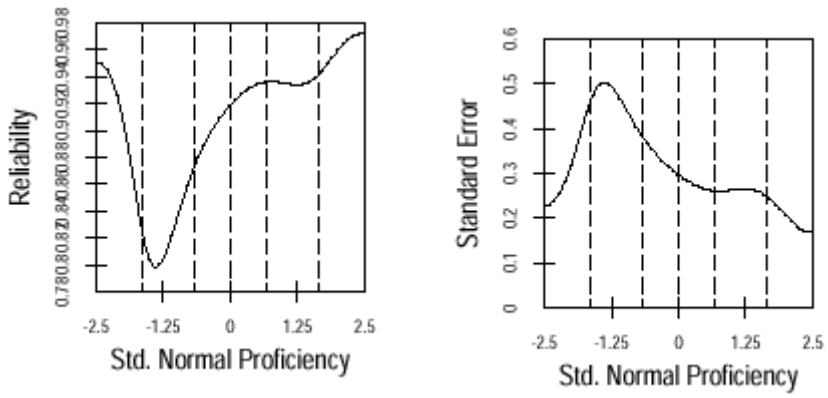

Figure 9. The Conditional Reliability and SEM for the CES-D Plotted Against the Standard Normal Score

To this point, methods have been described for ascertaining: (a) if the items are combined to create one scale score, do they measure just one latent variable? (b) how much of the observed variation is true variation?, and (c) how precisely do the items measure and how does this precision change across the levels of the continuous latent variable?

5. Measurement Decisions

5.1 Sensitivity, Specificity, and Predictive Values

Next, the focus will be on using the scale scores to categorize individuals or groups of individuals based on their responses to the test or measure. Specifically, this presentation will be on methods for determining the accuracy of decisions (i.e., categorizations).

Sensitivity, specificity, and predictive values are all used as evidence of the accuracy or correctness of a decision (i.e., validity) to categorize individuals – in this case, as depressed or not depressed. Before one can calculate these values, one must determine by some means other than the measure of interest, who in the sample is depressed and who is not. When evaluating a self-report scale, this might be determined on the basis of a gold standard such as a Diagnostic and Statistical Manual-IV (DSM-IV) diagnosis. The 2 x 2 table in Figure 10 shows the relationship between these values and how they are calculated.

		SCALE DECISION		
		Not Depressed	Depressed	
GOLD STANDARD	Not Depressed	True Negatives	False Positives	Specificity
	Depressed	False Negatives	True Positives	Sensitivity
		NPV	PPV	

Sensitivity = true positives / (false negatives + true positives)
Specificity = true negatives / (true negatives + false positives)
Positive Predictive Value (PPV) = true positives / (false positives + true positives)
Negative Predictive Value (NPV) = true negatives / (true negatives + false negatives)

Figure 10. A Two-by-Two Table Depicting Possible Measurement Decisions

Sensitivity and specificity tend to be of greater interest to the researcher who is interested in the accuracy of a scale in identifying both depressed and nondepressed individuals. Sensitivity is the percentage of depressed people in the sample that the depression scale correctly identified as depressed. For example, if sensitivity is 85%, then the scale was able to identify correctly 85% of the depressed people in the sample. Specificity is the percentage of nondepressed people in the sample that the depression scale correctly identified as nondepressed. Thus, if specificity is 79%, then the scale was able to identify correctly 79% of the nondepressed people in the sample.

In the case of depression, one may be interested in obtaining maximum sensitivity (as opposed to maximum specificity, or maximum sensitivity and specificity). Highest possible sensitivity levels typically are desired whenever the condition is serious and should not be missed, is treatable, and when false positive results do not lead to serious psychological or economic trauma to the patient or medical system.

Predictive values tend to be of greater interest to the clinician, physician, and patient because they indicate how accurately the scale can predict the presence or absence of depression when it is not known whether the person is depressed. The positive predictive value (PPV) is the percentage of individuals who are truly depressed out of those that the scale identified as depressed. Thus, a PPV of 56% means that only 56% of the people that the scale identified as depressed really were depressed. The negative predictive value (NPV) is the percentage of people who are truly not depressed out of those that the scale identifies as nondepressed. A NPV of 72% means that 72% of the people that the scale identified as not depressed really were not depressed.

PPVs are strongly influenced by both the prevalence of condition (i.e., the number of people per 100,000 who are depressed at the time of the study) and specificity. An examination of the chart in Example 1 shows that as prevalence increases, so does PPV.

Example 1: Sensitivity = 95% & Specificity = 95%

Prevalence (%)	PPV (%)
1	16.1
5	50.0
50	95.0

Small differences in specificity levels can also strongly influence PPVs. A change in specificity from 95% to 99% results in large increases in PPVs as is evident by comparing the PPVs in Example 2 to those in Example 1.

Example 2: Sensitivity = 95% & Specificity = 99%

Prevalence (%)	PPV (%)
1	49.0
5	83.0
50	99.0

However, even large changes in sensitivity have little impact on PPVs. This can be seen by comparing the PPVs in Example 1 to those in Example 3 when sensitivity drops from 95% to 75%.

Example 3: Sensitivity = 75% & Specificity = 95%

Prevalence (%)	PPV (%)
1	13.0
5	43.0
50	94.0

5.2 Receiver Operating Characteristic Curves

Receiver operating characteristic curves (ROC curves) graph the relationship between true positives (sensitivity) and false positives (1-specificity) for all of the possible scores of a depression scale. To demonstrate ROC curves with the example data, a fictitious gold standard is applied and ROC curves are shown using the software SIMSTAT. The results presented should be interpreted only as demonstrating the technique and not a property of the CES-D, per se. Figure 11 is the ROC curve.

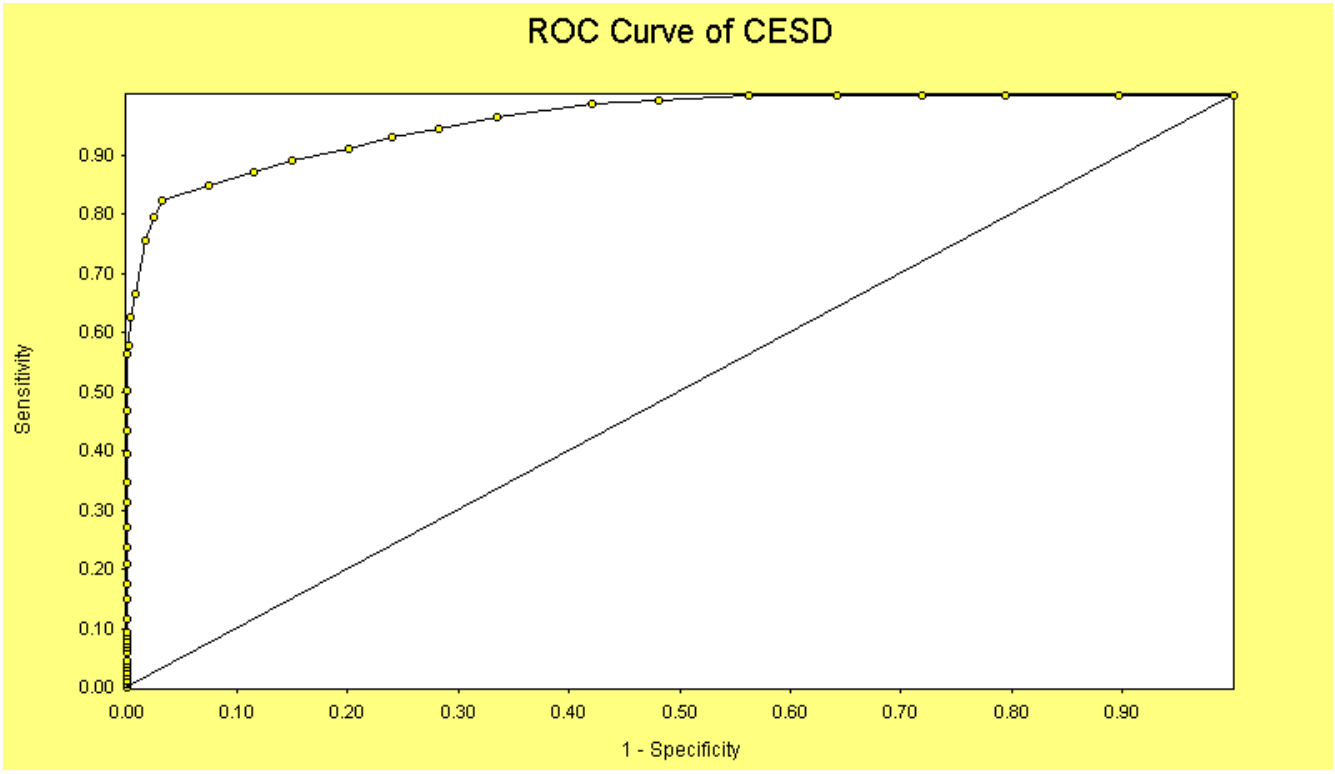

Figure 11. The ROC Curve For The Example Data

The more clearly a scale is able to discriminate between depressed and nondepressed individuals, the farther the curve will deviate from the line of no information (the solid diagonal line in Figure 11) toward the upper left corner of the graph. The area under the curve (AUC) is interpreted as an estimate of the probability that a randomly chosen depressed person will have a higher test score than a randomly chosen nondepressed person. The line of no information has an AUC probability of .50, whereas a perfect test would have an AUC probability of 1.00. Calculating the standard error of the AUC tells one if the AUC for a test is significantly different from the line of no information; that is, does the test provide statistically significantly more information than not administering the test. For example, physicians might want to know if administering a depression scale improves the accuracy of their diagnosis over just observing patients when they come in. In Figure 11, the AUC is 0.9583, which is statistically significantly different from the line of no information.

ROC curves can also be used to compare two scales. For example, assume test A has a sensitivity of 58% and a specificity of 88%, and test B has a sensitivity of 68% and a specificity of 76%. If one wants maximum sensitivity, then test B is better. However, if one wants maximum efficiency (i.e., maximum sensitivity and specificity), then it is difficult to determine which test is better. Calculating the AUC and its standard error allow you to determine if these two tests are significantly different from the line of no information, and if they are significantly different from one another.

6. Validity Theory With An Eye Toward Measurement Practice

Let us now turn to bringing all of the preceding statistical techniques to bear on the quality of a measure, test score validation.

6.1 Evaluating The Measures: Validity And Scale Development

Measurement or test score validation is an ongoing process wherein one provides evidence to support the appropriateness, meaningfulness, and usefulness of the specific inferences made from scores about individuals from a given sample and in a given context. The concept, method, and process of validation are central to constructing and evaluating measures used in the human sciences, for without validation, any inferences made from a measure are potentially meaningless.

The above definition highlights two central features in contemporary thinking about validation. First, it is not the measure per se that is being validated but rather the inferences one makes from a measure. This distinction between the validation of a scale and the validation of the inferences from scores obtained from a scale may appear subtle at first blush but, in fact, it has significant implications for measurement and testing.

The second central feature in the above definition is the clear statement that inferences made from all empirical measures, irrespective of their apparent objectivity, have a need for validation. That is, it matters not whether one is using an observational checklist, an "objective" educational, economic, or health indicator such as number of students finishing grade 12, or a more psychological measure such as a self-report depression measure, one must be concerned with the validity of the inferences.

In recent years, there has been a resurgence of thinking about validity in the field of measurement and testing. This resurgence has been motivated partly by the desire to expand the traditional views of validity to incorporate developments in qualitative methodology and partly by concerns about the consequences of decisions made as a result of the measurement process.

Let us now contrast the traditional and more modern views.

6.2 The Traditional View of Validity

The traditional view of validity focuses on:

- validity as a property of the measurement tool,
- a measure is either valid or invalid,
- various types of validity -- usually four – with the test user, evaluator or researcher typically assuming only one of the four types is needed to have demonstrated validity,
- validity as defined by a set of statistical methodologies, such as correlation with a gold-standard, and
- reliability is a necessary, but not sufficient, condition for validity.

The traditional view of validity can be summarized in the following table:

Type of Validity	What does one do to show this type of validity?
Content	Ask experts if the items (or behaviors) tap the construct of interest.
Criterion-related:	
A. Concurrent	Select a criterion and correlate the measure with the criterion measure obtained in the present
B. Predictive	Select a criterion and correlate the measure with the criterion measure obtained in the future
Construct (A. Convergent and B. Discriminant):	Can be done several different ways. Some common ones are (a) correlate to a "gold standard", (b) factor analysis, (c) multi-trait multi-method approaches

Table 1. The traditional categories of validity

The process of validation then simply becomes picking the most suitable strategy from Table 1 and conducting the statistical analyses. The basis for much validation research has been a correlation with the "gold standard"; this correlation is commonly referred to as a validity coefficient.

6.3 The Expanded (Modern) View Of Validity

The purpose of the more modern view of validity is to expand upon the conceptual framework and power of the traditional view of validity.

Thus, in the expanded, modern conception of validity, the traditional features listed above can be described as follows:

- validity is no longer a property of the measurement tool but rather of the inferences made from the scores,
- validity statements are not dichotomous (valid/invalid) but rather are described on a continuum,
- construct validity is the central most important feature of validity,
- there are no longer various types of validity but rather different sources of evidence that can be gathered to aid in demonstrating the validity of inferences,
- validity is no longer defined by a set of statistical methodologies, such as correlation with a gold-standard but rather by an elaborated theory and supporting methods,
- consequences of test decisions and use (such as unanticipated negative consequences of legitimate test use and/or interpretation that can be traced back to problems such as construct under-representation and/or construct irrelevant variance) are important considerations in the validation process, and
- there is debate as to whether reliability is a necessary but not sufficient condition for validity; it seems that this issue is better cast as one of measurement precision so that one strives to have as little measurement error as possible in their inferences.

In a broad sense, then, validity is about evaluating the inferences made from a measure. All of the methods discussed in the paper (e.g., factor analysis, reliability, item analysis, item response modeling, ROC curves) are directed at building the evidential basis for establishing valid inferences.

Glossary of Terms

Binary scores: The dichotomous scores resulting from item responses. Binary scores arise from true/false or agree/disagree responses. Periodically, an ordinal score is re-coded in a binary format.

Classical test theory: A statistical theory and set of procedures that are based on the notion that an observed score is comprised of a true score and an error score. Procedures commonly associated with classical test theory are reliability, item-total correlations, coefficient alpha, and test-retest correlations. It is called "classical" in reference to its historical precedent.

Confirmatory factor analysis: (CFA) A statistical method generally based on a strong theoretical and/or empirical foundation that allows the researcher to specify an exact factor model for a set of variables in advance.

Construct: A label used to describe a related set of intangible or non-concrete characteristics or qualities in which individual's differ. A label applied to a dimension along which individuals vary.

Continuum of variation: (also called latent variable). A continuum on which individuals vary.

Covariance matrix: A symmetric matrix that has the variances along the major diagonal and the covariances on the off-diagonal elements. A correlation matrix is a standardized covariance matrix.

Covariance structure modeling: A statistical technique that aims to reproduce an observed covariance matrix from a set of relations on observed (and potentially) unobserved (or latent) variables. A powerful statistical method for: (a) testing confirmatory factor analysis models, or (b) taking into account measurement error in modeling.

DIF: DIF occurs when examinees from different groups show differing probabilities of success on (or endorsing) the item *after matching on the underlying ability* that the item is intended to measure.

Essential tau-equivalence: A property of true scores such that one item's true score for one individual is equal to another by an additive constant. As a psychometric framework it also allows for the item error variances to be unequal.

Error score: The difference between the true score and observed score.

Examinees: Those individuals who take a test or complete a questionnaire or scale. We use the terms examinees, test-takers, and respondents synonymously.

Exploratory factor analysis: (EFA) A statistical method used to identify the factor structure or model for a set of variables.

Generalizability theory: An approach to estimating of reliability coefficients that allows one to study various sources of error variance. The theory is founded on the use of analysis of variance to estimate reliability and errors of measurement.

Item bias: Item bias occurs when examinees of one group are less likely to answer an item correctly (or endorse an item) than examinees of another group because of some characteristic of the test item or testing situation that is not relevant to the test purpose. DIF is required, but not sufficient, for item bias.

Item characteristic curve (ICC): A curve showing the probability of a correct response as a function of the trait being measured. A description of the relationship between the continuum of variation on the latent variable and the probability of selecting the keyed (i.e., correct) response.

Item discrimination: The degree to which a test item differentiates between people having various levels of ability or knowledge of the material tested. When an external criterion is used as the basis for evaluating the ability of an item to discriminate along the continuum of variation it is called *item validity*.

Item response modeling: (also called item response theory) A collection of statistical models and methods for making sense out of data obtained in the context of psychological measurement.

Latent variable: (also called latent trait, or unobserved variable) A hypothetical trait or dimension underlying test performance. By using specified statistical models, one can derive *latent trait scores*, which are quantities indicating the test taker's position on this latent continuum of variation.

Measurement error: Inconsistencies in scores from occasion to occasion attributable to the effects of variables operating in a non-systematic (i.e., random) manner.

Ordinal scores: Quantitative scores resulting from item responses. The scores are ordered categories reflecting a quantitative the unobserved (latent) variable. An example are Likert or Likert-like responses.

Random Effects ANOVA: In the context of analysis of variance (ANOVA), the term random effects denotes a factor(s) in an ANOVA design with levels that were not deliberately arranged by the researcher. Factors with levels that are deliberately arranged by the researcher are called fixed effects. Instead, levels of random factors are sampled from a population of possible samples. ANOVAs with fixed effects factors are often called Model I ANOVAs, whereas those with random effects factors are called Model II ANOVAs, and those with both fixed and random effects are called mixed model ANOVAs.

Reliability: The proportion of observed score variance that is true score variance. In practice, this quantity is estimated by computing internal consistency, test-retest, and/or alternate forms reliability coefficients.

Standard error of measurement: An index of the extent to which an individual's scores vary over a number of parallel tests. It is also an index of the amount of measurement error in test scores; theoretically, the standard deviation of the distribution of observed scores around an individual's true score. It is used to estimate the range in which the true score will fall or the amount of change expected on retesting.

Structural equation modeling (SEM): SEM is a comprehensive statistical approach to testing hypotheses about relations among observed and latent variables. This is a commonly used term synonymous with covariance structure modeling.

True score: The expected value (or average) of an individual's observed scores over repeated hypothetical testing of an item or scale.

Bibliography

Byrne B.M. (1998). *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming*. Mahwah, Lawrence Erlbaum Associates. [This book deals with both basic and advanced applications of structural equation modeling, including confirmatory factor analysis]

Crocker L. and Algina J. (1986). *Introduction to Classical & Modern Test Theory*, Fort Worth: Holt, Rinehart, and Winston, Inc. [This book provides an overview of measurement theory]

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. and Strahan, E. J. (1999). Evaluating the Use of Exploratory Factor Analysis in Psychological Research. *Psychological Methods, 4,* 272-299. [An overview article with good guidelines for the practice of exploratory factor analysis]

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*, Thousand Oaks: Sage. [This book provides a thorough description of item response modeling]

Pedhazur E.J. and Schmelkin L.P. (1991). *Measurement, Design, and Analysis: An Integrated Approach.* Hillsdale: Lawrence Erlbaum Associates. [This book provides an integrated discussion of measurement and statistics.]

Ramsay J.O. (2000). *TESTGRAF: A program for the graphical analysis of multiple choice test and questionnaire data.* McGill University. Unpublished computer program manual. [The software and manual for this form of nonparametric item response modeling can be obtained at the following web page http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html]

Shavelson, R. J. and Webb, N. M. (1991). *Generalizability Theory*, Thousand Oaks: Sage. [This book provides a thorough description of generalizability theory]

Traub R.E. (1994). *Reliability for the social sciences: Theory and applications*, Thousand Oaks: Sage. [This book provides a thorough description of classical reliability theory]

Traub, R. E., & Rowley, G. L. (1980). Reliability of Test Scores and Decisions. *Applied Psychological Measurement, 4,*517-545. [This paper describes various methods for establishing reliability in the context of classifying individuals and decision-making]

Zumbo B.D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistical Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense of Canada. [This handbook provides a discussion and methodology for investigating whether item bias is present in your scale. The book and software can be obtained at http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html]

Biographical Sketches

Bruno D. Zumbo, Ph.D. Dr. Zumbo is Professor of measurement, evaluation and research methodology and Associate Member of the Department of Statistics at the University of British Columbia. He is also adjunct professor of Psychology at Simon Fraser University, Senior Research Scholar at the Institute for Social Research and Evalution at the University of Northern British Columbia. His research interests are in psychometric models, statistical science, and the mathematical foundations of measurement.

Anita M. Hubley, Ph.D. Dr. Hubley is a professor of measurement, evaluation and research methodology at the University of British Columbia. Her research interests are in psychometrics and adult development. On aspects of measurement, Dr. Hubley has published on validity theory, neuropsychological testing, and the measurement of depression in the elderly.

Michaela N Gelin, B.A., M.A. Ms. Gelin completed her Masters degree in the Measurment, Evaluation and Research Methodology Program at the University of British Columbia. She also took her Bachelor degree in Psychology and Diploma in Counseling at the University of British Columbia.