THE CONSTRUCTION AND USE OF PSYCHOLOGICAL TESTS AND MEASURES

Bruno D. Zumbo, Michaela N. Gelin and Anita M. Hubley
Measurement, Evaluation, and Research Methodology Program, Department of Educational and Counselling Psychology, and Special Education, 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 ongoing 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. This scale is useful as a demonstration because it is commonly used in the life and social sciences for

both obtaining scores and classifying individual respondents. The latter purpose necessitates methods that help determine the cut-off score for classification (e.g. sensitivity, specificity, and receiver operating characteristic (ROC) curves).

1. Introduction

The topic of this article 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, coverage is by necessity selective and 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 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 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 readers have 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 five-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 multi-category 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 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 epidemiological 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 respondents to reflect upon their 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 0 (not even one day) to 3 (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 interested only in the presence or absence of depressive symptomatology rather than a degree of symptomatology 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 article, 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 general social and policy researchers, although our example focuses on a psychological dysfunction, the methods also apply to scales of opinions and attitudes (e.g. a measure of 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 epidemiological 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 seven 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. Readers should also 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 3 and 4 of this article present methods to answer each of these three questions, respectively, and focus on the descriptive purpose of measurement described above. Section 5 will concern itself with the techniques of the decision-making purpose of measurement. As a whole, this article concerns itself, in its essence, with validation; therefore, the article 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 were 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 **X** is a ($p \times 1$) column vector of scores for person *i* on the *p* items, Λ is a ($p \times 1$) 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 ($p \times 1$) 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:

$$\Sigma = \Lambda\Lambda' + \Psi, \qquad (2)$$

where Σ is the ($p \times p$) population covariance matrix among the items and Ψ is a ($p \times p$) 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 $p \times p$ 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, Σ(θ), for the model. Accordingly, there are various estimation methods such as generalized least-squares or maximum likelihood with their own criteria 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. Σ = Σ(θ)). 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

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

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 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.

-
-
-

Bibliography

Byrne B.M. (1998). *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming*, 412 pp. Mahwah, N.J.: Lawrence Erlbaum. [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 and Modern Test Theory*, 527 pp. New York: Holt, Rinehart, and Winston. [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*, 174 pp. Thousand Oaks, Calif.: 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*, 819 pp. Hillsdale, N.J.: Lawrence Erlbaum. [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 accessed at http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html.]

Shavelson R.J. and Webb N.M. (1991). *Generalizability Theory*, 137 pp. Thousand Oaks, Calif.: Sage. [This book provides a thorough description of generalizability theory.]

Traub R.E. (1994). *Reliability for the Social Sciences: Theory and Applications*, 174 pp. Thousand Oaks, Calif.: Sage. [This book provides a thorough description of classical reliability theory.]

Traub R.E. and 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): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores.* Ottawa: 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 accessed at *http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html*.]

Biographical Sketches

Bruno D. Zumbo, Ph.D. Dr. Zumbo is professor and coordinator of the measurement, evaluation, and research methodology program and associate member of the Department of Statistics and the Department of Psychology at the University of British Columbia, Canada. He is also adjunct professor of psychology at Simon Fraser University, and senior research scholar at the Institute for Social Research and Evaluation 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, Canada. 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. Ms. Gelin is a master’s student in the measurement, evaluation, and research methodology program at the University of British Columbia, Canada. She has completed her bachelor’s degree in psychology and a diploma in counseling at the University of British Columbia.