Psychometrics

Denny Borsboom and Dylan Molenaar, University of Amsterdam, Amsterdam, The Netherlands

This article is a revision of the previous edition article by J.O. Ramsay, volume 18, pp. 12416–12422, © 2001, Elsevier Ltd.

Abstract

Psychometrics is a scientific discipline concerned with the construction of measurement models for psychological data. In these models, a theoretical construct (e.g., intelligence) is systematically coordinated with observables (e.g., IQ scores). This is often done through latent variable models, which represent the construct of interest as a latent variable that acts as the common determinant of a set of test scores. Important psychometric questions include (1) how much information about the latent variable is contained in the data (measurement precision), (2) whether the test scores indeed measure the intended construct (validity), and (3) to what extent the test scores function in the same way in different groups (measurement invariance). Recent developments have focused on extending the basic latent variable model for more complex research designs and on implementing psychometric models in freely available software.

Definition

Psychometrics is a scientific discipline concerned with the question of how psychological constructs (e.g., intelligence, neuroticism, or depression) can be optimally related to observables (e.g., outcomes of psychological tests, genetic profiles, neuroscientific information). This problem is most often approached through the construction of measurement models, in which the construct of interest is represented as a latent variable that acts as a common determinant of a set of observable variables (Bollen, 2002; Sijtsma, 2011). Psychometrics is a highly interdisciplinary field, with connections to statistics, data theory, econometrics, biometrics, measurement theory, and mathematical psychology. Psychometricians may be involved in the design of psychological tests, the formalization of psychological theory, and the construction of data-analytic models. The latter activity has, in the past century, been the main focus of the discipline, especially through the development of test theory and associated latent variable modeling techniques. The main professional organization of psychometricians is the *Psychometric Society*, which publishes the journal *Psychometrika* and organizes the annual *International Meeting of the Psychometric Society*.

History

The birth of psychometrics is usually situated at the end of the nineteenth century, when Francis Galton set up the Antropometric Laboratory with the intention to determine psychological attributes experimentally. Among the first constructs of interest to be subjected to measurement were keenness of sight, color sense, and judgment of eye (Galton, 1884). Galton attempted to measure such attributes through a variety of tasks, recording performance accuracy as well as reaction times. In the early twentieth century, the interest in measuring human qualities intensified greatly when the United States implemented a program to select prospective soldiers using tests (the Army Alpha and Beta) that purported to measure a range of abilities deemed relevant for military performance. Such tests produced a great deal of data, which led to questions that inspired the birth of psychometric theory as we currently know it: how should we analyze psychological test data? Which properties determine the quality of a psychological test? How may we find out whether a test is suited for its purpose?

Two important properties of tests were almost immediately identified (Kelley, 1927). The first referred to the question of whether a test produced *consistent* scores when applied in the same circumstances. This question introduced the notion of *reliability*. In general, a measurement instrument is reliable to the extent that it yields the same outcomes when applied to persons with the same standing of the measured attribute under the same circumstances (e.g., to what extent does an IQ test produce the same score if administered to people with the same level of intelligence?). The second important question is whether the test measures what it purports to measure. This question defines the concept of *validity*. Although different views of validity exist (see later text), the textbook definition is that a measurement instrument is valid if and only if it measures what it should measure (e.g., an IQ test is valid if and only if it actually measures intelligence). The concepts of reliability and validity are still among the most important ones in the evaluation of psychological tests, although they have a very different status in psychometric theory. The definition of reliability is widely accepted, but the definition of validity is widely contested.

The development of reliability theory in the first half of the twentieth century culminated in the work of Lord and Novick (1968), who presented the currently accepted definition of reliability as a signal-to-noise ratio (the ratio of true score variance to observed score variance). The concept may be somewhat differently conceptualized in different theoretical frameworks (e.g., as measurement precision in latent variable theory (Mellenbergh, 1996) and as generalizability in generalizability theory (Cronbach et al., 1972)). However, in such cases Lord and Novick's definition typically follows as a special case, which indicates the consistency of the general psychometric framework. Reliability can, under appropriate assumptions, be estimated in various ways; for instance, from the correlation between two test halves, from the average correlation between test items, and from the correlation

between two administrations of the same test at different times. Strictly speaking, such reliability estimates signify properties of test scores rather than of tests themselves (e.g., administering the test to different populations will ordinarily lead to different values for reliability). Although some discussion may exist about how to optimally estimate reliability, or about which coefficient should be preferred in a given context (Sijtsma, 2009), there is little discussion about how to define reliability theoretically or on how to determine it empirically.

The opposite holds for validity. There are no widely accepted methods to determine whether a test is valid (or to estimate its degree of validity, for theories that conceptualize validity as a matter of degree). In fact, there is no agreement on what validity refers to (Borsboom et al., 2004; Lissitz, 2009; Sijtsma, 2011; Newton, 2012). These disagreements result from the fact that the question of whether a test measures what it purports to measure raises fundamental questions about the nature of psychological constructs themselves. For instance, can psychological constructs literally be measured, in the same way that length and mass are (Michell, 1999)? If not, what notion of measurement should be invoked? Should psychological constructs be given a realist interpretation, or are they merely summaries of data (Borsboom, 2005; Cronbach and Meehl, 1955; Trout, 1999)? Does it make sense to talk about the validity of tests at all, or should we rather talk about the validity of interpretations of test scores or of uses of test scores (Newton, 2012)? Should the societal consequences of test use be counted in the evaluation of validity (Messick, 1989)? Perhaps the only conclusion about validity that is widely accepted is that it is the most problematic among the psychometric concepts.

It is an interesting fact of scientific history that psychometric theory and practice developed largely in the absence of definitive answers to these fundamental questions. Perhaps as a result of the inability to come to grips with the slippery constructs of psychological science, psychometrics instead took its lead from statistics and from general ideas on how measurement should be conceptualized. In doing so, it became a largely technical discipline, which identified as its main task the construction of measurement models for psychological data. Usually, such models contain a 'stand-in' for the psychological construct to be measured (e.g., the expectation of the observed scores – the 'true score' – or a latent variable that is assumed to 'underlie' the responses to test items); they do not, however, contain psychological theory.

Although the detachment of modeling techniques and substance matter is occasionally lamented in psychometric circles, it is undeniable that the statistical models developed along statistical, data-analytic lines define some of the most important contributions of psychometric theory to science in general: classical test theory (Lord and Novick, 1968), latent class analysis (Lazarsfeld and Henry, 1968), the congeneric model (Jöreskog, 1971), the modern test theory models of Rasch (1960) and Birnbaum (1968), and the nonparametric item response models (Mokken, 1971). After these models had been constructed, the development of software to fit and estimate them became one of the main topics of psychometric research, and psychometricians played a leading role in the development of estimation algorithms (e.g., Bock and Aitkin, 1981), model fit tests (Andersen, 1973), software for test analysis (Zimowski et al., 1996; Thissen, 1983; Wilson et al., 1991; Bowler et al., 2007), and general latent variable modeling (e.g., Jöreskog and Sörbom, 1974; Muthén and Muthén, 1998; Bentler, 2000; Arbuckle, 2010; Vermunt and Magidson, 2000). This development took place during the last three decades of the twentieth century, at the end of which most of the basic psychometric models could be estimated and fitted.

Main Concepts of Standard Psychometric Theory

Psychometric measurement models relate a latent structure to a set of observed variables by mapping positions in the latent structure to distributions or densities of the observed variables. This is usually done by specifying the conditional distribution function of the observables, given the latent structure. Thus, the general framework can be thought of in terms of a simultaneous regression of the observed variables on a latent variable or a set of latent variables. Many different models can be derived from this idea through variations on (1) the form of the latent structure (e.g., a continuous line or a set of latent classes), (2) the form of the regression function (e.g., a step function or a logistic function), and (3) the distribution or density appropriate to the observations (e.g., a binomial distribution or a normal density). For instance, if the latent structure is a unidimensional continuum, the regression function is linear, and the observables follow a normal density, the resulting model is the linear common factor model (Jöreskog, 1971); if the latent structure is a unidimensional continuum, the regression function is logistic, and the observables follow a binomial distribution, we get the two-parameter logistic model (Birnbaum, 1968); and if the latent structure is categorical and the observed variables are binary, we obtain the latent class model (Lazarsfeld and Henry, 1968). Mellenbergh (1994b) provides a systematic overview of the connections between these and other models.

The latent structure is generally viewed as a representation of the construct to be measured (e.g., intelligence), while the observed scores typically represent concrete behavioral responses (e.g., answers to items in an IQ test). As a whole, the psychometric model thus coordinates the correspondence between observational and theoretical terms, and in that sense is a measurement model. It is important to note that the notion of measurement intended here is broad and covers the entire spectrum from categorical unordered models to quantitative continuous models. Thus the meaning of the term 'measurement' is not limited to the classical interpretation of the term, which implies quantitative continuous structure (Michell, 1986), although models that are consistent with stricter interpretations may be derived from the general model as a special case (e.g., Rasch, 1960). Thus, the psychometric model is a measurement model in the sense that it coordinates theory with observation, but not in the sense that it assumes that human behavior can be successfully analyzed in terms of quantitative laws.

Measurement precision. Reliability relates to the psychometric model via the concept of measurement precision. The measurement precision of test scores is inversely related to the

variance of the observed scores conditional on a given position in the latent structure (Mellenbergh, 1996). Thus, the higher the variance of the conditional distribution of the observables, the lower the measurement precision of the observables. Note that measurement precision may, but need not, be identical for different positions of the latent structure. For instance, in the linear factor model, which has homoscedastic residuals, measurement precision is identical for all values of the latent variable; in contrast, in the Rasch (1960) model, it is highest for the latent position where the logistic regression of the observable has its inflection point and lower for positions father away from that point; note also that, at the point where this curve has its lowest measurement precision, it yields maximal information. The reliability of test scores, as proposed in classical test theory (Lord and Novick, 1968), is an unconditional index of measurement precision that can be straightforwardly derived from the conditional definition (Mellenbergh, 1996).

Item Response Theory. If the observed variables are responses to test items (e.g., items in an IQ test), then the measurement model falls under the scope of item response theory (IRT), a subfield of psychometrics that has come to play an important role in the analysis of educational tests. In IRT, the function that specifies the regression of the observed variables on the latent structure is known as an item characteristic curve (ICC). Usually, IRT models assume a unidimensional and continuous latent structure, which means the ICC is a smooth curve as in Figure 1. Because items in educational testing are typically scored dichotomously (1: correct, 0: incorrect), the ICC is bounded from above and from below, and hence is often modeled using a nonlinear function. The slope of the ICC at a given point on the latent scale is proportional to the ability of the item to discriminate between positions above and below that point, and thus determines the amount of information that the item yields at that point. Plotting the item information against the latent variable then gives the item information function (IIF).

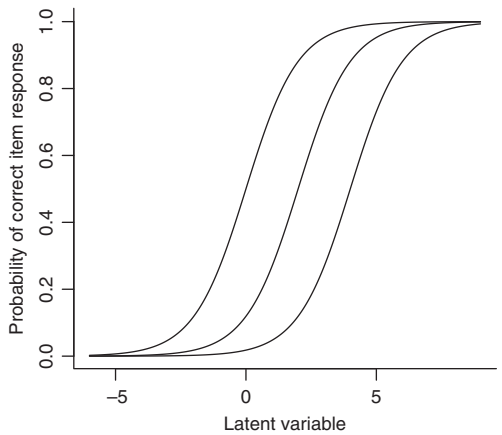

Figure 1 Item response theory relates the probability of a correct response to an item to a continuous latent variable through item characteristic curves (ICCs). When ICCs are parallel, as in this figure, items located further to the right have a lower probability of a correct item response for a given level of the latent variable and are typically interpreted as more difficult.

Adaptive testing. The IIF plays an important role in psychometrics because it can be used to regulate the selection of items. This idea forms the basis for adaptive testing (Van der Linden and Glas, 2000), a strategy of adaptively assembling tests that has become increasingly important with the advent of computerized test administration. In adaptive testing, items are administered sequentially and selected for administration adaptively, i.e., on the basis of the previous item responses of the respondent. This typically works as follows: at each point in the item administration process, examinee ability (the position in the latent structure) is estimated on the basis of the item responses given so far. The next item to be administrated is then chosen on the basis of the slope of the IIF at the estimated examinee ability, so that the administered items yield optimal information. In this way, tests can be shortened without affecting reliability.

Measurement invariance. Test scores are often used to select individuals (e.g., in job selection, student placement). Typically, these selection processes operate on populations that are heterogeneous with respect to background variables like sex and ethnicity. In these cases, considerations of fairness suggest that the test should function in the same way across different subpopulations, in the sense that it should not produce systematic bias in the test scores against a certain group. Such bias may, for instance, arise when an IQ test contains questions that are easier for test takers with a specific background regardless of their intelligence level. This may, for instance, occur if the test contains general knowledge questions that draw on a specific cultural background, so that they are more difficult for ethnic minorities for reasons independent of intelligence level. The concept of measurement invariance (Mellenbergh, 1989; Meredith, 1993) formalizes this idea and allows for testing it against empirical data.

Alternative psychometric models. Although the latent variable model is the primary race horse of current psychometrics, it is not the only available model to connect theoretical terms to observations. An important alternative in the psychometric literature concerns multidimensional scaling (MDS). MDS is a psychometric tool to infer the number of underlying dimensions in proximity data, i.e., data consisting of measures of similarity among pairs of stimuli (e.g., the degree in which different facial expressions are judged to be similar). Depending on whether the measurement scales of the similarity measures are either continuous or ordinal, one speaks of metric MDS (Torgerson, 1952) or nonmetric MDS (Shepard, 1962), respectively. In MDS, individual differences are taken into account by weighting the underlying dimensions differently across subjects (Carroll and Chang, 1970). By doing so, each subject receives a different weight for each dimension indicating which dimensions are more important for a subject in deciding which stimuli are similar. These weights enable identification of different subtypes of subjects. An important instance of MDS with individual differences is unfolding analyses, which is suitable for preference data. In unfolding analyses (Coombs, 1964), each subject is assumed to have an ideal point on the dimensions underlying the preference data. When a given stimulus is close to the subjects ideal point, that stimulus is preferred more.

In addition to MDS, other possibilities to make the connection between theory and data are by (1) representing the

construct as a common effect of the observed variables (Bollen and Lennox, 1991), (2) taking the construct to be a universe of behaviors, of which the observables are assumed to be a sample (Cronbach et al., 1972; McDonald, 1999), and (3) interpreting the construct as a causal system in which the observables influence each other (Cramer et al., 2010).

Recent Developments

The past couple of years, advances in psychometrics have mainly focused on extensions of the traditional measurement models to deal with more complex situations. Important extensions include the incorporation and development of multilevel and random effects structures in IRT models (Fox and Glas, 2001), factor models (Rabe-Hesketh et al., 2004), and latent class models (Lenk and DeSarbo, 2000). In these models, item parameters – considered fixed effects in the traditional measurement models – may themselves become random variables. This enables psychometric analyses of nested data, which are common in large-scale assessments. These extended models have been formulated within integrated frameworks that include the different models as special cases (Moustaki and Knott, 2000; Skrondal and Rabe-Hesketh, 2004).

Other extensions were motivated by the increasing popularity of computerized test administration. This causes response times to become available to the researcher in addition to the ordinary responses. Various suitable models have been proposed including models in which response times are modeled as collateral information (Van der Linden, 2007, 2009; Van Breukelen, 2005) or as manifestation of the underlying decision process (Tuerlinckx and De Boeck, 2005; Van der Maas et al., 2011).

Furthermore, recent advances in computer technology enabled more refined estimation techniques and model fit assessment, particularly for multidimensional IRT models (Béguin and Glas, 2001; Reckase, 2009), factor analyses of non-normal and categorical data (Satorra and Bentler, 2001; Molenaar et al., 2012), cognitive diagnosis models (Junker and Sijtsma, 2001; De la Torre and Douglas, 2004), unfolding analyses (Busing et al., 2005), and nonlinear factor models (Klein and Moosbrugger, 2000; Lee and Zhu, 2002).

On the conceptual side, three discussions have dominated the psychometric literature of the past decade. One concerns the status of psychometric measurement models and the relation between psychometrics and psychology (Michell, 1999; Borsboom, 2006). A second important discussion concerns the usefulness of Cronbach's alpha as a measure of reliability, which has been forcefully contested (Zinbarg et al., 2005; Sijtsma, 2009). Third, the proper definition of validity remains a contentious topic that continues to generate debate (Lissitz, 2009; Newton, 2012).

With respect to psychometric software, recent developments have mainly been inspired by the rise of the open source statistical software program R (R Development Core Team, 2012). This program enables psychometricians to develop their own model estimation packages and share these with other researchers. This trend is illustrated by two recent special issues of the *Journal of Statistical Software* (2007, 2012) that focused exclusively on psychometrics in R. Among the most popular R packages for psychometric analyses are *ltm* (Rizopoulos, 2006) and *eRm* (Mair and Hatzinger, 2007), which can fit various IRT models to data; *mokken* (Van der Ark, 2012), which may be used to fit nonparameteric IRT models to data; *sem* (Fox, 2006) and *lavaan* (Rosseel, 2012), which enables structural equation modeling; *smacof* which involves MDS (De Leeuw and Mair, 2009); and *qgraph* (Epskamp et al., 2012), which produces network visualizations of psychometric data and models. In addition, the *OpenMx* package (Boker et al., 2010) is a more general R package that can be used to fit various (extensions of) parametric models, including advanced possibilities like mixtures and joint modeling of discrete and continuous data.

See also: Classical (Psychometric) Test Theory; Reliability: Measurement; Selection Bias, Statistics of.

Bibliography

Andersen, E.B., 1973. A goodness of fit test for the Rasch model. Psychometrika 38 (1), 123–140.
Arbuckle, J.L., 2010. IBM SPSS Amos 19 User's Guide. SPSS, Chicago, IL.
Béguin, A.A., Glas, C.A.W., 2001. MCMC estimation and some model-fit analysis of multidimensional IRT models. Psychometrika 66 (4), 541–561.
Bentler, P.M., 2000. EQS 6 Structural Equations Program Manual. Multivariate Software Inc., Encino, CA.
Birnbaum, A., 1968. Some latent trait models and their use in inferring an examinee's ability. In: Lord, F.M., Novick, M.R. (Eds.), Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA, pp. 397–479.
Bock, R.D., Aitkin, M., 1981. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika 46 (4), 443–459.
Bollen, K.A., 2002. Latent variables in psychology and the social sciences. Annual Review of Psychology 53 (1), 605–634.
Bollen, K.A., Lennox, R., 1991. Conventional wisdom on measurement: a structural equation perspective. Psychological Bulletin 110 (2), 305–314.
Boker, S., Neale, M.C., Maes, H.H., Wilde, M., Spiegel, M., Brick, T., et al., 2010. OpenMx: an open source extended structural equation modeling framework. Psychometrika 76 (2), 306–317.
Borsboom, D., 2005. Measuring the Mind: Conceptual Issues in Contemporary Psychometrics. Cambridge University Press, Cambridge.
Borsboom, D., 2006. Attack of the psychometricians. Psychometrika 71 (3), 425–440.
Borsboom, D., Mellenbergh, G.J., Van Heerden, J., 2004. The concept of validity. Psychological Review 111 (4), 1061–1071.
Bowler, D.R., Miyazaki, T., Gillan, M.J., Ohno, T., Brázdová, V., et al., 2007. CONQUEST β. 1: User manual.
Busing, F.M.T.A., Groenen, P.J.F., Heiser, W., 2005. Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. Psychometrika 70, 71–98.
Carroll, J.D., Chang, J.J., 1970. Individual differences and multidimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrika 35, 282–319.
Coombs, C., 1964. A Theory of Data. Wiley, New York.
Cramer, A.O.J., Waldorp, L.J., van der Maas, H., Borsboom, D., 2010. Comorbidity: a network perspective. Behavioral and Brain Sciences 33 (2–3), 137–193.
Cronbach, L.J., Gleser, G.C., Nanda, H., Rajaratnam, N., 1972. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. Wiley, New York.
Cronbach, L.J., Meehl, P.E., 1955. Construct validity in psychological tests. Psychological Bulletin 52 (4), 281–302.
De la Torre, J., Douglas, J.A., 2004. Higher order latent trait models for cognitive diagnosis. Psychometrika 69 (3), 333–353.
De Leeuw, J., Mair, P., 2009. Multidimensional scaling using majorization: SMACOF in R. Journal of Statistical Software 31 (3), 1–30.

Epskamp, S., Cramer, A.O.J., Waldorp, L.J., Schmittmann, V.D., Borsboom, D., 2012. qgraph: network visualizations of relationships in psychometric data. Journal of Statistical Software 48 (4), 1–18.

Fox, J., 2006. Teacher's corner: structural equation modeling with the sem package in R. Structural Equation Modeling 14, 465–486.

Fox, J.-P., Glas, C.A.W., 2001. Bayesian estimation of a multilevel IRT model using Gibbs sampling. Psychometrika 66 (2), 271–288.

Galton, F., 1884. Measurement of character. Fortnightly Review 36, 179–185.

Jöreskog, K.G., 1971. Statistical analysis of sets of congeneric tests. Psychometrika 36 (2), 109–133.

Jöreskog, K.G., Sörbom, D., 1974. LISREL III. Scientific Software International, Inc., Chicago, IL.

Junker, B.W., Sijtsma, K., 2001. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. Applied Psychological Measurement 25 (3), 258–272.

Kelley, T.L., 1927. Interpretation of Educational Measurements. Macmillan, New York.

Klein, A., Moosbrugger, H., 2000. Maximum likelihood estimation of latent interaction effects with the LMS method. Psychometrika 65 (4), 457–474.

Lazarsfeld, P.F., Henry, N.W., 1968. Latent Structure Analysis. Houghton Mifflin, Boston, MA.

Lee, S.-Y., Zhu, H.-T., 2002. Maximum likelihood estimation of nonlinear structural equation models. Psychometrika 67 (2), 189–210.

Lenk, P., DeSarbo, W., 2000. Bayesian inference for finite mixtures of generalized linear models with random effects. Psychometrika 65 (1), 93–119.

Lissitz, R.W., 2009. The Concept of Validity: Revisions, New Directions, and Applications. Information Age, Charlotte, NC.

Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA.

Mair, P., Hatzinger, R., 2007. Extended Rasch modeling: the eRm package for the application of IRT models in R. Journal of Statistical Software 20 (9), 1–20.

McDonald, R.P., 1999. Test Theory: A Unified Treatment. Lawrence Erlbaum Associates, Inc., Mahwah, NJ.

Mellenbergh, G.J., 1989. Item bias and item response theory. International Journal of Educational Research 13 (2), 127–143.

Mellenbergh, G.J., 1994a. A unidimensional latent trait model for continuous item responses. Multivariate Behavioral Research 29 (3), 223–236.

Mellenbergh, G.J., 1994b. Generalized linear item response theory. Psychological Bulletin 115 (2), 300–307.

Mellenbergh, G.J., 1996. Measurement precision in test score and item response models. Psychological Methods 1 (3), 293–299.

Meredith, W., 1993. Measurement invariance, factor analysis, and factorial invariance. Psychometrika 58 (4), 525–543.

Messick, S., 1989. Validity. In: Linn, R.L. (Ed.), Educational Measurement. American Council on Education and National Council on Measurement in Education, Washington, DC, pp. 13–103.

Michell, J., 1986. Measurement scales and statistics: a clash of paradigms. Psychological Bulletin 100 (3), 398–407.

Michell, J., 1999. Measurement in Psychology: A Critical History of a Methodological Concept. Cambridge University Press, Cambridge.

Mokken, R.J., 1971. A Theory and Procedure of Scale Analysis. Mouton, The Hague.

Molenaar, D., Dolan, C.V., de Boeck, P., 2012. The heteroscedastic graded response model with a skewed latent trait: testing statistical and substantive hypotheses related to skewed item category functions. Psychometrika 77, 455–478.

Moustaki, I., Knott, M., 2000. Generalized latent trait models. Psychometrika 65 (3), 391–411.

Muthén, L.K., Muthén, B.O., 1998. Mplus User's Guide. Authors, Los Angeles.

Newton, P.E., 2012. Clarifying the consensus definition of validity. Measurement: Interdisciplinary Research & Perspective 10 (1–2), 1–29.

R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.Rproject.org.

Rabe-Hesketh, S., Skrondal, A., Pickles, A., 2004. Generalized multilevel structural equation modelling. Psychometrika 69 (2), 167–190.

Rasch, G., 1960. Probabilistic Models for Some Intelligence and Attainment Tests. Paedagogiske Institut, Copenhagen.

Reckase, M.D., 2009. Multidimensional Item Response Theory. Springer, London.

Rizopoulos, D., 2006. ltm: An R package for latent variable modeling and item response theory analyses. Journal of Statistical Software 17 (5), 1–25.

Rosseel, Y., 2012. lavaan: An R package for structural equation modeling. Journal of Statistical Software 48 (2), 1–36.

Satorra, A., Bentler, P.M., 2001. A scaled difference chi-square test statistic for moment structure analysis. Psychometrika 66 (4), 507–514.

Sijtsma, K., 2009. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika 74 (1), 107–120.

Sijtsma, K., 2011. Introduction to the measurement of psychological attributes. Measurement. Journal of the International Measurement Confederation 44 (7), 1209–1219.

Shepard, R.N., 1962. Analysis of proximities: multidimensional scaling with an unknown distance function, I. Psychometrika 27, 125–140.

Skrondal, A., Rabe-Hesketh, S., 2004. Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Chapman & Hall/CRC, Boca Raton, FL.

Thissen, D., 1983. MULTILOG: Multiple Category Item Analysis and Test Scoring Using Item Response Theory [Computer Software]. Scientific Software International, Chicago, IL.

Tuerlinckx, F., De Boeck, P., 2005. Two interpretations of the discrimination parameter. Psychometrika 70 (4), 629–650.

Trout, J.D., 1999. Measurement. In: Newton-Smith, W.H. (Ed.), A Companion to the Philosophy of Science. Blackwell, Oxford, pp. 265–276.

Torgerson, W.S., 1952. Multidimensional scaling: I. Theory and method. Psychometrika 17, 401–409.

Van der Ark, L.A., 2012. New developments in mokken scale analysis in R. Journal of Statistical Software 48 (5), 1–27.

Van Breukelen, G.J.P., 2005. Psychometric modeling of response speed and accuracy with mixed and conditional regression. Psychometrika 70 (2), 359–376.

Van der Linden, W.J., 2007. A hierarchical framework for modeling speed and accuracy on test items. Psychometrika 72 (3), 287–308.

Van der Linden, W.J., 2009. Conceptual issues in response-time modeling. Journal of Educational Measurement 46 (3), 247–272.

Van der Linden, W.J., Glas, C.A.W. (Eds.), 2000. Computerized Adaptive Testing: Theory and Practice. Kluwer, Norwell, MA.

Van der Maas, H.L.J., Molenaar, D., Maris, G., Kievit, R.A., Borsboom, D., 2011. Cognitive psychology meets psychometric theory: on the relation between process models for decision making and latent variable models for individual differences. Psychological Review 118 (2), 339–356.

Vermunt, J.K., Magidson, J., 2000. Latent Gold: User's Manual. Statistical Innovations Inc., Boston.

Wilson, D.T., Wood, R., Gibbons, R., 1991. TESTFACT: Test Scoring, Item Statistics, and Item Factor Analysis [Computer Software]. Scientific Software International, Chicago.

Zimowski, M.F., Muraki, E., Mislevy, R.J., Bock, R.D., 1996. BILOG-MG: Multiple-group IRT Analysis and Test Maintenance for Binary Items [Computer Software]. Scientific Software International, Chicago.

Zinbarg, R.E., Revelle, W., Yovel, I., Li, W., 2005. Cronbach's α, Revelle's β, and McDonald's ωH: their relationships with each other and two alternative conceptualizations of reliability. Psychometrika 70 (1), 1–11.