Psychological Test and Assessment Modeling, Volume 52, 2010 (3), 272-297

Scale construction and evaluation in practice: A review of factor analysis versus item response theory applications

Janke C. ten Holt[1], Marijtje A. J. van Duijn[2] & Anne Boomsma[2]

Abstract

In scale construction and evaluation, factor analysis (FA) and item response theory (IRT) are two methods frequently used to determine whether a set of items reliably measures a latent variable. In a review of 41 published studies we examined which methodology – FA or IRT – was used, and what researchers' motivations were for applying either method. Characteristics of the studies were compared to gain more insight into the practice of scale analysis. Findings indicate that FA is applied far more often than IRT. Many times it is unclear whether the data justify the chosen method because model assumptions are neglected. We recommended that researchers (a) use substantive knowledge about the items to their advantage by more frequently employing confirmatory techniques, as well as adding item content and interpretability of factors to the criteria in model evaluation; and (b) investigate model assumptions and report corresponding findings. To this end, we recommend more collaboration between substantive researchers and statisticians/psychometricians.

Key words: factor analysis, item response theory, test construction, scale evaluation

[1] *Correspondence concerning this article should be addressed to:* Janke C. ten Holt, PhD, Grote Rozenstraat 31, 9712 TG Groningen, The Netherlands; email: j.c.ten.holt@rug.nl

[2] Department of Sociology, University of Groningen, The Netherlands

In the process of scale construction and evaluation, statistical modeling is used to assess the extent to which a group of items can be considered to measure the latent variable(s) researchers are interested in. Factor analysis (FA) and item response theory (IRT) are two types of models used for scale analysis. In the present study we aim to evaluate the use of FA and IRT for the construction and evaluation of scales and tests in practice. Of primary interest is the researchers' motivation for choosing either methodology. It is of secondary (methodological) interest to investigate how the chosen analysis is performed and what results are reported.

The theoretical relationship between FA and IRT has been well documented (Takane & De Leeuw, 1987; see also Kamata & Bauer, 2008, and Mehta & Taylor, 2006), demonstrating that certain variants of FA and IRT are equivalent, thus enabling the computation of FA model parameters from IRT parameters and vice versa (see Brown, 2006, p. 398 ff., for a comprehensive demonstration). Furthermore, in past research, the results of FA and IRT have been compared on simulated data (e.g., Knol & Berger, 1991; Wirth & Edwards, 2007), on empirical data (e.g., Glöckner-Rist & Hoijtink, 2003; Moustaki, Jöreskog, & Mavridis, 2004), or on both simulated and empirical data (e.g., Jöreskog & Moustaki, 2001). FA and IRT have also been compared in their usefulness for investigations of measurement equivalence (e.g., Meade & Lautenschlager, 2004; Raju, Laffitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993). An overview of these studies will not be given here. Instead, the present study focuses on what is done *in the practice* of scale development: What method is used, why is it used, and how is it used? We thus provide a descriptive overview of the current status of scale construction and evaluation. We will also consider whether and where there is room for improvement.

We conducted a review of a sample of published articles. We first searched for journals in the fields of psychology and education that contain many articles reporting on the construction or evaluation of a scale as the main topic. *Psychological Assessment* (PA), *European Journal of Psychological Assessment* (EJPA), and *Educational and Psychological Measurement* (EPM) met this criterion. In these journals, all articles concerning scale construction or evaluation published in 2005 were selected for review, yielding a total of 46 articles. In the majority of articles, either a FA or an IRT analysis was conducted. Six articles that contained other analyses were excluded from this review, due to their lack of relevance for the present comparison. Three of these articles concerned a reliability generalization analysis (cf. Vacha-Haase, 1998). In the other three articles, a classical test theory (CTT) analysis, a generalizability analysis, and a multidimensional scaling analysis were performed, respectively. Of the remaining 40 articles, one contained separate analyses of two distinct scales, both of which were included in this study. Hence, a total of 41 studies were selected for consideration.

First, we describe how often each model is applied. Second, explicit motivation given for model choice is discussed. Third, an attempt is made to reveal implicit motives by examining various characteristics of the data and the models. Fourth, we review how the statistical analyses were performed and reported upon. Finally, the findings of our study are summarized and discussed.

Applied model

FA, IRT, and both FA and IRT are applied in 32 (78 %), six (15 %), and three (7 %) studies, respectively, illustrating the dominance of FA in the practice of scale analysis. These and all percentages mentioned hereafter, refer to all 41 studies included in the review.

In four FA studies, the applied model has an equivalent IRT counterpart. In these studies, the model is estimated using polychoric correlations, which is equivalent (cf. Takane & De Leeuw, 1987) to using the 2-parameter normal-ogive IRT model (Birnbaum, 1968), or, for polytomous items, the graded response model (Samejima, 1969), either in the normal-ogive or the logistic form; the latter with a scaling constant for approximating the normal distribution function by the logistic (cf. Molenaar, 1974). For none of the models applied in the IRT studies does an FA equivalent exist. Of the three studies where both FA and IRT are applied, there are two cases that use equivalent models. In one case (Vigneau & Bors, 2005), the Rasch IRT model is applied to dichotomous items. This is equivalent (except for the logistic/normal approximation) to applying a factor model to the matrix of tetrachoric correlations, and restricting all loadings to be equal. In the other case (Wang & Russell, 2005), the graded response IRT model is applied.

In the remainder of the paper, we distinguish between FA and IRT in accordance with the authors' terminology. We do so to emphasize the practice as it is presented by the researchers, and because there are only a few studies with equivalent FA and IRT models (six of the 41 studies). Table 1 provides an overview of a number of aspects of the studies that are discussed.

Explicit motives for model choice

The first step in gaining information about researchers' motives for applying a certain model is to simply record what investigators themselves say about their motivation. Unfortunately, that is not much.

We distinguish studies where some motives are given for the selection of a (sub)model from studies where no motives are given at all. In addition, we distinguish studies where the model choice is discussed in detail, mentioning both FA and IRT, our primary interest.

In 24 studies (59 %), no motives whatever for model choice are given. In 14 studies (34 %), some motives are given concerning the choice of the model. In four of the six IRT studies, the benefits of IRT over CTT are described. In some FA studies, the choice of exploratory FA (EFA) versus confirmatory FA (CFA) is defended. Arguments provided in favor of EFA are: "A small items-to-subjects ratio," "not expecting to replicate a factor structure," and "the absence of a previous factor-analytically derived factor structure." Arguments provided in favor of CFA are: "the need for an in-depth analysis of the hypothesized factor structure of the scale," "the possibility of testing a theoretical model," and of "testing competing models by means of comparative fit indices."

In three of the reviewed studies (7 %), the model choice is motivated, mentioning both FA and IRT. In two of these (Vigneau & Bors, 2005; Wang & Russell, 2005) both FA

and IRT are applied, in the other one (Hong & Wong, 2005) only IRT is applied. Vigneau and Bors (2005) and Hong and Wong (2005) both mention a skewed item response distribution as an argument for applying the Rasch model, because it is "insensitive to the shape of the item distributions," whereas in standard (i.e., linear) FA, items are assumed to have a multivariate normal distribution if maximum likelihood (ML) estimation is employed.

Because studies where both FA and IRT are applied are of particular interest for the present comparison, we briefly describe each of them. Vigneau and Bors (2005) do not state clearly why they use *both* FA and IRT. It is mentioned that "the IRT model is better suited for the analysis of the dichotomous data." They do not explain why FA is also performed, but perhaps they also want results comparable to previous studies. They perform three separate factor analyses: on product-moment, tetrachoric, and corrected phi-correlations (phi/phi_{max}), which, as they note, have all been used before for the scale under investigation. In this study, based on the FA results, it cannot be decided whether the data are one- or two-dimensional, whereas the IRT analysis indicates that a unidimensional model does not describe the data well.

Clark, Antony, Beck, Swinson, and Steer (2005) apply IRT at an exploratory stage of scale construction. They determine item discrimination at various levels of the latent variable by graphically examining item response functions. Although not explicitly mentioned, from the references given it can be deduced that Ramsay's (2000) Testgraf model is used here. The structure of the scale is investigated with a principal component analysis. Clark et al. do not elaborate on their reasons for applying both FA and IRT.

Wang and Russell (2005) perform a differential item functioning (DIF) analysis, i.e., an inspection of whether items function equivalently in different populations of respondents. They describe FA and IRT as "complementary approaches," with IRT better suited for testing equivalence of item parameters (see also Meade & Lautenschlager, 2004), and FA more appropriate for multidimensional model testing. It is remarkable that the CFA is applied on product-moment correlations rather than polychoric correlations, since the latter would be equivalent to the graded response IRT model that was used. In fact, this relationship is never mentioned in the study.

Since model choice is not explicitly motivated in the majority of the studies, we next discuss a number of study characteristics – including aims, some descriptive statistics, and software use – in the hope of revealing some implicit motivations.

Characteristics of the data and applied models

Comparison of characteristics of FA and IRT studies

We classify the 41 studies in our analysis into five types, based on their primary aims: evaluation, new scale, translation, DIF, and short form. In 18 of the studies (44 %), an existing scale is evaluated. The focus of these studies is usually on the latent variable structure of the data, examining which items are substantially associated with each other,

indicating that they measure the same construct. Another interest of these evaluation studies is to estimate how reliably the items measure the latent variable.

In 10 of the studies (24 %), a new scale is constructed. Researchers in these studies often report the process of writing a large number of items, followed by a systematic reduction of the item set in a number of steps, one of which is a psychometric evaluation by means of FA and/or IRT.

In eight of the studies (20 %), a scale is translated and the psychometric properties of this translated scale are analyzed. In three studies (7 %), a DIF analysis is performed to investigate whether items in the scale are responded to differentially by distinct groups of participants. Finally, in two studies (5 %), a short form of an existing scale is constructed and analyzed, with the goal of creating a compact version of the scale consisting of only a small number of items.

From Table 1 it can be seen that the type of study is not clearly related to the type of analysis being performed: The relative frequency of the application of FA and IRT is the same for new scale, evaluation, and translation studies. One could argue that in DIF and short-form studies, IRT is applied more often, but these types of studies occurred too infrequently in this sample to draw any general conclusions.

The number of item categories varies from two to eight. As is apparent from Table 1, 5-point scales are most popular for FA studies, but other numbers of categories are also common. IRT studies and dichotomous items are not strongly associated, contrary to what might have been expected. In one study, items vary in the number of categories, and in two studies no information about item categories is provided.

Thirty-four of our studies (83 %) consider multidimensional scales, ranging from two to 15 dimensions. In six of these, multiple models with varying numbers of dimensions are tested. Seven studies (17 %) consider unidimensional scales. These include one FA study and five of the six IRT studies. It seems that, in practice, IRT is primarily applied to investigate unidimensional scales.

The ratio of number of items to number of factors varies between 4 and 36 with a median of 8. In most studies, each factor is represented by five to 15 items. This number is well above the recommended minimum of four or five items per factor for small samples (Marsh & Hau, 1999; Marsh, Hau, Balla, & Grayson, 1998). In the IRT studies, the item/factor ratio is larger than in the FA studies. A confounding factor could be the number of dimensions in the model, as a smaller number of factors with a fixed number of items increases the item/factor ratio.

The sample sizes in the studies vary between 118 and 9,160 with a median of 553. There are no noticeable differences between FA and IRT studies here, other than that more extreme values are encountered in FA studies, but this could be a result of the greater number of FA studies in the sample of articles. Because of the large variation in sample size and the limited number of studies, it is not possible to make any generalizations beyond the reviewed studies about the differences in sample size between FA and IRT studies.

Table 1:
Overview of Study Characteristics

	Type of applied analysis		
Characteristic	FA (n = 32)	IRT (n = 6)	FA & IRT (n = 3)
Motives provided			
no	24		
some	8	5	1
explicit		1	2
Type of study			
Evaluation	15	2	1
New scale	8	1	1
Translation	7	1	
DIF	1	1	1
Short form	1	1	
# Item categories			
2	3	1	1
3	3		1
4	5	2	1
5	8	1	
6–8	10	1	
varying[a]		1	
no info	3		
# Dimensions			
1	1	5	1
2	4		1
3	8		
4	4		
5	3		1
6–15	6	1	
varying[b]	6		
Item/factor ratio			
median	7.3	18.0	12.6
(MAD)[c]	(1.67)	(3.00)	(0.10)
Sample size			
min.	118	205	506
max.	9160	4306	2151
median	577	982	512
(MAD)	(368)	(740)	(6)
Respondent/item ratio			
median	17.9	41.9	20.5
(MAD)	(10.16)	(33.10)	(6.42)
Exploratory vs. confirmatory			
expl.	8	2	
conf.	13	4	1
both	11		2

Note. Numbers in the table represent frequencies of studies, except for the row entries min., max., median, and MAD.
[a]A scale consisting of items that differ in the number of categories. [b]Models with various numbers of dimensions are tested. [c]MAD: median absolute deviation from the median.

The ratio of number of respondents to number of items varies between 4.6 and 1,077 with a median of 18.6. In most studies, there are about 20 respondents per item. This number surpasses the ratios of 5 or 10, recommended as lower bounds in the literature (Bentler, 1989; Mueller, 1996; Nunnally, 1978). It should be noted, though, that such guidelines are too simple. As Brown (2006, p. 413 ff.) notes, many more characteristics of the data and the model should be taken into account to determine a sufficient number of respondents for proper inference. The number of estimated parameters, to name one, is greater for IRT models than for standard factor models, the former thus requiring more respondents. Brown suggests choosing a sample size by conducting a power analysis, using either the method proposed by Satorra and Saris (1985) or a Monte Carlo method, in both cases selecting the sample size associated with an 80 % likelihood of rejecting a false null hypothesis for the specified model.

In 10 studies (24 %), the applied analysis is exploratory; a confirmatory analysis is reported in 18 studies (44 %); and in 13 studies (32 %), a combination of exploratory and confirmatory analyses is applied. As can be seen from Table 1, there are no noticeable differences between FA and IRT studies here.

The software used for scale analysis, as reported in the studies, is shown in Table 2. For EFA, either general statistical software (`sas, spss, statview, systat`) is used (four studies) or no information is provided (15 studies), presumably also indicating the use of general software. CFA and IRT analyses are almost always accomplished using specialized software, and information about the software used is almost always provided. For CFA, `lisrel` (Jöreskog & Sörbom, 1996) is the most popular, but other packages such as `amos` (Arbuckle, 1995–2006), `eqs` (Bentler, 1995), and `mplus` (L. K. Muthén & Muthén, 1998–2010) are also used. Researchers who apply IRT use a wide variety of software. In each study, a different program is employed, except for the Mokken scaling program (`msp`), which is used twice.

The use of software shows that EFA is more accessible to researchers than CFA or IRT, since it can be carried out with general statistical software. One does not have to obtain and learn how to use a new computer program to do EFA.

It is remarkable that for 17 analyses, no information is provided on the software being used. This is against the policy of the American Psychological Association [APA] (2001, p. 280; 2010, p. 210): Although reference entries are not necessary for standard software and programming languages such as `sas` and `spss`, the proper name of the software and the version number should always be reported in the text.

Table 2:
Overview of Software Use

	Type of applied analysis			
	FA (n = 32)		IRT	FA & IRT
Software	EFA	CFA	(n = 6)	(n = 3)
lisrel		12		1
amos		4		
eqs		2		
mplus		2		
sca		1		
noharm				1
msp			2	
rsp				1
testgraf			1	
multilog				1
parscale			1	
winsteps			1	
poly-sibtest			1	
equate				1
dfitps6				1
sas	1	1		1
spss	1			
statview	1			
systat	1			
No information	15	2		1

Note. Numbers in the table represent frequencies of studies. In some studies, multiple software packages are used to estimate different kinds of models.

Types of studies

Various goals of the analyses are mentioned in the studies. They include the examination of the psychometric properties, the factor structure, and group differences for a scale. In three studies (all IRT), researchers express interest in the values of item parameters (such as item difficulty and item discrimination).

When the only analysis is CFA, researchers always (in all 13 studies) mention the goal: testing whether the data fit a hypothesized structure. This objective is also mentioned in two of the eight EFA studies. Additionally, in some studies, the choice of specific details of an analysis, e.g., the estimation method or a multilevel approach, is defended.

In Table 3 some descriptive statistics are given for the three most prevalent types of studies: new scale development, evaluation of a scale, and translation of a scale. Not surprisingly, researchers constructing a new scale more often apply an exploratory

Table 3:
Selection of Characteristics for the Three Most Prevalent Study Types

	Type of study		
Characteristic	New scale (n = 10)	Evaluation (n = 18)	Translation (n = 8)
Cross-validation			
no	4	13	5
yes	6	5	3
Exploratory vs. confirmatory			
expl.	5	3	3
conf.	2	11	3
both	3	4	2
Sample size			
median	462	680	386
(MAD)	(302)	(426)	(81)

Note. Numbers in the table represent frequencies of studies, except for the row entries median and MAD.

model, whereas researchers evaluating an existing scale more often apply a confirmatory model. However, most, if not all, new scales are developed on a theoretical basis, which would make a confirmatory analysis a reasonable choice. In most studies reporting an exploratory analysis, researchers do have clear hypotheses about the structure of the data.

It is not clear why many researchers would rather carry out an exploratory analysis than test a hypothesized structure based on substantive knowledge about the items and their mutual relationships, even though the latter approach is far more powerful. Perhaps worries about the presence of secondary factors lure them into an exploratory examination of the factor structure of their data. There is debate about this issue, with some experts advocating a more liberal application of exploratory techniques (e.g., Bandalos & Finney, 2010). However, we argue that if theory about the underlying structure is available, it is preferable to use it and suboptimal to neglect that knowledge.

A second reason for applying exploratory rather than confirmatory models could be that CFA (often) requires dedicated software, whereas EFA can be done with general statistical software. The use of specialized software requires an investment – either in time, money, or both – that researchers might not be willing to make. This reluctance does not necessarily concern financial costs, since there is a variety of free software available that can handle CFA. We mention mx (Neale, Boker, Xie, & Maes, 2003), which is being redeveloped as openmx (OpenMx Development Team, 2010), and the `R` packages `sem` (Fox, Kramer, & Friendly, 2010; Fox, 2006) and `lavaan` (Rosseel, 2010), the latter currently being at a promising, initial stage of development. In addition, free demo versions of both `lisrel` and `mplus` are available, though they are limited with respect to the number of observed variables they can handle.

Sample sizes tend to be somewhat larger for evaluation studies than for other types of studies. This could be due to the fact that an evaluation study is often performed as a secondary analysis of data collected in a large study to investigate properties of the latent variable the scale is supposed to measure. But again, given the large variation in sample size and the limited number of studies, it is not clear how general the observed pattern is.

Statistical analyses reported

Descriptive statistics

Item means are reported in five of the 32 FA studies, in one of the three studies where both FA and IRT are applied, and in four of the six IRT studies. In a fifth IRT study, item difficulty parameters are reported. This difference between FA and IRT applications is to be expected, since one of the focal points of IRT is assessment of item difficulty, of which the item mean is an indicator. In contrast, when the linear factor model is applied, observed variables are often implicitly assumed to be measured as deviations from their means, except when multiple groups are compared (cf. Brown, 2006, p. 54). In addition, researchers applying IRT are traditionally more interested in *item* characteristics, whereas researchers who favor FA are more focused on the multidimensional *structure* of the data (cf. Harman, 1968).

In 38 of the studies (93 %), parameters are estimated. The other three studies are non-parametric IRT applications. When reporting parameter estimates, it is recommended to also report corresponding standard error estimates, because they contain information about the variability, hence the reliability of the parameter estimates (e.g., Boomsma, 2000; McDonald & Ho, 2002). This is especially important for studies with small sample sizes or small respondent/item ratios, where standard errors can be relatively large. Nevertheless, standard errors are only provided in four of the 38 studies.

In 27 of the studies (66 %), some information is given about the distribution of the estimated latent variable scores in the sample. In 13 of these, information about unweighted sum scores is reported; in one study (Hong & Wong, 2005), latent trait estimate information is provided; and in two studies, the average item scores are given. In 11 of these studies, it is unclear what kind of latent variable estimate is employed. Information about the distribution usually consists of means and standard deviations (20 studies), but additional distributional properties (e.g., skewness and kurtosis) are also discussed in six studies. In one study (Glutting, Watkins, & Youngstrom, 2005), only the mean is provided.

Correlations between the latent variable estimates are reported in 29 of the 34 multidimensional studies.

Model assumptions

Application of a model is only useful when its assumptions are not violated beyond specific robustness criteria. The use of ML estimation in FA, for instance, theoretically requires item responses to have a multivariate normal distribution (Bollen, 1989, p. 107). As another example, the Rasch model in IRT assumes unidimensionality of the item responses.

In 19 of the 32 FA studies, model assumptions are not examined or mentioned at all. In nine FA studies, model assumptions are properly investigated. The distribution of the item responses is examined and reported upon, and adequate methods are used, such as robust estimators or the use of an appropriate correlation matrix. In four FA studies, model assumptions are considered only marginally: Item distributions are not investigated or described, but a robust estimator is used nevertheless; or researchers describe that they also analyzed their data using robust estimators but do not report these results because the nonrobust analysis gave virtually the same results. The reason for this practice is not entirely clear but perhaps, because of the similar results, it is implicitly argued that the use of the robust estimator is not necessary. Moreover, researchers sometimes mention that results from standard methods are more easily comparable with results from other studies. However, this statement is only true when the necessary assumptions are sufficiently satisfied. When the assumptions of a method do not hold, researchers ought to choose an appropriate method, based on the characteristics of the available data only, and report its results.

In four of the six IRT studies, model assumptions are properly examined. The unidimensionality assumption is checked in three studies. Item response functions are examined twice for monotonicity (Sabourin, Valois, & Lussier, 2005; Rivas, Bersabé, & Berrocal, 2005) and once for similarity between observed and estimated functions (Wang & Russell, 2005). Hong and Wong (2005) applying the Rasch rating scale model check the assumption of equal spacing of item categories across items. In two IRT applications, assumptions are not given any attention.

In the three studies where both FA and IRT are applied, model assumptions are checked properly once, are given some attention once, and are not investigated at all once.

In nine studies (22 %), robustness of the estimation method is discussed. Sometimes robust statistics are reported: twice Satorra-Bentler's χ^2 test statistic (Grothe et al., 2005; Shevlin & Adamson, 2005), once an extension of Yuan-Bentler's T_2^* test statistic to multilevel models (Zimprich, Perren, & Hornung, 2005), and once `mplus`'s weighted least squares mean and variance adjusted fit statistic (WLSMV; Leite & Beretvas, 2005). In one study (Toland & De Ayala, 2005), the need for a robust estimator is mentioned, but could not be satisfied and thus not applied, because none was available for the specific method.

In eight of the 26 studies where CFA is applied, the covariance matrix is analyzed. The matrix of product-moment (PM) correlations and the matrix of polychoric (PC) correlations are used to estimate the model in four studies each. In two PM and in two PC studies, ML estimation is employed rather than WLS, even though WLS is recommended for

analyzing these matrices, as ML is known to produce erroneous standard error estimates and χ^2-based fit measures when applied to correlation matrices (see, e.g., Cudeck, 1989). In 10 studies, it is unclear what matrix is used for the analysis, which is certainly not a good reporting practice. In one of the 20 studies where EFA is applied, polychoric correlations are analyzed. The other EFA studies either use product-moment correlations or provide no information, which probably also indicates the use of product-moment correlations.

Peculiarities

When model assumptions are violated, an analysis may produce unexpected results. Peculiar results may also occur due to other problems, such as model underidentification. Remarkably, none of the studies report peculiarities such as nonconvergence of the estimation procedure or the occurrence of Heywood cases. It is, however, strongly recommended (e.g., Bandalos & Finney, 2010; Boomsma, 2000; De Ayala, 2010) to pay attention to unexpected features of an analysis first, before performing any other evaluation of the results of that analysis. If no peculiarities are encountered, one should also report this fact.

The lack of reported peculiarities could be explained in a number of ways. Researchers could fail to notice peculiar outcomes; or would prefer not to report them, because they think that journal editors might not be inclined to accept papers including studies with peculiarities (cf. the "file-drawer" problem; Scargle, 2000).

Model fit and modification

In 26 of the studies, CFA is part of the analysis. Model fit is formally tested in all of these, except for one (Arrindell et al., 2005), where the multiple group method (Holzinger, 1944) is applied, for which no formal test of model fit exists. The measures most often reported for examining model fit are the root mean square error of approximation (RMSEA), the goodness-of-fit index (GFI), the comparative fit index (CFI), and the Tucker-Lewis or nonnormed fit index (TLI or NNFI). Other reported measures are the (standardized) root mean residual [(S)RMR], the normed fit index (NFI), the incremental fit index (IFI), and the adjusted goodness-of-fit index (AGFI). In addition, the χ^2 test statistic is usually reported with corresponding degrees of freedom and *p*-value, most often with the annotation that this fit statistic is very sensitive to sample size. It should be noted, however, that the χ^2 test statistic is even more sensitive to nonnormality of the observed variables (Boomsma, 1983). When competing models are compared, reported measures are the χ^2 difference test, Akaike's information criterion (AIC) or a consistent version of the AIC (CAIC), and the expected cross-validation index (ECVI).

Factor loadings are reported in 23 of the 26 studies that include a CFA, and are discussed as a criterion for model fit evaluation in six of those studies. The items are then usually

required to load significantly on the factor or have a loading higher in absolute value than a criterion value such as .40.

Regarding the choice of fit criteria, researchers often refer to one or more statistical publications. In five studies where criteria are applied, no references are given. In 13 studies, Hu and Bentler (1995, 1998, or 1999) are mentioned. Browne and Cudeck (1993), Hoyle and Panter (1995), Hatcher (1994), and Kline (1998) are referred to in four, three, three, and two studies, respectively. In four studies, other literature is mentioned.

When EFA is conducted, the fit of the model is not formally tested. Instead, researchers use specific criteria to determine the number of factors underlying the items. In 18 of the 21 EFA studies, item factor loadings are examined to determine whether items belong to a factor. Factor loadings greater than .30–.40 in absolute value are usually interpreted as salient. The number of factors is commonly determined by a combination of criteria: examination of a scree plot (15 studies), parallel analysis (10 studies), and/or the eigenvalue > 1 criterion (eight studies). In four studies, the percentage of explained variance by a factor is mentioned, without an explicit criterion on how to evaluate it. In only five studies, the interpretability of the factors is mentioned as a criterion. These results, once again, suggest that many researchers do not use their substantive knowledge about the items to evaluate the structure of the data.

In the IRT studies, formal tests of model fit are never provided. When the Mokken model is applied, Loevinger's *H*-value is reported as an indication of the scale's strength. In three IRT studies, unidimensionality is tested.

The difference between the FA and IRT studies in assessing model fit reflects a difference in traditions. In FA, it has become standard practice to report a collection of indices and compare them to cutoff values, a process that has been criticized in the literature (e.g., Chen, Curran, Bollen, Kirby, & Paxton, 2008; Marsh, Hau, & Wen, 2004). IRT model fit measures seem to be less well known among researchers, while at the same time claims about the correctness of a model are also more modest than is the case with FA.

In six of the studies where CFA is applied, the model is modified at some point. In three studies, both modification indices and item content are used to modify the model; in one study (Zapf, Skeem, & Golding, 2005), only modification indices are used; in one study (Toland & De Ayala, 2005), only item content is considered; and in one study, the number of factors is adapted after a parallel analysis (Heinitz, Liepmann, & Felfe, 2005). In the IRT studies, the model is never modified, other than by removing a number of items from the scale.

Some items are discarded in 18 of the studies (44 %). Criteria for item retention include: item content (13 studies); factor loading greater than a certain cutoff value, usually .30 or .40 in absolute value (11 studies); loading on an additional unintended factor smaller than about .20–.40 in absolute value (eight studies); sufficient item discrimination (two studies); and lack of decrease of the scale's Cronbach's coefficient alpha (α; Cronbach, 1951) when the item is excluded (three studies). It should be noted that α typically increases with the number of items in a scale (e.g., Cortina, 1993), and therefore the latter

criterion is not very useful. In one study (Beyers, Goossens, Calster, & Duriez, 2005), the criteria used for item retention are not made explicit.

Reliability

Reliability refers to the consistency of measurement – that part of a measure that is free of random error (Bollen, 1989, p. 206 ff.). The reliability of a scale is estimated in 35 of the studies (85 %), usually by computing Cronbach's α for each subscale. It is remarkable that α is still the most commonly used reliability measure, even though other coefficients, like the lower bounds proposed by Guttman (1945), provide greater lower bounds to reliability (e.g., Jackson & Agunwamba, 1977; see also Zinbarg, Revelle, Yovel, & Li, 2005). Moreover, research on the behavior of α (e.g., Cortina, 1993; see also Sijtsma, 2009a, for a historical overview) does not seem to be well known among applied researchers. Cortina criticized the practice of comparing α to a cutoff value, like .70 or .80, without any consideration of context, since the interpretation of α depends on many factors, such as test length and sample homogeneity. Recently, the use of α was criticized and discussed again (Bentler, 2009; S. B. Green & Yang, 2009a, 2009b; Revelle & Zinbarg, 2009; Sijtsma, 2009a, 2009b), with recommendations for alternative reliability estimators and software to employ them. Furthermore, confidence intervals for α (Iacobucci & Duhachek, 2003; Koning & Franses, 2003) are hardly ever reported (three studies, 7 %), even though they provide a means of comparing α values from different studies.

In the studies where a nonparametric Mokken IRT analysis is applied, reliability is estimated based on the so-called P(++) matrices. In parametric IRT, one traditionally focuses on item and test information functions, because the measurement error of a scale is a function of the latent variable values. However, in only one of the three parametric IRT applications (Caprara, Steca, Zelli, & Capanna, 2005) are information curves examined to assess reliability. Furthermore, in none of the studies are reliability measures developed within the IRT tradition, such as marginal reliability (B. F. Green, Bock, Humphreys, Linn, & Reckase, 1984) or EAP reliability (Adams, 2005), reported.

In 12 of the 34 studies reporting on a multidimensional scale, a composite reliability measure is provided. In most cases, this is done by computing α for the entire set of items taken together, even though more sophisticated composite reliability measures, like weighted ω (e.g., Bacon, Sauer, & Young, 1995; McDonald, 1970; see also Raykov & Shrout, 2002) are available. Weighted ω is reported in only one study (Clark et al., 2005). Reliance on the value of α as a lower bound to composite reliability is not justified, as shown by Raykov (1998). He argued that α can be an overestimation (rather than an underestimation) of the composite reliability of a scale when the items have correlated errors.

Validity

Cross-validation is performed in 17 of the studies (41 %). In 10 of these, one sample is used to calibrate a proposed structure of the data, and a second, independent sample is used for validation purposes. In seven studies, one original sample is randomly split in half to create a calibration and a validation sample. Cross-validation is never applied in the strict sense of imposing the parameter estimates found in the calibration sample on the validation sample, and evaluating the fit (e.g., Camstra & Boomsma, 1992; Cudeck & Browne, 1983). In the examined studies, the common procedure of cross-validation is a CFA on a second sample to test the final model structure that was found by applying an EFA or a CFA on a first sample.

In studies aiming to evaluate a scale, cross-validation is performed remarkably less often than in other types of studies (see Table 3). Since an evaluation study is a further examination of an existing scale, researchers might consider their study to be a cross-validation of an earlier study, and reason that an extra cross-validation is unnecessary.

In 24 of the studies (59 %), some external validation of the scale is performed. This is usually accomplished by examining correlations of the scale under investigation with other measures of the construct (convergent validity) or with measures of related but distinct constructs (divergent validity).

Expert coauthor

As a final aspect for comparison, we evaluate some study characteristics in terms of the involvement of researchers with methodological expertise. For this purpose the website or online curriculum vitae of each author was examined. Unfortunately, for six of the studies no information about the authors could be found online. For the remaining 35 studies, we distinguish between (a) studies that are (co)authored by a methodological expert or a psychometrician, (b) studies where the contributions of an expert are acknowledged in an author note, (c) studies where one of the authors shows a research interest in psychometrics or quantitative methods, and (d) studies without any involvement of a methodological expert. Some descriptive statistics regarding this distinction may be found in Table 4.

IRT analyses are only performed in studies where one of the authors is a methodological expert, or where an expert's contribution is acknowledged in a note. This also holds for studies where both FA and IRT are applied. It seems that applied researchers without a specific methodological research interest are not sufficiently familiar with IRT to apply it without consulting an expert.

Most of the studies in *Educational and Psychological Measurement* (EPM) are coauthored by a methodological expert. This is not the case for the studies in the other two journals. Furthermore, all of the studies where explicit motives are provided for the choice of methodology – discussing both FA and IRT – are coauthored by a methodo-

Table 4:
Selection of Characteristics for Studies With Varying Degree of Involvement of a Methodological Expert

	Methodological expert as coauthor?				
Characteristic	Yes	Acknowl.[a]	Research interest[b]	No	Unknown[c]
Type of analysis					
FA	11		3	12	6
IRT	5	1			
FA & IRT	2	1			
Journal					
EJPA[d]	4		1	2	6
EPM[e]	12		1	2	
PA[f]	2	2	1	8	
Motives					
no	10			8	6
some	5	2	3	4	
explicit	3				

Note. Numbers in the table represent frequencies of studies.
[a]A methodological expert acknowledged for valuable contributions in a note. [b]One of the authors has a research interest in quantitative methods. [c]Information about the authors could not be retrieved via the Internet. [d]European Journal of Psychological Assessment. [e]Educational and Psychological Measurement. [f]Psychological Assessment.

logical expert. However, it is worth noting that in more than half of the studies coauthored by a methodological expert, no such motives are given. This is rather disappointing and reflects, perhaps, a conflict between authors on the substance of the paper, possibly influenced by limitations in the length of papers accepted by journals.

Discussion

A possible limitation of our study is the lack of generalizability of our findings. Although the studies we reviewed are hardly a random sample from all published factor analysis (FA) and item response theory (IRT) applications in psychology and education, we do believe that they provide a – perhaps limited – impression of the current status of scale construction and evaluation in practice. In fact, we believe that this impression is relatively favorable, since the selected journals offer authors the opportunity to report on statistically sound research, making it likely that the studies reviewed here are among the methodologically stronger ones.

We found that FA is applied far more often than IRT. As researchers do not sufficiently explicate their model choice, we are forced to make some educated guesses about possible explanations for this fact. Researchers may not feel obliged to give motivation and

prefer to use their limited publishing space for different matters, or perhaps they feel uncertain about their choice. In the latter case, the guidelines for applying quantitative methods in the social sciences in a book edited by Hancock and Mueller (2010) might be a useful reference for both authors and reviewers. These guidelines concern model choice as well as reporting practice.

Expectations about the dimensionality of the scale could be a motive to apply FA instead of IRT, even though this motive is not explicitly stated in the studies. Researchers' lack of familiarity with software for multidimensional IRT models could well be an important reason for IRT's relative unpopularity. FA software is better known: EFA (including principal component analysis) can be conducted using most general statistical packages; for CFA, the lisrel program is the most popular and best known. Software for multidimensional IRT models is highly specialized and not very easy to find. Some multidimensional IRT software packages are limited to dichotomous items only, e.g., `testfact` (Wilson, Wood, & Gibbons, 1984) and `mirte` (Carlson, 1987); polytomous items can be handled by, e.g., `mplus`, `conquest` (Wu, Adams, & Wilson, 1998), the `stata` program `glamm` (Rabe-Hesketh, Skrondal, & Pickles, 2004), and the `R` package `mcmc` (Martin, Quinn, & Park, 2007).

In our opinion, researchers could take far better advantage of their theoretical knowledge and/or expectations by incorporating their a priori knowledge of the items and scales in the analyses. This should be reflected (a) by more frequent application of confirmatory techniques, especially in the construction of new scales; and (b) by adding interpretability of factors and content of items to the criteria used for model evaluation.

The issue of applying exploratory versus confirmatory techniques has been discussed by a number of authors (Ferrando & Lorenzo-Seva, 2000; Floyd & Widaman, 1995; Gerbing & Hamilton, 1996; McCrae, Zonderman, Costa, Bond, & Paunonen, 1996). From these studies, although differing somewhat in scope, it can be concluded that EFA performs reasonably well at recovering a hypothesized factor structure. Results of CFA and EFA may often be different, with CFA fit measures indicating an unsatisfactory fit of structures uncovered by EFA. The use and interpretation of fit indices, however, is still (or again) under debate (e.g., Chen et al., 2008; Marsh et al., 2004; Saris, 2008; Vernon & Eysenck, 2007).

It is troublesome that in fewer than half of the studies (46 %) are model assumptions investigated to check whether the chosen scaling model is appropriate for the measurement level and distribution of the data, even though well-known guidelines such as those of Wilkinson and the Task Force on Statistical Inference (1999) encourage researchers to do so. Our results are better in this respect than those reported by Osborne (2008), who reviewed 96 articles in the field of educational psychology published in 1998–1999, and found that merely 8.3 % reported testing the assumptions of the statistical tests that were used.

Most, if not all, scales in psychological and educational research use ordered categorical item responses, invalidating the assumption of a linear relation between the items and the latent variable as posed in ordinary FA. Use of this linear model as a pragmatic approach to the analysis of a scale should always be preceded by careful inspection of the distribution of the data. A fair amount of research has dealt with the consequences of applying a

linear factor model to polytomous data (e.g., Boomsma, 1983; Coenders, Satorra, & Saris, 1997; Flora & Curran, 2004; Hoogland, 1999; Jöreskog & Moustaki, 2001; Moustaki et al., 2004; B. O. Muthén & Kaplan, 1985, 1992). From these studies it could be concluded that categorical and ordinal data raise no serious problems as long as the distribution of the item variables is approximately normal, illustrating the importance of examining model assumptions. When the distributional assumptions are violated, alternative, robust estimators are proposed that might require specialized software and consultation with methodological experts.

The final question is: What can we learn from the present study, other than *"most scale research uses FA, some uses IRT, and hardly any uses both"?* Most importantly, we learn that far too often models are applied without proper justification. Model assumptions could and should be investigated more frequently. If limited publication space is a bottleneck, authors could consider referring to a website where the results of their analyses would be available for interested fellow researchers. Journal editors may want to encourage such practice by providing journal web space. If expertise is a factor that is lacking, a more frequent collaboration between substantive researchers and statisticians/psychometricians should be encouraged, requiring an active role for both parties. Finally, the education in methodology and statistics for (future) scale developers in the fields of psychology and education might need some reconsideration and reinforcement, again requiring an active role for both substantive researchers and methodological experts.

Author note

This research was supported by the Netherlands Organisation for Scientific Research [NWO], file 400–04–137. The authors thank Charles Lewis, Marieke Timmerman, and two anonymous reviewers for their helpful comments and suggestions.

References

References marked with an asterisk indicate studies included in the review.

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*, 162-172.

* Aluja, A., Blanch, A., & García, L. F. (2005). Dimensionality of the Maslach Burnout Inventory in school teachers: A study of several proposals. *European Journal of Psychological Assessment, 21*, 67-76.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Arbuckle, J. (1995-2006). Amos 7.0 user's guide [Computer software manual]. Chicago: SPSS.

* Arrindell, W., Akkerman, A., Bagés, N., Feldman, L., Caballo, V. E., Oei, T. P., et al. (2005). The short-EMBU in Australia, Spain, and Venezuela. *European Journal of Psychological Assessment, 21*, 56-66.

* Aycicegi, A., Dinn, W. M., & Harris, C. L. (2005). Validation of Turkish and English Versions of the Schizotypal Personality Questionnaire–B. *European Journal of Psychological Assessment, 21*, 34-43.

Bacon, D. R., Sauer, P. S., & Young, M. (1995). Composite reliability in structural equation modeling. *Educational and Psychological Measurement, 55*, 394-406.

Bandalos, D. L., & Finney, S. J. (2010). Factor analysis: Exploratory and confirmatory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 93-114). New York: Routledge.

Bentler, P. M. (1989). EQS structural equations program manual [Computer software manual]. Los Angeles, CA: BMDP Statistical Software.

Bentler, P. M. (1995). EQS program manual [Computer software manual]. Encino, CA: Multivariate Software.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74*, 137-143.

* Beyers, W., Goossens, L., Calster, B. V., & Duriez, B. (2005). An alternative substantive factor structure of the Emotional Autonomy Scale. *European Journal of Psychological Assessment, 21*, 147-155.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-424). Reading, MA: Addison-Wesley.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality*. Amsterdam: Sociometric Research Foundation.

Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling, 7*, 461-483.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Beverly Hills, CA: Sage.

* Calvete, E., Estévez, A., Arroyabe, E. L. de, & Ruiz, P. (2005). The Schema Questionnaire–Short Form. *European Journal of Psychological Assessment, 21*, 90-99.

Camstra, A., & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis. *Sociological Methods & Research, 21*, 89-115.

* Caprara, G. V., Steca, P., Zelli, A., & Capanna, C. (2005). A new scale for measuring adults' prosocialness. *European Journal of Psychological Assessment, 21*, 77-89.

Carlson, J. E. (1987). Multidimensional item response theory estimation: A computer program [Computer software manual]. Retrieved from http://www.act.org/research/reports/pdf/ACT RR87-19.pdf.

* Cashin, S. E., & Elmore, P. B. (2005). The Survey of Attitudes Toward Statistics Scale: A construct validity study. *Educational and Psychological Measurement, 65*, 509-524.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research, 36*, 462-494.

* Clark, D. A., Antony, M. M., Beck, A. T., Swinson, R. P., & Steer, R. A. (2005). Screening for obsessive and compulsive symptoms: Validation of the Clark-Beck Obsessive-Compulsive Inventory. *Psychological Assessment, 17*, 132-143.

Coenders, G., Satorra, A., & Saris, W. E. (1997). Alternative approaches to structural modeling of ordinal data: A Monte Carlo study. *Structural Equation Modeling, 4*, 261-282.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104.

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

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105*, 317-327.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18*, 147-167.

De Ayala, R. J. (2010). Item response theory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 155-171). New York: Routledge.

* De Frias, C. M., & Dixon, R. A. (2005). Confirmatory factor structure and measurement invariance of the Memory Compensation Questionnaire. *Psychological Assessment, 17*, 168-178.

Ferrando, P. J., & Lorenzo-Seva, U. (2000). Unrestricted versus restricted factor analysis of multidimensional test items: Some aspects of the problem and some suggestions. *Psicológica, 21*, 301-323.

* Fletcher, R., & Hattie, J. (2005). Gender differences in physical self-concept: A multidimensional differential item functioning analysis. *Educational and Psychological Measurement, 65*, 657-667.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*, 466-491.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*, 286-299.

Fox, J. (2006). Structural equation modeling with the sem package in R. *Structural Equation Modeling, 13*, 465-486.

Fox, J., Kramer, A., & Friendly, M. (2010). sem (R package Version 0.9–20): Structural equation models [Computer software]. Retrieved from CRAN.R-project.org/package=sem.

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling, 3*, 62-72.

* Ghaderi, A. (2005). Psychometric properties of the Self-Concept Questionnaire. *European Journal of Psychological Assessment, 21*, 139-146.

Glöckner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling, 10*, 544-565.

* Glutting, J. J., Watkins, M. W., & Youngstrom, E. A. (2005). ADHD and college students: Exploratory and confirmatory factor structures with student and parent data. *Psychological Assessment, 17*, 44-55.

Green, B. F., Bock, D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-360.

Green, S. B., & Yang, Y. (2009a). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74*, 121-135.

Green, S. B., & Yang, Y. (2009b). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*, 155-167.

* Grothe, K. B., Dutton, G. R., Jones, G. N., Bodenlos, J., Ancona, M., & Brantley, P. J. (2005). Validation of the Beck Depression Inventory–II in a low-income African American sample of medical outpatients. *Psychological Assessment, 17*, 110-114.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 23*, 297-308.

Hancock, G. R., & Mueller, R. O. (Eds.). (2010). *The reviewer's guide to quantitative methods in the social sciences*. New York: Routledge.

Harman, H. H. (1968). *Modern factor analysis* (2nd ed.). Chicago: University of Chicago.

Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling.* Cary, NC: SAS Institute.

* Heinitz, K., Liepmann, D., & Felfe, J. (2005). Examining the factor structure of the MLQ: Recommendation for a reduced set of factors. *European Journal of Psychological Assessment, 21*, 182-190.

Holzinger, K. J. (1944). A simple method of factor analysis. *Psychometrika, 9*, 257-262.

* Hong, S., & Wong, E. C. (2005). Rasch rating scale modeling of the Korean version of the Beck Depression Inventory. *Educational and Psychological Measurement, 65*, 124-139.

Hoogland, J. J. (1999). *The robustness of estimation methods for covariance structure analysis*. Unpublished doctoral dissertation, University of Groningen.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 158-198). Thousand Oaks, CA: Sage.

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 76-99). Thousand Oaks, CA: Sage.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 4*, 424-453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Iacobucci, D., & Duhachek, A. (2003). Advancing alpha: Measuring reliability with confidence. *Journal of Consumer Psychology, 13*, 478-487.

* Inglés, C. J., Hidalgo, M. D., & Méndez, F. X. (2005). Interpersonal difficulties in adolescence: A new self-report measure. *European Journal of Psychological Assessment, 21*, 11-22.

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. *Psychometrika, 42*, 567-578.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36*, 347-387.

Jöreskog, K. G., & Sörbom, D. (1996). LISREL 8: User's reference guide [Computer software manual]. Chicago: Scientific Software International.

* Joseph, S., Linley, P. A., Andrews, L., Harris, G., Howle, B., & Woodward, C. (2005). Assessing positive and negative changes in the aftermath of adversity: Psychometric evaluation of the Changes in Outlook Questionnaire. *Psychological Assessment, 17*, 70-80.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15*, 136-153.

Kline, R. (1998). *Principles and practice of structural equation modeling*. New York: The Guilford Press.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26*, 457-477.

Koning, A. J., & Franses, P. H. (2003, June). *Confidence intervals for Cronbach's coefficient alpha values* (Tech. Rep. No. ERS-2003–041-MKT). Rotterdam, the Netherlands: Erasmus Research Institute of Management. Retrieved from http://publishing.eur.nl/ir/repub/asset/431/ERS-2003-041-MKT.pdf.

* Kotov, R., Schmidt, N. B., Zvolensky, M. J., Vinogradov, A., & Antipova, A. V. (2005). Adaptation of panic-related psychopathology measures to Russian. *Psychological Assessment, 17*, 242-246.

* Le, H., Casillas, A., Robbins, S. B., & Langley, R. (2005). Motivational and skills, social, and self-management predictors of college outcomes: Constructing the Student Readiness Inventory. *Educational and Psychological Measurement, 65*, 482-508.

* Leite, W. L., & Beretvas, S. N. (2005). Validation of scores on the Marlowe-Crowne Social Desirability Scale and the Balanced Inventory of Desirable Responding. *Educational and Psychological Measurement, 65*, 140-154.

* Longley, S. L., Watson, D., & Noyes, R., Jr. (2005). Assessment of the hypochondriasis domain: The Multidimensional Inventory of Hypochondriacal Traits (MIHT). *Psychological Assessment, 17*, 3-14.

* Lowe, P. A., & Reynolds, C. R. (2005). Factor structure of AMAS-C scores across gender among students in collegiate settings. *Educational and Psychological Measurement, 65*, 687-708.

* Marsh, H. W., Ellis, L. A., Parada, R. H., Richards, G., & Heubeck, B. (2005). A short version of the Self Description Questionnaire II: Operationalizing criteria for short-form evaluation with new applications of confirmatory factor analyses. *Psychological Assessment, 17*, 81-102.

Marsh, H. W., & Hau, K.-T. (1999). Confirmatory factor analysis: Strategies for small sample sizes. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 251-306). Thousand Oaks, CA: Sage.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*, 181-220.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*, 320-341.

Martin, A. D., Quinn, K. M., & Park, J. H. (2007). MCMCpack: Markov chain Monte Carlo (MCMC) package. R package Version 0.9–2 [Computer software]. Retrieved from http://mcmcpack.wustl.edu.

McCrae, R. R., Zonderman, A. B., Costa, P. T., Jr., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology, 70*, 552-566.

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology, 23*, 1-21.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation models. *Psychological Methods, 7*, 64-82.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7*, 361-388.

Mehta, P. D., & Taylor, W. P. (2006, June). *On the relationship between item response theory and factor analysis of ordinal variables: Multiple group case.* Paper presented at the 71st annual meeting of the Psychometric Society, HEC Montreal, Canada.

Molenaar, I. W. (1974). De logistische en de normale kromme [The logistic and the normal curve]. *Nederlands Tijdschrift voor de Psychologie, 29*, 415-420.

Moustaki, I., Jöreskog, K. G., & Mavridis, D. (2004). Factor models for ordinal variables with covariate effects on the manifest and latent variables: A comparison of LISREL and IRT approaches. *Structural Equation Modeling, 11*, 487-513.

Mueller, R. O. (1996). *Basic principles of structural equation modeling: An introduction to LISREL and EQS.* New York: Springer.

* Muller, J. J., Creed, P. A., Waters, L. E., & Machin, M. A. (2005). The development and preliminary testing of a scale to measure the latent and manifest benefits of employment. *European Journal of Psychological Assessment, 21*, 191-198.

Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*, 171-189.

Muthén, B. O., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45*, 19-30.

Muthén, L. K., & Muthén, B. O. (1998–2010). Mplus user's guide (6th ed.) [Computer software manual]. Los Angeles, CA: Author.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2003). Mx: Statistical modeling (6th ed.) [Computer software manual]. Retrieved from http://www.vipbg.vcu.edu/~vipbg/software/mxmanual.pdf. Richmond: Virginia Commonwealth University, Department of Psychiatry.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

OpenMx Development Team. (2010). OpenMx documentation [Computer software manual]. Retrieved from http://openmx.psyc.virginia.edu/docs/OpenMx/ latest/OpenMxUserGuide.pdf.

Osborne, J. W. (2008). Sweating the small stuff in educational psychology: how effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. *Educational Psychology, 28*, 151-160.

* Pett, M. A., & Johnson, M. J. M. (2005). Development and psychometric evaluation of the Revised University Student Hassles Scale. *Educational and Psychological Measurement, 65*, 984-1010.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). GLLAMM manual. U.C. Berkeley division of biostatistics working paper series. Working paper 160 [Computer software manual]. Retrieved from http://www.bepress.com/ucbbiostat/paper160.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*, 517-529.

Ramsay, J. O. (2000). TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data [Computer software manual]. Retrieved from ftp://ego.psych.mcgill.ca/pub/ramsay/testgraf/.

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement, 22*, 375-385.

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling, 9*, 195-212.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*, 145-154.

* Rivas, T., Bersabé, R., & Berrocal, C. (2005). Application of the Double Monotonicity Model to polytomous items, scalability of the Beck depression items on subjects with eating disorders. *European Journal of Psychological Assessment, 21*, 1-10.

Rosseel, Y. (2010). lavaan (R package Version 0.3–1): Latent variable analysis [Computer software]. Retrieved from http://CRAN.R-project.org/package=lavaan.

* Sabourin, S., Valois, P., & Lussier, Y. (2005). Development and validation of a brief version of the Dyadic Adjustment Scale with a nonparametric item analysis model. *Psychological Assessment, 17*, 15-27.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 36*, 109-133.

Saris, W. E. (2008, September). *Tests of structural equation models do not work: What to do?* Paper presented at the 7th International Conference on Social Science Methodology, Naples, Italy.

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50*, 83-90.

Scargle, J. D. (2000). Publication bias: The "file-drawer" problem in scientific inference. *Journal of Scientific Exploration, 14*, 91-106.

* Shevlin, M., & Adamson, G. (2005). Alternative factor models and factorial invariance of the GHQ-12: A large sample analysis using confirmatory factor analysis. *Psychological Assessment, 17*, 231-236.

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

Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika, 74*, 169-173.

* Simms, L. J., Casillas, A., Clark, L. A., Watson, D., & Doebbeling, B. N. (2005). Psychometric evaluation of the restructured clinical scales of the MMPI2. *Psychological Assessment, 17*, 345-358.

* Stepleman, L. M., Darcy, M. U., & Tracey, T. J. (2005). Helping and coping attributions: Development of the Attribution of Problem Cause and Solution Scale. *Educational and Psychological Measurement, 65*, 525-542.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

* Toland, M. D., & De Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement, 65*, 272-296.

* Tomás-Sábado, J., & Gómez-Benito, J. (2005). Construction and validation of the Death Anxiety Inventory (DAI). *European Journal of Psychological Assessment, 21*, 108-114.

* Torff, B., Sessions, D., & Byrnes, K. (2005). Assessment of teachers' attitudes about professional development. *Educational and Psychological Measurement, 65*, 820-830.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*, 6-20.

* Van der Pas, S., Van Tilburg, T., & Knipscheer, K. C. P. M. (2005). Measuring older adults' filial responsibility expectations: Exploring the application of a vignette technique and an item scale. *Educational and Psychological Measurement, 65*, 1026-1045.

* Verdugo, M. A., Prieto, G., Caballo, C., & Peláez, A. (2005). Factorial structure of the Quality of Life Questionnaire in a Spanish sample of visually disabled adults. *European Journal of Psychological Assessment, 21*, 44-55.

Vernon, T., & Eysenck, S. (Eds.). (2007). Special issue on structural equation modeling. *Personality and Individual Differences, 42*(5).

* Vigneau, F., & Bors, D. A. (2005). Items in context: Assessing the dimensionality of Raven's Advanced Progressive Matrices. *Educational and Psychological Measurement, 65*, 109-123.

* Wang, M., & Russell, S. S. (2005). Measurement equivalence of the Job Descriptive Index across Chinese and American workers: Results from confirmatory factor analysis and item response theory. *Educational and Psychological Measurement, 65*, 709-732.

* Weeks, J. W., Heimberg, R. G., Fresco, D. M., Hart, T. A., Turk, C. L., Schneier, F. R., et al. (2005). Empirical validation and psychometric evaluation of the brief Fear of Negative Evaluation Scale in patients with social anxiety disorder. *Psychological Assessment, 17*, 179-190.

* Wiebe, J. S., & Penley, J. A. (2005). A psychometric comparison of the Beck Depression Inventory-II in English and Spanish. *Psychological Assessment, 17*, 481-485.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist, 54*, 594-604.

Wilson, D., Wood, R., & Gibbons, R. D. (1984). Testfact: Test scoring, item statistics, and item factor analysis [Computer software]. Mooresville, IN: Scientific Software.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*, 58-79.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). ConQuest: Multi-aspect test software [Computer software manual]. Melbourne, Australia: Australian Council for Educational Research.

* Zapf, P. A., Skeem, J. L., & Golding, S. L. (2005). Factor structure and validity of the MacArthur Competence Assessment Tool – Criminal Adjudication. *Psychological Assessment, 17*, 433-445.

* Zimprich, D., Perren, S., & Hornung, R. (2005). A two-level confirmatory factor analysis of a modified Rosenberg Self-Esteem Scale. *Educational and Psychological Measurement, 65*, 465-481.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123-133.