Developing Multidimensional Likert Scales using Item Factor Analysis: The Case of Four-Point Items[1]

Abstract

This study compares the performance of two approaches in analysing fourpoint Likert rating scales with a factorial model: the classical factor analysis (FA) and the item factor analysis (IFA). For FA, maximum likelihood and weighted least squares estimations using Pearson correlation matrices among items are compared. For IFA, diagonally weighted least squares and unweighted least squares estimations using items polychoric correlation matrices are compared. Two hundred and ten conditions were simulated in a Monte Carlo study considering: one to three factor structures (either, independent and correlated in two levels), medium or low quality of items, three different levels of item asymmetry and five sample sizes. Results showed that IFA procedures achieve equivalent and accurate parameter estimates; in contrast, FA procedures yielded biased parameter estimates. Therefore, we do not recommend classical FA under the conditions considered. Minimum requirements for achieving accurate results using IFA procedures are discussed.

Introduction

The Likert Rating Scale (Likert, 1932; Likert, Roslow & Murphy 1934) is a simple procedure to generate measurement instruments which is widely used by social scientists to measure a variety of latent constructs, therefore, meticulous statistical procedures have been developed to design and validate these scales (see e.g.: DeVellis, 1991; Spector, 1992), however, most of them ignore the ordinal nature of observed responses and assume the presence of continuous observed variables measured at interval level. Although a very interesting debate about the robustness to ordinal data of parametric statistical techniques to analyze Likert Scales still be present (Jamieson, 2004; Carifio & Perla, 2007; Norman, 2010), evidence shows that, under relatively common circumstances, classical Factor Analysis (FA) yield inaccurate results characterizing the internal structure of the scale or selecting of the most informative items within each factor (Berstein & Teng, 1989; DiStefano, 2002; Holgado–Tello, Chacón–Moscoso, Barbero–García & Vila–Abad, 2010). Fortunately, Item Factor Analysis (IFA) provides an alternative that avoids these problems (Wirth & Edwards, 2007) because it addresses and recognizes the ordinal nature of observed variables.

Although the relevance of IFA for developing Likert Scales has been acknowledged (Flora and Curran 2004), there is some debate regarding the specific estimation procedures to employ, especially in the case of polytomous items (Savalei and Rhemtulla 2013), and an alternative estimation procedure that could allow the use of FA in ordinal data instead of IFA has not been ruled out.

Thus, this article aims to address this gap by presenting the results of a simulation study comparing the performance of the most recommended IFA estimation procedures and

[1] Asún, R. A., Rdz-Navarro, K., & Alvarado, J. M. (2015). Developing Multidimensional Likert Scales Using Item Factor Analysis The Case of Four-point Items. *Sociological Methods & Research (*Published online before print January 13, 2015)

some alternatives in classical FA. Given that the performance of estimation procedures depends on the number of item response categories (Beauducel and Herzberg 2006; Dolan 1994; Savalei and Rhemtulla 2013), this research will focus on four-point items, whose consequences have been little investigated despite it being the most widely employed format for Likert Scales when the intermediate category is suspected to be inadequate.

The number of response categories on Likert items

Since Rensis Likert suggested the scaling procedure which now bears his name, a strong debate have been placed with regard to the optimal number of categories to present to the subjects responding the questionnaire. Interestingly, the evidence found in literature support highly contrasting positions: some researchers suggest that larger numbers of response categories enable reaching higher levels of reliability (Garner, 1960) and validity (Hancock & Klockars, 1991; Loken, Pirie, Virnig, Hinkle & Salmon, 1987); while others suggest that the number of response categories is not related to the reliability of the scale (Boote, 1981; Brown, Wilding & Coulter, 1991) and its validity (Chang, 1994; Matell & Jacoby, 1971). Overall, the evidence tend to indicate that: i) researchers should avoid presenting few response categories (two or three) to the subjects as it could decrease the validity of the scale and the subjects may feel they are not able to express their true opinion when responding the questionnaire (Preston & Colman, 2000); and ii) benefits of increasing the number of response categories will vanish if more than seven-points are presented to the subjects, because they might not be able to discriminate among them (Miller, 1956).

For those reasons, most of the Likert scales employ 4 up to 7 response categories and, five or seven-points are the most common format used in applied research (Cox III, 1980). The preference for an odd number of response categories reflects a tendency to choose items that allow subjects to define their position as ‘neutral’ with respect to the construct intended to measure (Preston & Colman, 2000).

Nevertheless, the intermediate category may affect the validity of results because: i) subjects could use this category for reasons different than having an intermediate opinion, for example, the subject have no opinion, does not want to express his/her true opinion, does not understand the question, is facing a ‘not applicable’ question, among others (Kulas, Stachowski, & Haynes, 2008; Raaijmakers, van Hoof, Hart, Verbogt & Wollebergh, 2000); ii) a relationship among social desirability and the intermediate category option has been reported in previous literature (Garland, 1991); iii) it is a cumbersome task to semantically express the idea of neutrality in the continuum of response categories (González-Romá & Espejo, 2003); and iv) in certain occasions, the information contributed by an intermediate category is not informative (Andrich, 1978)

Therefore, a four-points response format is highly attractive when social desirability is suspected to affect the construct intended to measure, subjects are heterogeneous in their capacities to discriminate among categories (i.e. sample is drawn from a general population) or when the interview administration method (e.g.: face-to-face) makes it difficult to employ a larger number of response categories.

However, when considering a four-point response format, researchers should bear in mind that as the number of response categories decreases, the resemblance of observed items with variables measured at interval level is likely to be vanished, therefore, statistical analysis like classical FA shall yield inaccurate results.

Likert Scales and Classical Factor Analysis

The FA has been widely acknowledged as a central procedure to develop Likert scales (Nunnally, 1978). Thus, the conventional wisdom indicates that, when a unidimensional scale is desired and the subjects' responses to a set of items are available, items could be selected using Pearson correlations among the item and total scale and/or selecting the items that maximize the reliability and internal consistency of the scale using Cronbach's Alpha[2] (DeVellis, 1991) and afterwards, FA could be employed to assess the internal structure of the scale. If a multidimensional construct is measured, researchers tend to begin the process using FA to assess the internal structure of the data (confirming or modifying their initial ideas about it) and then proceed selecting the items that better reflect each factor using factor loadings or the same statistical analyses employed for the unidimensional case but within each dimension separately (Spector, 1992).

One of the problems of this scenario is that classical FA assumes continuous observed variables measured at interval level and the estimation procedures frequently employed in FA, such us Maximum Likelihood estimation (ML), assume multivariate normal distribution of observed responses. Contrastingly, items in a Likert scale are coded using a procedure known as *integer scoring* (González-Romá & Espejo, 2003), which assigns integer successive numbers to each response category (i.e. 1, 2, 3, …, *n*), therefore, items can be regarded only as ordinal measurements, in the best case scenario.

Several authors have argued that statistical validity does not depend on levels of measurement (Gaito, 1980; Lord, 1953; Velleman & Wilkinson, 1993), that statistical analyses are robust to ordinal data (Norman, 2010) and furthermore, that Likert scales produce interval level of measurement (Carifio & Perla, 2007). However, measurement theory clearly states that is not possible to infer quantities from ordinal attributes (Mitchell, 2009). This implies that, even though the assumption of interval level of measurement in certain cases might work well, this assumption could be highly problematic especially when multivariate normality is not met.

This situation is particularly problematic for classical FA because, when applied to discontinuous data, the correlation among observed variables will depend on the real amount of association and the frequencies of observed responses. Therefore, items with different response frequencies will show artificially attenuated correlations (McDonald, 1999) and this will lead to: i) the emergence of spurious factors due to artificially higher correlations among items with lower response frequencies, increasing the dimensional complexity of the instrument (Berstein & Teng, 1989) and; ii) underestimation of factor loadings of items with asymmetric response frequencies (DiStefano, 2002) which will increase the probability of inaccurate selection.

[2] Despite its popularity, Cronbach's Alpha has been strongly criticized on its general interpretation as an internal consistency and reliability measure and as a method to select items. See for example: Sijtsma, 2009.

Although some solutions have been proposed to this problem, such us creating *item parcels* in order to achieve a larger number of response categories (Hau & Marsh, 2004), IFA is the alternative that better preserve the logic of FA applied to items, treating each of them as independent indicators.

The IFA

During the last 40 years researchers have been developing methods that allows FA to deal with dichotomous and ordinal variables (Christofferson, 1975; Christofferson, 1977; McDonald, 1982; Muthén, 1978; Muthén, 1984; Muthén, 1989). Most of the proposals are based on a three-step methodology.

First, it is assumed that each categorical observed variable is just a rough record of a true underlying continuous and normally distributed variable which is the response that subjects would have give if the instrument were not restricted to a limited number of ordinal alternatives. Therefore, *threshold* (τ) scores are estimated; they represent the value that would have allowed ordinalization of the underlying continuous variables.

Formally, if an item has *m* ordered response categories (1, 2, 3, …, *m*), *z* is the ordinal response given by the subject in the item and z^* is the true underlying score the subject should had; the link between *z* and z^* will be:

$$\textit{If}\quad \tau_{i-1} < z^* < \tau_i \quad \rightarrow \quad z = \mathrm{i} \qquad (1)$$

Where *m*-1 threshold parameters will fragment the scale of z^*:

$$-\infty < \tau_1 < \tau_2 < \ldots < \tau_{m-1} < +\infty \qquad (2)$$

Second, using threshold parameters and bivariate distribution among variables, tetrachoric or polychoric correlations are estimated (in case of dichotomous or polytomous observed variables respectively) to reflect the association among underlying continuous variables.

Finally, a factorial model is adjusted and factor loadings – *lambda* (λ) – for each item are estimated using procedures that minimizes the differences among observed tetra or polychoric correlation matrix and the matrix reproduced by the model.

Three estimation procedures have been advised for this type of data: i) Weighted Least Squares (WLS; Muthén, 1984) which minimizes the residual matrix weighted by the variance-covariance matrix of tetra or polychoric correlations estimates; ii) Diagonally Weighted Least Squares (DWLS; Muthén, du Toit, & Spisic, 1997) which minimizes the residual matrix weighted by the variances of the tetra or polychoric correlation estimates and; iii) Unweighted Least Squares (ULS; Muthén, 1993) which minimizes the unweighted residual matrix.

Previous studies have shown that IFA tend to produce more accurate estimations compared to classical FA (using ML estimation) in dichotomous or ordinal data with few response alternatives and that both procedures tend to converge when five or more response alternatives are available (Beauducel & Herzberg, 2006; DiStefano, 2002;

Dolan, 1994; Holgado–Tello, Chacón–Moscoso, Barbero–García & Vila–Abad, 2010; Rhemtulla, Brosseau-Liard & Savalei, 2012).

However, when using IFA different estimation procedures will have different performances; for example, although WLS have outstanding asymptotic properties, when applied to ordinal data it requires very large samples to evidence them and in small samples it evidences convergence problems and yields bias and unstable parameter estimates (Flora & Curran, 2004).

Regarding ULS and DWLS, information nowadays is scarce and somewhat inconsistent; for example, Rigdon and Ferguson (1991) found no difference among these two procedures, while Forero, Maydeu-Olivares, and Gallardo-Pujol (2009) found that DWLS shows higher convergence rates (CRs) than ULS, but ULS was more robust to the toughest conditions (small samples, asymmetric distributions, and dichotomous responses). However, this case research did not differentiate dichotomous from polytomous data results, hence it is not possible to know which one will produce better results on Likert scales with more than two response categories. Moreover, Yang-Wallentin, Jo¨reskog, and Luo (2010) found slight differences among DWLS and ULS, while Rhemtulla et al. (2012) found that both procedures yielded equivalent CRs and proper solutions, but ULS yielded lower type I error rates.

Thus, considering information cumulated nowadays, it is not possible define which is the best estimation procedure to analyze four-points Likert rating scales because, albeit the majority of research conclude that the number of response categories affect the effectiveness of estimation procedures in different ways (Beauducel & Herzberg, 2006; Dolan, 1994; Savalei & Rhemtulla, 2012), only a few studies have assessed this response format and most of them analyzed either the dichotomous case or an odd number of response categories (i.e.: three or five).

In addition, while WLS is not recognized as an option for estimating IFA parameters, it should be noted that it was developed as an alternative for ML when multivariate normality is not met (for this reason, WLS is also known as asymptotically distribution free), in classical FA based on Pearson correlations (Browne 1984); and its performance has not been tested in the context of ordinal data, namely, assuming that ordinal responses are measured at interval level and directly estimating Pearson correlations among items. Considering that WLS is available in several well-known software programs, such as AMOS (Arbuckle 2010) and LISREL (Jöreskog and Sörbom 2006), its performance is of great interest because it could be a simpler alternative to IFA for applied research.

Therefore, in order to provide orientation for applied research to analyze or validate Likert scales with items of four-points, a Monte Carlo study was conducted to compare the performance of IFA estimation procedures, namely: DWLS and ULS (hereinafter "DWLS$_{PO}$" and "ULS$_{PO}$" to indicate that estimations are made on polychoric correlations) against classical FA procedures, namely: WLS and ML (hereinafter "WLS$_{PE}$" and "ML$_{PE}$" to indicate that estimations are made on Pearson correlations among items) where ML$_{PE}$ will be considered the 'baseline' to compare the potential improvements of the other three.

We expect to contribute providing useful information that clarifies the consequences that the selection of an estimation procedure have for factorial models and help applied researchers to improve their practices to achieve more reliable and valid instruments.

Method

Simulation procedure

Data was generated using the software PRELIS 2 (Jöreskog and Sörbom, 2002) for the following factorial multidimensional model:

$$X_{ij} = \sum_{k=1}^{k} \lambda_{jk} \times F_k + \left(1 - \sum_{k=1}^{k} \lambda_{jk}^2\right)^{0.5} \times e_j \qquad (3)$$

Where X_{ij} is the simulated response of subject i to item j, λ_{ik} is the factor loading of item i in factor k (a simple structure was generated with no cross-loadings, thus $\lambda_{jk}=0$ for item reflecting another factor), F_k are underlying latent factors created from a standard normal distribution (factors could be independent or linearly associated) and e_j is the random measurement error of each item generated from a standard normal distribution.

Given that continuous X_j variables were generated, they were recoded into 4 response categories according to the desired proportion of subjects within each category (this process will be explained later) to represent four-point Likert items.

Simulated conditions

Data was generated for one, two and three dimensional structures as they are commonly found in applied research. For multidimensional conditions, three degrees of correlation among factors were created to represent common situations in applied research, namely: nil (ρ=0), low (ρ=.3) and high (ρ=.6).

In order to increase the probability to get well-specified factors (cf. Fabrigar, Wegener, MacCallum & Strahan, 1999), six items were created for each dimension; thus, 6, 12 and 18 items items were created for unidimensional, bidimensional and threedimensional conditions respectively.

To assess the robustness of each estimation procedure to the quality of the scale, factor loadings were adjusted to represent low (λ=.3) and medium (λ=.6) quality items.

Continuous items were recoded into 4 categories forming distributions with different degrees of asymmetry to assess the performance of each procedure on different the distribution of responses. Thus, three distribution types were created, as shown in Figure 1: Type I items represent symmetric distributions, Type II items represent mild asymmetry (g_1=1.1) and Type II items represent high asymmetry (g_1=1.7) of responses. Higher levels of asymmetry were not considered because they imply lower number of empirically selected alternatives.

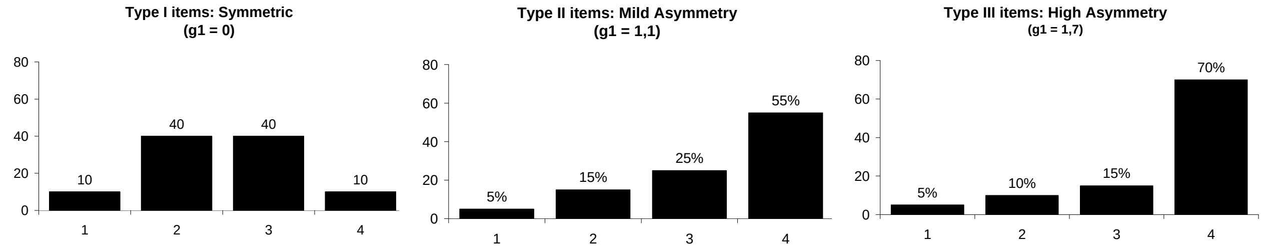

Figure 1. Types of item distribution.

Finally, sample sizes were adjusted to represent variation from small to large sample size commonly employed in applied research, namely: 100, 200, 500, 1000 y 2000 subjects.

Following Harwell, Stone, Hsu and Kirisci (1996) criteria, 500 replications were created for conditions with larger expected variance (i.e.: 100 and 200 subjects conditions or 500 subjects in a three-dimensional structurs with highly asymmetric items) and 250 replications for the rest.

Overall, 210 conditions were adjusted: 180 were multidimensional structures (two and three factors x three levels of correlation among then x two sizes of lambda parameters x three levels of asymmetry x five sample sizes) and 30 were unidimensional structures (two sizes of lambda parameters x three levels of asymmetry x five sample sizes).

Analysis of the effectiveness of estimation procedures

To determine the performance of each estimation procedure (DWLSPO, ULSPO, WLSPE and MLPE) when using four-points Likert type items, a Confirmatory Factor Analysis (CFA) was implemented using LISREL 8.8 (Jöreskog & Sörbom, 2006).

Each procedure was assessed on its capacity to produce unbiased and stable parameter estimates for the factorial model. Hence, we evaluated (i) CR and admissible solutions obtained for each procedure. For simplicity, hereinafter CR and admissible solutions will be referred to simply as CR. Nonconvergent solutions are those for which the estimation procedure does not reach a solution after 250 iterations, while nonadmissible solutions are those yielding values outside range or Heywood cases (e.g., negative variances, standardized l parameters greater than one). As suggested by previous research (Flora and Curran 2004), nonconvergent and nonadmissible solutions will not be considered for further analyses; (ii) relative bias of lambda estimates (RBL), which is the percentage of underestimation or overestimation of real l parameters averaged across replicates within each condition; (iii) standard deviation of lambda estimates (SDL) which is the standard deviation (SD) of l estimates within each condition; (iv) absolute bias of correlation (ABC) which is the magnitude of overestimation or underestimation of the correlation among factors in absolute values averaged across replicates within each condition (relative bias of correlation among factors is discarded because for nil correlation its value is not defined); and (v) standard deviation of correlations (SDC) which is the SD of the correlation estimate among factors averaged across all replicates in each condition.

Data analysis combines multivariante ANOVA tests, effect size estimation using partial eta-squared statistic (η^2_p) and descriptive analyses of results. For descriptive analyses, effect sizes are considered as moderate or large for values exceeding .25 (Ferguson, 2009), achieving less than 80% of valid replicates in each condition is considered unacceptable CR (Forero & Maydeu-Olivares, 2009) and as relevant we will consider bias greater than 5% and for SD those greater than 0.1 (Hoogland & Boomsma, 1998).

Results

Preliminary results showed that neither the complexity of the factorial model (i.e. number of simulated factors) nor the presence and magnitude of correlation among factors had a statistically significant effect explaining the differences among estimation procedures therefore, those results are omitted from this report.

CR

The CR is highly relevant for applied research because it reflects the probability of achieving an acceptable solution when selecting a statistical procedure.

Table 1 shows that estimation procedures considered in this study had no significant effect on the capacity to achieve valid solutions. This result is very interesting since we considered classical FA procedures that currently are not recommended in literature but, when using ordinal data, their CR results were similar to IFA procedures.

Table 1

Analysis of variance of convergence rate

Variable	F (df[a])	η^2_p
EP	1.67 (3)	.01
Size of lambda	554.62 (1)**	.41
Asymmetry	10.37 (2)**	.03
Sample Size	168.92 (4)**	.46
EP x lambda	1.50 (3)	.01
EP x asymmetry	0.01 (6)	.00
EP x sample size	0.25 (12)	.00

Note. EP=Estimation Procedure. F(df)=Fischer-Snedecor F & degrees of freedom. η^2_p=partial eta squared.

[a] Error degrees of freedom=808.

* p<.05; ** p<.01.

Consequently, Figure 2 shows that procedures had similar performances on CRs across the 210 conditions. However, it should be noted that ML$_{PE}$ tends to yield a slightly lower proportion of convergent replicates when compared to other procedures and that WLSPE evidenced better results compared to ML$_{PE}$. Considering that no significant interaction effect was found among estimation procedures and sample size (see Table 1), this result implies that the convergence of WLS$_{PE}$ is not affected by small sample sizes and seems to contradict previous studies using WLS with tetra or polychoric correlation matrices—WLS$_{PO}$—(DiStefano 2002; Flora and Curran 2004); therefore, to confirm that this unexpected result was correct and not the effect of our simulation procedure, we decided to test WLS$_{PO}$ in our data and, as expected, it yielded lower CRs than other procedures for samples lesser than 500 subjects, which was not observed for WLSpe.

Variables that evidenced a significant and meaningful effect size on CR were: i) the magnitude of lambda parameters, where low quality of the items (λ=.3) yielded unacceptable CR (69.7%) which experienced an important improvement (reaching almost perfect CR) when the quality of items was higher (λ=.6) and; ii) the sample size where unacceptable CR was found for samples of 100 subjects (57.8%) but improved to a satisfactory level (95.6%) for samples of 500 and to optimal (99.2%) for samples of 1000 subjects. Overall and regardless of the estimation procedure, acceptable CR can be achieved for sample sizes greater or equal to 500 subjects if the quality of the items is low, however 100 subjects are enough to estimate a model when the quality of the items is high (λ=.6).

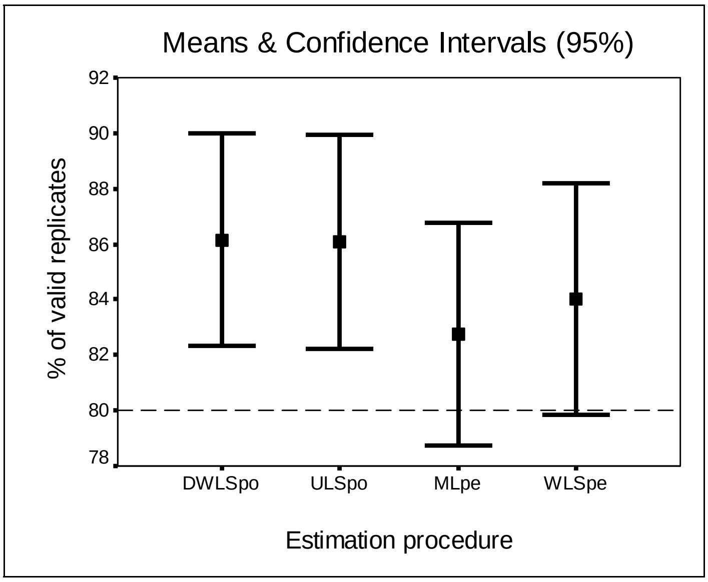

Figure 2. Means and confidence intervals of valid replicates by estimation procedure

Relative Bias of Lambdas

Lambda parameters are a key result for Likert scales because only if correct factor loadings among the items and its factors ensures correct elimination of less-informative items to build a uni or multidimensional scale.

Table 2

Analysis of variance of relative bias of lambdas

Variable	F (df^a)	η^2_p
EP	385.92 (3)**	.59
Size of lambda	174.10 (1)**	.18
Asymmetry	54.49 (2)**	.12
Sample Size	257.76 (4)**	.56
EP x lambda	3.70 (3)*	.01
EP x asymmetry	34.35 (6)**	.20
EP x sample size	33.04 (12)**	.33

Note. EP=Estimation Procedure. F(df)=Fischer-Snedecor F & degrees of freedom. η^2_p=partial eta squared.

[a.] Error degrees of freedom=808.

* p<.05; ** p<.01.

As shown in Table 2, estimation procedures had a statistically significant and large effect on RBL. To examine this effect in detail, Figure 3 shows the performance of each procedure. There we can appreciate that DWLSPO and ULSPO yielded relatively accurate results (somewhat better in ULSPO) with a slight overestimation of the true parameter. Surprisingly, WLSPE performed reasonable fine evidencing low underestimation bias (less than 5%), which is just a bit larger than the bias evidenced by IFA procedures. Accordingly, unlike MLPE which yielded biased parameter estimates, WLSPE could be considered an alternative procedure to achieve relatively unbiased lambda parameter estimates for Likert type items. However, the magnitude of the interaction effects among estimation procedures and samples sizes as well as item asymmetry (see Table 2) evidence that situation could be more complex.

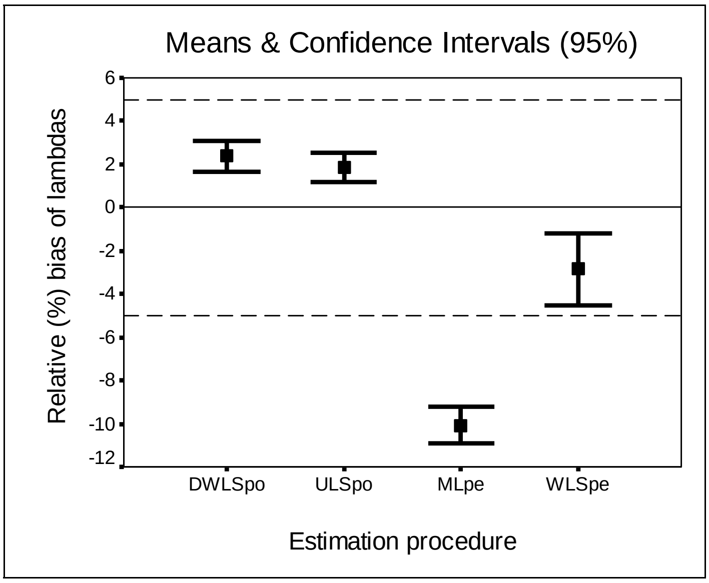

Figure 3. Means and confidence intervals of relative bias of lambdas by estimation procedure

In fact, as shown in Figure 4, WLSPE achieved equivalent results to ULSPO and WLSPO for symmetric items and samples of 200 subjects. Smaller samples tend to yield unacceptable overestimations and contrastingly, samples greater or equal to 500 subject yielded unacceptable underestimated parameter estimates. Moreover, a detailed analysis of WLSPE allowed us to determine that its bias near cero in samples of 200 subjects is just the spurious result of the compensation of bias with opposite signs. Thus, for samples of 200 subjects, WLSPE overestimate the lambda parameters when quality of the item is low (λ=.3) and this bias tend to decrease as the asymmetry of items increases, while for good quality of the items (λ=.6) it overestimates the true parameter and this bias tend to increase as asymmetry of the items increases. Therefore, WLSPE is not a reliable procedure to estimate factor loadings in any case when Likert type items are considered.

In addition, by observing Figure 4, we can conclude that ULSPO and DWLSPO procedures showed similar performances (ULSPO seems slightly better), both are relatively robust to items' asymmetry and that samples of 200 subjects seems to be enough to reach acceptable results, although 500 subjects are required to get optimal accuracy.

In contrast, MLPE tend to underestimate lambda parameters in all conditions, especially when items are not symmetric and, surprisingly, increasing sample size only allows the stabilization of the underestimation bias around 10% but does not solve the problem.

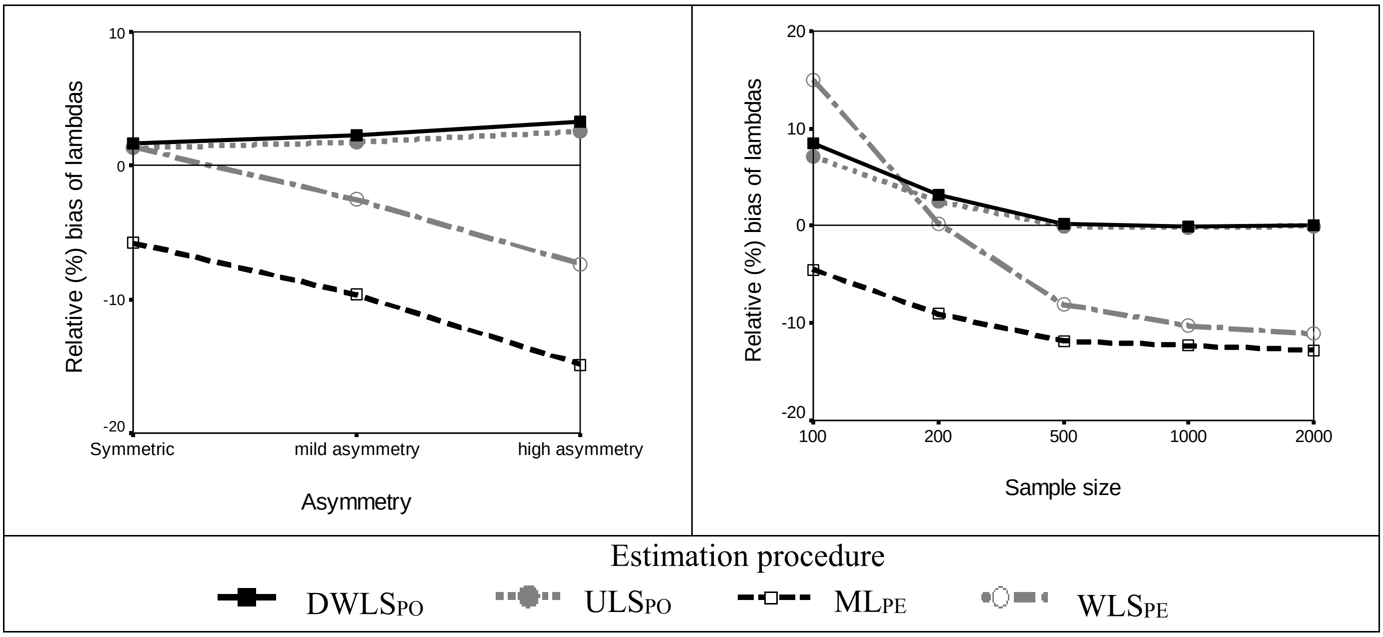

Figure 4. Relative bias of lambdas by asymmetry and sample size by estimation procedure

Standard Deviation of Lambdas

The SDL is a relevant indicator of the stability of parameter estimates achieved by a statistical procedure. Therefore, large SD values evidence that an estimation procedure yields very different parameter estimates when facing equivalent data and its estimations are not precise; in contrast, those evidencing a small standard deviation will be more precise estimating the parameter.

As shown in Table 3, estimation procedures had a statistically significant effect on the stability of parameter estimates; however its effect size is almost irrelevant. Hence, estimation procedures are not different in their degrees of instability to estimate the parameter and descriptive analysis showed that all procedures presented results within the acceptable range.

Table 3

Analysis of variance of standard deviation of lambdas estimation

Variable	F (df^a)	η^2_p
EP	4.35 (3)**	.02
Size of lambda	3204.52 (1)**	.80
Asymmetry	162.94 (2)**	.29
Sample Size	2431.55 (4)**	.92
EP x lambda	1.37 (3)	.01
EP x asymmetry	0.43 (6)	.03
EP x sample size	2.27 (12)**	.03

Note. EP=Estimation Procedure. F(df)=Fischer-Snedecor F & degrees of freedom. η^2_p=partial eta squared.

$^{a.}$ Error degrees of freedom=808.

* p<.05; ** p<.01.

Variables having at least a moderate effect on instability of parameter estimates are the asymmetry of items, the magnitude of lambda parameters and sample sizes. However, differences with regard to asymmetry of the item are negligible (e.g. for highly asymmetric items SD=0.09 while for symmetric items SD=0.07). Regarding to the magnitude of lambda parameters, when the quality of the items was low (λ=.3) parameters are estimated right at the upper limit of acceptable instability (SD=0.11), while for items with higher quality (λ=.6) parameter estimates are stable (SD=0.06). Finally, for samples equal or lower than 100 subjects, large instability of estimates is observed (SD=0.15) and it tend to reach completely acceptable values for samples of 500 or larger (SD=0.07)

Absolute Bias of Correlations

Improper estimation of correlation among factors can lead to an erroneous representation of the dimensional structure of the construct intended to measure. Hence, estimation procedures should be examined on this matter.

Table 4

Analysis of variance of bias of factor correlation estimation

Variable	F (df^a)	η^2_p
EP	27.04 (3)**	.11
Size of lambda	4.24 (1)*	.01
Asymmetry	6.89 (2)**	.02
Sample Size	2.96 (4)*	.02
EP x lambda	8.42 (3)**	.04
EP x asymmetry	1.47 (6)	.01
EP x sample size	5.75 (12)**	.09

Note. EP=Estimation Procedure. F(df)=Fischer-Snedecor F & degrees of freedom. η^2_p=partial eta squared.

$^{a.}$ Error degrees of freedom=808.

* p<.05; ** p<.01.

Table 4 shows that a statistically significant relation was found among the estimation procedures and ABC; although its effect size was mild, empirical absolute bias were within the range -0.02 and 0.02, hence only slight differences were found since ML$_{PE}$ yielded negative values and WLS$_{PE}$ and IFA procedures (DWLS$_{PO}$ and ULS$_{PO}$) yielded positive values.

Significant effects were found for several variables in Table 4, however the single relevant effect was a two -way interaction among the estimation procedures and sample size. Figure 5 allows observing that this effect was basically a slight bias for small samples sizes which decreases as sample size increases, where ML$_{PE}$ tends to underestimate the correlation while WLS$_{PE}$ tend to overestimate it and DWLS$_{PO}$ and ULS$_{PO}$ are robust to small sample sizes.

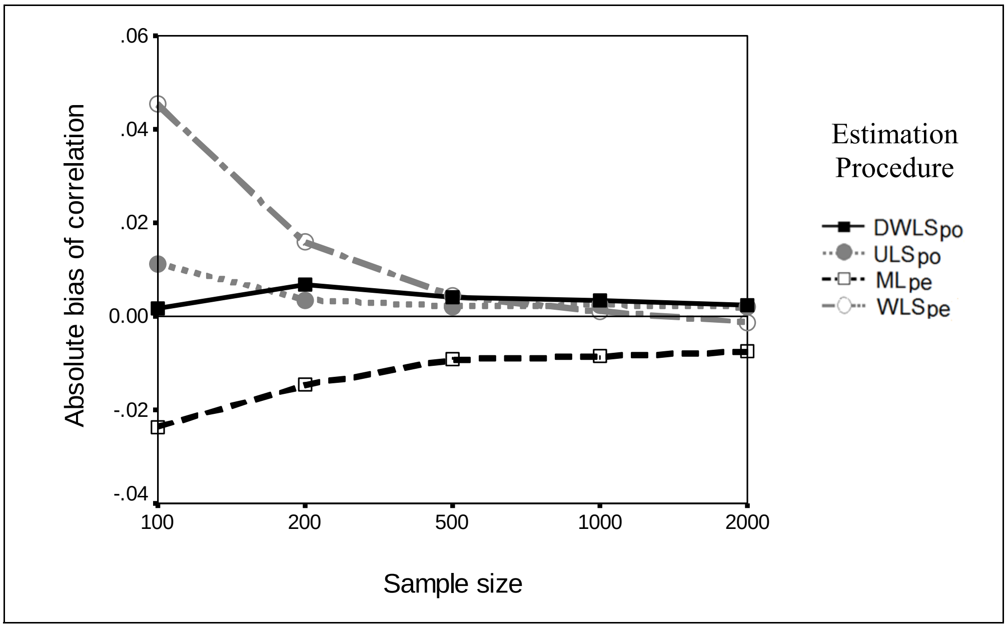

Figure 5. Absolute bias of correlation estimate by sample size by estimation procedure

Standard Deviation of Correlations

Table 5 allows determining that no statistically significant or meaningful difference was found between estimation procedures either when treated as main or two-way interaction effects. In fact, all estimation procedures tend to estimate the correlation among factors with the same degree of instability which was above the acceptable level (i.e. SD>0.1).

Table 5

Analysis of variance of standard deviation of factor correlation estimation

Variable	F (df^a)	η^2_p
EP	0.38 (3)	.00
Size of lambda	1669.83 (1)**	.71
Asymmetry	30.46 (2)**	.08
Sample Size	614.02 (4)**	.78
EP x lambda	1.19 (3)	.01
EP x asymmetry	0.18 (6)	.00
EP x sample size	0.58 (12)	.01

Note. EP=Estimation Procedure. F(df)=Fischer-Snedecor F & degrees of freedom. η^2_p=partial eta squared.

$^{a.}$ Error degrees of freedom=808.

* p<.05; ** p<.01.

In addition, Table 5 shows that no interaction effect was found among procedures and other independent variables, which indicate that no procedure outperforms the others in any situation.

Only two statistically significant and relevant effects were found for SDC: the magnitude of lambda parameters and the sample size. As shown in previous analyses, best results were found for items of good quality and poorer for those with lower quality (e.g. when λ=.3 SDC=0.18 and for λ=.6 SDC=0.08), while heterogeneity of estimations was larger for smaller samples than larger ones (e.g. when n=100 SDC=0.23 and for n=2000 SDC=0.06).

Overall, results shows that to reach an acceptable level of heterogeneity (SDC<0.1) samples of 2000 subjects are required when the quality of the items is low (λ=.3) while a sample of 500 subject could be enough if the quality of the items is medium (λ=.6).

Conclusions

This study aimed to determine the best procedure to analyze factorial models of four-points Likert type items on uni and multidimensional scenarios. We compared IFA procedures against classical FA procedures and overall, we found that IFA procedures outperformed the classical perspective.

According to our findings, although all procedures evidenced similar capacity to produce valid solutions and stable lambda and correlation parameter estimates, ULSPO and DWLSPO yielded remarkable lower bias in both parameter estimates and were robust to the toughest scenarios: asymmetric item distributions, low quality of items (λ=.3) and small sample sizes.

It has been clearly confirmed that employing classical estimation procedures in ordinal data with four response alternatives is inappropriate and counterproductive. This is consistent with previous research evidencing underestimation of key parameters in the model when classical FA procedures are employed (Beauducel & Herzberg, 2006; DiStefano, 2002; Dolan, 1994; Holgado–Tello, Chacón–Moscoso, Barbero–García & Vila–Abad, 2010; Rhemtulla, Brosseau-Liard & Savalei, 2012).

However on this matter, two points must be highlighted: i) first, that using classical FA with WLS estimation is never a viable option for ordinal data given results presented here using Pearson correlation matrices and considering its poor results on tetra and polychoric correlation matrices reported in previous research (Flora & Curran, 2004) and; ii) secondly, that the poor performance of MLPE could be due to the employment of product-moment Pearson correlations and not to the ML estimation procedure itself, because several studies have shown that using ML estimation on tetra or polychoric correlation matrices, yield fairly similar results to DWLSPO and ULSPO, specially in large samples (Dolan, 1994; Rigdon & Ferguson, 1991; Yang-Wallentin, Jöreskog & Luo, 2010).

According to our findings we must state that IFA should be considered the standard procedure to analyze four-point ordinal items because its lower bias guarantees a more accurate selection of items for the final scale and thus, the generation of more valid and reliable instruments.

In addition, when comparing the relative quality of IFA procedures (DWLSPO and ULSPO), there are hardly any differences. In fact, although ULSPO seems better than DWLSPO, this advantage is too small to make any meaningful differences for applied research. These findings are consistent with those reported by Rigdon and Ferguson (1991) and Yang-Wallentin et al. (2010) and somewhat divergent from those reported by Forero et al. (2009), as the advantage in favor of ULSPO they reported could be due to the dichotomous items they considered and the lack of separation among results could have overlooked the dilution of this effect for a larger number of response alternatives. Therefore, applied researchers can select ULSPO or DWLSPO to analyze multidimensional Likert scales.

Our main advice for applied research is facilitated because IFA procedures are widely implemented for exploratory or confirmatory purposes in several well-known software such us: Factor (Lorenzo-Seva & Ferrando, 2006) which allows exploratory IFA, LISREL (Jöreskog & Sörbom, 2006) which allows confirmatory IFA and M-Plus (Muthén & Muthén, 2011) which allows exploratory and confirmatory IFA.

In addition to our main research questions, our inquiry was also concerned about the minimal requirements to employ IFA procedures on four-point Likert type items. Concerning to this matter, our research allow us to sustain that if a researcher expect that the quality of the items in the scale will be low (λ=.3), a sample of 500 subjects might be selected in order to ensure a large probability to achieve admissible results (i.e. a convergent solution and with no Heywood cases) and relatively unbiased and stable estimation of key parameters in the model. Evidently, if the items are suspected to reflect the latent construct in a better fashion (λ=.6), accurate estimations can be reached for small samples (200 or even 100 subjects) if items distributions are symmetric or mildly asymmetric.

To sum up, these research results allow us to sustain that classical FA was not robust to the discontinuity of data represented by the case of four-point Likert rating scales; therefore, its employment must be strongly discouraged for this particular scenario, although it could work in other scenarios with a larger number of response alternatives (Beauducel & Herzberg, 2006; Dolan, 1994; Rhemtulla, Brosseau-Liard & Savalei, 2012).

Although these findings and directions are highly interesting and promising for applied research, at least three important limitations of this study need to be addressed to avoid inferences beyond its limits.

First, this research only considered confirmatory IFA models, therefore, further research is still needed to evaluate if these findings could be extended to exploratory models.

Second, we only considered four-point Likert type items which, to some extent, can not be completely extrapolated to higher or lower number of response categories. Given that, as the number of response categories increases, different procedures tend to yield better results and evidence similar performances (Beauducel & Herzberg, 2006; Dolan, 1994; Savalei & Rhemtulla, 2012), a careful research and analysis of three-point Likert scales scenario still needed and could be highly interesting considering that dichotomous case have been widely investigated.

Finally, this research only considered highly 'ideal' situations (e.g. homogeneous quality of the items, no cross-loadings, no missing data). Therefore, further examination of estimation procedures in more complex situations closest to applied research has its merits, for example: heterogeneous quality of items, weak and strong mixed factors and different number of items per factor, among others.

Referencias

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43(4),* 561-573.

Arbuckle, J.L. (2010). Amos (Version 19.0) [Computer Program]. Chicago: SPSS, An IBM Company.

Beauducel, A., & Herzberg, P.Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling:* a *Multidisciplinary Journal, 13(2),* 186-203.

Bernstein, I., & Teng, G. (1989). Factoring items and factoring scales are different: spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105(3),* 467-477.

Boote, A.S. (1981). Reliability testing of psychographic scales: Five-point or seven-point? Anchored or labeled? *Journal of Advertising Research, 21,* 53-60.

Brown, G., Wilding, R.E., & Coulter, R.L. (1991). Customer evaluation of retail salespeople using the SOCO scale: A replication, extension, and application. *Journal of the Academy of Marketing Science, 9,* 347-351.

Browne, M.W. (1984). Asymptotic distribution free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 127–141.

Carifio, J., & Perla, R.J. (2007). Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes. *Journal of Social Sciences, 3(3),* 106-116.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18,* 205-215.

Cox III, E.P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of marketing research, 17,* 407-422.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40(1)*, 5–32.

Christoffersson, A. (1977). Two-step weighted least squares factor analysis of dichotomized variables. *Psychometrika, 42(3)*, 433–438.

DeVellis, R. (1991). *Scale development, theory and applications*. Newbury Park: Sage.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling:* a *Multidisciplinary Journal, 9,* 327-346.

Dolan, C.V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: a comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47*, 309–326.

Fabrigar, L.R., Wegener, D.T., MacCallum R.C. & Strahan E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4(3)*, 272-299.

Ferguson, C.J. (2009). An effect size primer: a guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40(5)*, 532-538.

Flora, D.B. y Curran, P.J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9(4),* 466-491.

Forero, C.G., Maydeu-Olivares, A. y Gallardo-Pujol, D. (2009). Factor Analysis with Ordinal Indicators: A Monte Carlo Study Comparing DWLS and ULS Estimation. *Structural Equation Modeling:* a *Multidisciplinary Journal, 16*, 625–641.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87*, 564-567.

Garland, R. (1991). The mid-point on a rating scale: Is it desirable? *Marketing Bulletin, 2(1)*, 66-70.

Garner, W.R. (1960). Rating scales, discriminability and information transmission. *Psychological Review, 67,* 343-352.

González-Romá, V., & Espejo, B. (2003). Testing the middle response categories "Not sure", " In between" and "?" in polytomous items. *Psicothema, 15(2),* 278-284.

Hancock, G.R., & Klockars, A.J. (1991). The effect of scale manipulations on validity: Targeting frequency rating scales for anticipated performance levels. *Applied Ergonomics, 22,* 147-154.

Harwell, M., Stone, C.A., Shu, T.-C., & Kirisci, L. (1996). Montecarlo studies in item response theory. *Applied Psychological Measurement, 20(2),* 101-125.

Hau, K-T., & March, H. (2004). The use of items parcels in structural equation modelling: Non-normal data and small sample sizes. *British Journal of Mathematical Statistical Psychology, 57*, 327-351.

Holgado–Tello, F.P., Chacón–Moscoso, S., Barbero–García, I., & Vila–Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity, 44(1)*, 153-166.

Hoogland, J.J., & Boomsma, A. (1998). Robustness studies in covariance structural modeling: an overview and a meta-analysis. *Sociological Methods & Research, 26(3)*, 329–367.

Jamieson, S. (2004). Likert scales: how to (ab)use them. *Medical Education, 38,* 1212-1218.

Jöreskog, K.G., & Sörbom, D. (2002). *PRELIS 2: User's reference guide.* Lincolnwood: Scientific Software International, Inc.

Jöreskog K.G. & Sörbom, D. (2006). *LISREL 8.8: User's reference guide.* Lincolnwood: Scientific Software International, Inc.

Kulas, J.T., Stachowski, A.A., & Haynes, B.A. (2008). Middle response functioning in Likert-responses to personality items. *Journal of Business and Psychology, 22(3)*, 251-259.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22(140)*, 44-55.

Likert, R., Roslow, S., & Murphy, G. (1934). A simple and reliable method os scoring Thurstone Attitudes Scales. *The Journal of Social Psychology, 5(2)*, 228-238.

Loken, B., Pirie, P., Virnig, K.A., Hinkle, R. L., & Salmon, C. T. (1987). The use of 0-10 scales in telephone surveys. *Journal of the Market Research Society, 29*(3), 353-362.

Lord, F.M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8,* 750-751.

Lorenzo-Seva, U., y Ferrando, P.J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. Behavioral Research Methods, Instruments and Computers, 38(1), 88-91.

Matell, M.S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study 1: Reliability and validity. *Educational and Psychological Measurement, 31,* 657-674.

McDonald, R.P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6(4)*, 379–396.

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Mitchell, J. (2009). The psychometricians' fallacy: too clever by half? *British Journal of Mathematical Statistical Psychology, 62*, 41-55.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63(2),* 81-97.

Muthén, B.O. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43(4),* 551–560.

Muthén, B.O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variables indicators. *Psychometrika, 49(1)*, 115-132.

Muthén, B.O. (1989). Dichotomous factor analysis of symptom data. *Sociological Methods & Research, 18(1)*, 19-65.

Muthén, B.O. (1993). Goodness of fit with categorical and other nonnormal variables. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 205–234). Newbury Park, CA: Sage.

Muthén, L.K. & Muthén, B.O. (2011). *Mplus Version 6.11*. Los Angeles: Author.

Muthén, B.O., du Toit, S.H.C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript. Retrieved from http://pages.gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education, 15(5),* 625-632.

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

Preston, C.C., & Colman, A.M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104,* 1-15.

Raaijmakers, Q.A., van Hoof, A., Hart, H., Verbogt, T.F.M.A., & Wollebergh, W.A.M. (2000). Adolescents' midpoint response on Likert-tyep scale items: Neutral or missing values? *International Journal of Public Opinion Research, 12(2),* 208-216.

Rhemtulla, M., Brosseau-Liard, P.E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological methods, 17(3)*, 354-373.

Rigdon, E.E, & Ferguson Jr, C.E. (1991). The performance of the polichoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research, 28*, 491-497.

Savalei, V., & Rhemtulla, M. (2012). The performance of robust test statistics with categorical data. *British Journal of Mathematical and Statistical Psychology.* Advance on line publication.

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

Spector, P.E. (1992). *Summating rating scale construction: an introduction.* Newbury Park: Sage.

Velleman, P.F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician, 47,* 65-72.

Wirth, R.J., & Edwards, M.C. (2007). Item factor analysis: current approaches and future directions. *Psychological Methods, 12(1)*, 58-79.

Yang-Wallentin, F., Jöreskog, K., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling:* a *Multidisciplinary Journal, 17,* 392–423.