Article

Sociological Methods & Research
2016, Vol. 45(1) 109-133
© The Author(s) 2015
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0049124114566716
smr.sagepub.com
SAGE

Developing Multidimensional Likert Scales Using Item Factor Analysis: The Case of Four-point Items

Rodrigo A. Asún[1], Karina Rdz-Navarro[1], and Jesús M. Alvarado[2]

Abstract

This study compares the performance of two approaches in analysing four-point 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.

[1] Facultad de Ciencias Sociales, Universidad de Chile, Ñuñoa, Santiago, Chile
[2] Facultad de Psicología, Universidad Complutense de Madrid, Madrid, Spain

Corresponding Author:
Rodrigo Asún, Facultad de Ciencias Sociales, Universidad de Chile, Ignacio Carrera Pinto 1045, Ñuñoa, Santiago, Chile.
Email: rasun@uchile.cl

Keywords

Likert scales, item factor analysis, polychoric correlation, four-point items, classical factor analysis

The Likert Rating Scale (Likert 1932; Likert, Roslow, and Murphy 1934) is a simple procedure for generating measurement instruments which is widely used by social scientists to measure a variety of latent constructs, and meticulous statistical procedures have therefore been developed to design and validate these scales (DeVellis 1991; Spector 1992). However, most of these ignore the ordinal nature of observed responses and assume the presence of continuous observed variables measured at interval level. Although there is still much debate over the robustness to ordinal data of parametric statistical techniques for developing Likert Scales (Carifio and Perla 2007; Jamieson 2004; Norman 2010), evidence shows that, under relatively common circumstances, classical factor analysis (FA) yields inaccurate results characterizing the internal structure of the scale or selecting the most informative items within each factor (Bernstein and Teng 1989; DiStefano 2002; Holgado–Tello et al. 2010). Fortunately, item factor analysis (IFA) provides an alternative that avoids these problems (Wirth and 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 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.

Developing Likert Scales Using Four-point Items

The Number of Response Categories on Likert Items

Since Rensis Likert first suggested the scaling procedure that now bears his name, there has been considerable debate over the optimum number of categories to present to the subjects answering the questionnaire. Interestingly, the evidence found in the literature supports highly contrasting positions: Some researchers suggest that larger numbers of response categories achieve higher levels of reliability (Garner 1960) and validity (Hancock and Klockars 1991; Loken et al. 1987); while others suggest that the number of response categories is not related to the reliability of the scale (Boote 1981; Brown, Wilding, and Coulter 1991) or its validity (Chang 1994; Matell and Jacoby 1971). Overall, the evidence tends 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 answering the questionnaire (Preston and 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 four to seven response categories, and five or seven points are the most common format used in applied research (Cox 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 be measured (Preston and Colman 2000).

Nevertheless, the intermediate category may affect the validity of results because (i) subjects could use this category for other reasons than having an intermediate opinion, for example, the subject does not have an opinion, does not want to express his or her true opinion, does not understand the question, is facing a "not applicable" question, among others (Kulas, Stachowski, and Haynes 2008; Raaijmakers et al. 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 and Espejo 2003); and (iv) on certain occasions, the information contributed by an intermediate category is not informative (Andrich 1978).

Therefore, a four-point response format is highly attractive when social desirability is suspected to affect the construct intended to be measured, 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, observed items will no longer be similar to interval level of measurement variables; therefore, statistical analysis, such as classical FA, is likely to yield inaccurate results.

Likert Scales and Classical Factor Analysis

The FA has been widely acknowledged as a central procedure for developing 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 that maximize the internal consistency of the scale could be selected using either Pearson correlations among the item and total scale and/or Cronbach's α (DeVellis 1991), which remains popular despite the criticism it has received (Sijtsma 2009). FA could then 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 by 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 that are measured at interval level and the estimation procedures frequently employed in FA, such as maximum likelihood estimation (ML), assume multivariate normal distribution of observed responses. In contrast, items in a Likert scale are coded using a procedure known as *integer scoring* (González-Romá and 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 and Wilkinson 1993), that statistical analyses are robust to ordinal data (Norman 2010) and, furthermore, that Likert scales produce interval level of measurement (Carifio and Perla 2007). However, measurement theory clearly states that is not possible to infer quantities from ordinal attributes (Michell 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 (Bernstein and Teng 1989) and (ii) underestimation of factor loadings of items with asymmetric response frequencies (DiStefano 2002), which will increase the probability of inaccurate item selection.

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

The IFA

Over the last 40 years, researchers have been developing methods allowing FA to deal with dichotomous and ordinal variables (Christoffersson 1975, 1977; McDonald 1982; Muthén 1978, 1984, 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—the response that subjects would have given if the instrument had not been 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 $(1, 2, 3, \ldots, m)$, z is the ordinal response given by the subject in the item and z^* is the true underlying score the subject should have; the link between z and z^* will be:

$$\textit{If } \tau_{i-1} < z^* < \tau_i \rightarrow z = i. \tag{1}$$

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

$$-\infty < \tau_1 < \tau_2 < \ldots < \tau_{m-1} < +\infty. \tag{2}$$

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 minimize 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, and Spisic 1997) thats minimizes the residual matrix weighted by the variances of the tetra or polychoric correlation estimates; and (iii) unweighted least squares (ULS; Muthén 1993) that minimizes the unweighted residual matrix.

Previous studies have shown that IFA tends 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 and Herzberg 2006; DiStefano 2002; Dolan 1994; Holgado–Tello et al. 2010; Rhemtulla, Brosseau-Liard, and Savalei 2012).

However, when using IFA, different estimation procedures will have different performances; for example, although WLS has outstanding asymptotic properties, when applied to ordinal data it requires very large samples and in small samples it encounters convergence problems and yields bias and unstable parameter estimates (Flora and 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, Jöreskog, 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 the amount of information cumulated nowadays, it is not possible to define which is the best estimation procedure to analyze four-point Likert rating scales because, although the majority of research concludes that the number of response categories affects the effectiveness of estimation procedures in different ways (Beauducel and Herzberg 2006; Dolan 1994; Savalei and Rhemtulla 2013), only a few studies have assessed this response format and most of these looked at 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 guidelines 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)—with 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" for comparing the potential improvements of the other three.

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

Method

Simulation Procedure

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

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

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 four 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 were 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 ($\rho = 0$), low ($\rho = .3$), and high ($\rho = .6$).

In order to increase the probability of obtaining well-specified factors (Fabrigar et al. 1999), six items were created for each dimension; thus, 6, 12, and 18 items were created for unidimensional, bidimensional, and three-dimensional conditions, respectively.

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

Continuous items were recoded into four categories forming distributions with different degrees of asymmetry to assess the performance of each procedure on the different 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 a lower number of empirically selected alternatives.

Finally, sample sizes were adjusted to represent variation from small to large sample sizes commonly employed in applied research, namely, 100, 200, 500, 1,000, and 2,000 subjects.

Following Harwell et al. (1996) criteria, 500 replications were created for conditions with larger expected variance (i.e., 100 and 200 subjects

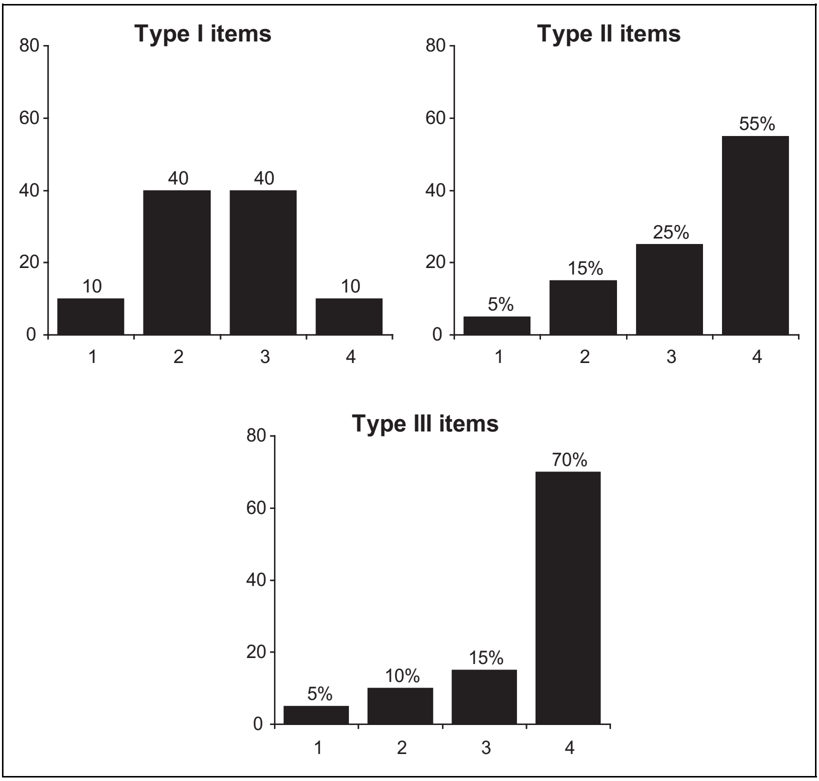

Figure 1. Types of item distribution.

conditions or 500 subjects in a three-dimensional structure with highly asymmetric items) and 250 replications for the rest.

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

Analysis of the Effectiveness of Estimation Procedures

To determine the performance of each estimation procedure (DWLS_{PO}, ULS_{PO}, WLS_{PE}, and ML_{PE}) when using four-point Likert type items, a confirmatory factor analysis was implemented using LISREL 8.8 (Jöreskog and 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 λ 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 λ parameters averaged across replicates within each condition; (iii) standard deviation of lambda estimates (SDL) which is the standard deviation (SD) of λ 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 multivariate analysis of variance tests, effect size estimation using partial eta-square statistic (η_p^2) 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 and Maydeu-Olivares 2009) and we will consider as relevant any bias greater than 5% and for SD those greater than 0.1 (Hoogland and 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.

Table 1. Analysis of Variance (ANOVA) of Convergence Rate (CR).

Variable	F (df [a])	η_p^2
EP	1.67 (3)	.01
Size of λ	554.62 (1)**	.41
Asymmetry	10.37 (2)**	.03
Sample size	168.92 (4)**	.46
EP × λ	1.50 (3)	.01
EP × Asymmetry	.01 (6)	.00
EP × Sample size	.25 (12)	.00

Note: EP = estimation procedure; F (df) = Fischer–Snedecor F and degrees of freedom; η_p^2 = partial eta squared.
[a]Error degrees of freedom = 808.
*p < .05. **p < .01.

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 the literature; however, when using ordinal data, their CR results were similar to IFA procedures.

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 WLS_{PE} 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 WLS_{PE}.

Variables that showed a significant and meaningful effect size on CR were (i) the magnitude of λ parameters, where low item quality (λ = .3) yielded unacceptable CR (69.7%), which significantly improved (to almost

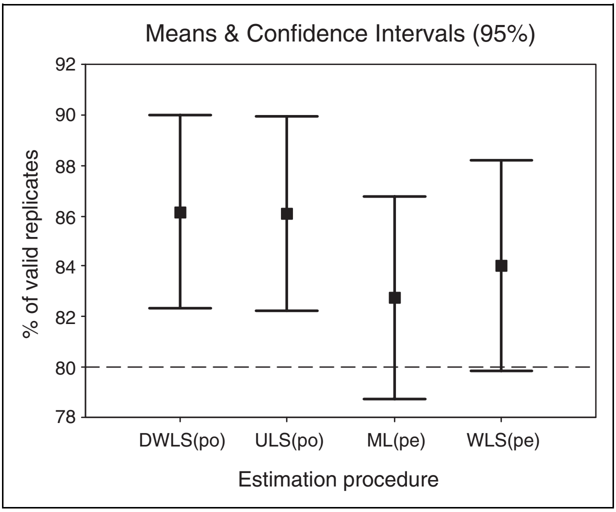

Figure 2. Means and confidence intervals of valid replicates by estimation procedure.

perfect CR) when the quality of items was higher ($\lambda = .6$) 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 level (99.2%) for samples of 1,000 subjects. Overall, and regardless of the estimation procedure, acceptable CR can be achieved for sample sizes greater than or equal to 500 subjects if the quality of the items is low; that said, 100 subjects are enough to estimate a model when the quality of the items is high ($\lambda = .6$).

Relative Bias of λs

Lambdas (λ) parameters are a key result for Likert scales because only correct factor loadings among the items and its factors ensure correct elimination of less informative items to build a uni- or multidimensional scale.

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. Here we can appreciate that DWLS_{PO} and ULS_{PO} yielded relatively accurate results (somewhat better in ULS_{PO}) with a slight overestimation of the true parameter. Surprisingly, WLS_{PE} performed reasonably well, evidencing low underestimation bias (less than

Table 2. Analysis of Variance (ANOVA) of Relative Bias of λs.

Variable	F (df ^{a})	η_p^2
EP	385.92 (3)**	.59
Size of λ	174.10 (1)**	.18
Asymmetry	54.49 (2)**	.12
Sample size	257.76 (4)**	.56
EP × λ	3.70 (3)*	.01
EP × Asymmetry	34.35 (6)**	.20
EP × Sample size	33.04 (12)**	.33

Note: EP = estimation procedure; F (df) = Fischer–Snedecor F and degrees of freedom; η_p^2 = partial eta squared.

[a]Error degrees of freedom = 808.

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

5%), which is only slightly larger than the bias evidenced by IFA procedures. Accordingly, unlike $\mathrm{ML_{PE}}$ that yielded biased parameter estimates, $\mathrm{WLS_{PE}}$ could be considered an alternative procedure to achieve relatively unbiased λ 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), show that the situation could be more complex.

In fact, as shown in Figure 4, $\mathrm{WLS_{PE}}$ achieved equivalent results to $\mathrm{ULS_{PO}}$ and $\mathrm{WLS_{PO}}$ for symmetric items and samples of 200 subjects. Smaller samples tend to yield unacceptable overestimations and, contrastingly, samples greater than or equal to 500 subject yielded unacceptable underestimated parameter estimates. Moreover, through a visual inspection of scatter plots of $\mathrm{WLS_{PE}}$, we were able to determine that its bias near zero in samples of 200 subjects is the result of an unstable performance where large biases of opposite signs are compensated. Thus, for samples of 200 subjects, $\mathrm{WLS_{PE}}$ overestimates the λ parameters when item quality is low ($\lambda = .3$), and this bias tends to decrease as the asymmetry of items increases; while for high item quality ($\lambda = .6$) it overestimates the true parameter and this bias tends to increase as item asymmetry increases. Therefore, $\mathrm{WLS_{PE}}$ is not a reliable procedure for estimating factor loadings in any case when Likert-type items are considered.

In addition, by observing Figure 4, we can conclude that $\mathrm{ULS_{PO}}$ and $\mathrm{DWLS_{PO}}$ procedures showed similar performances ($\mathrm{ULS_{PO}}$ seems slightly better), both are relatively robust to items' asymmetry and samples of 200 subjects seem to be enough to reach acceptable results—although 500 subjects are required to get optimum accuracy.

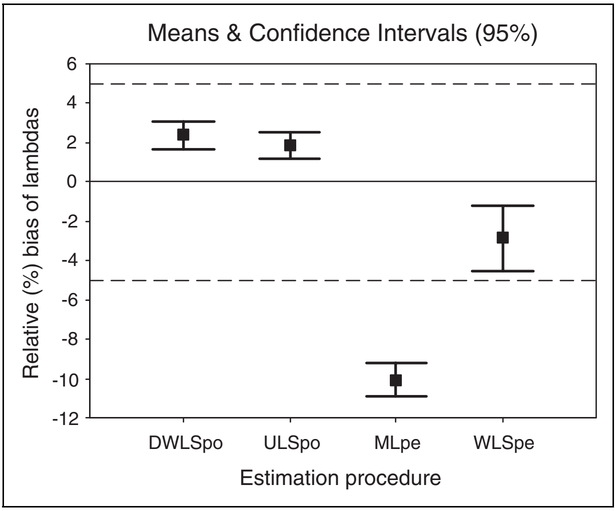

Figure 3. Means and confidence intervals of relative bias of λs by estimation procedure.

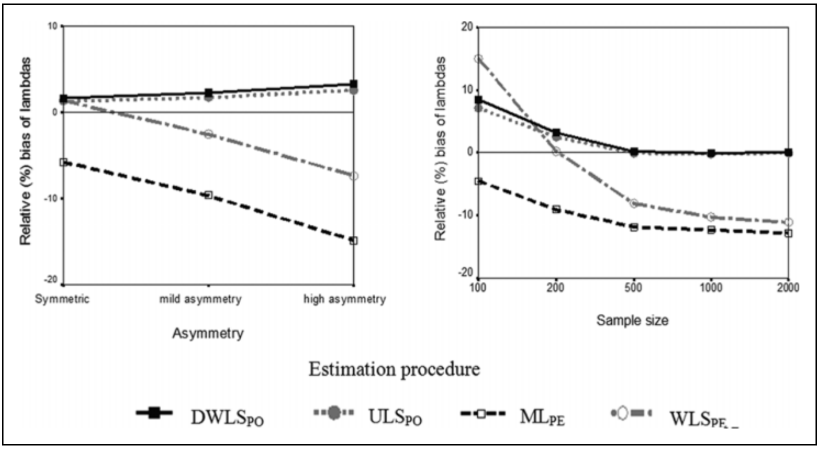

Figure 4. Relative bias of λs by asymmetry and sample size by estimation procedure.

Table 3. Analysis of Variance (ANOVA) of Standard Deviation (SD) of λs Estimation.

Variable	F (df [a])	η_p^2
EP	4.35 (3)**	.02
Size of λ	3204.52 (1)**	.80
Asymmetry	162.94 (2)**	.29
Sample size	2431.55 (4)**	.92
EP × λ	1.37 (3)	.01
EP × asymmetry	.43 (6)	.03
EP × sample size	2.27 (12)**	.03

Note: EP = estimation procedure; F (*df*) = Fischer–Snedecor F and degrees of freedom; η_p^2 = partial eta squared.
[a] Error degrees of freedom = 808.
*p < .05. **p < .01.

In contrast, ML_{PE} tends to underestimate λ 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% without solving the problem.

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 show that an estimation procedure yields very different parameter estimates when facing equivalent data and its estimations are not precise; in contrast, those demonstrating a small SD will be more precise when 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 when estimating the parameter, and descriptive analysis showed that all procedures presented results within the acceptable range.

Variables having at least a moderate effect on instability of parameter estimates are the asymmetry of items, the magnitude of λ parameters, and sample sizes. However, differences with regard to item asymmetry are negligible (e.g., for highly asymmetric items *SD* = 0.09, while for symmetric items *SD* = 0.07). Regarding the magnitude of λ 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 higher quality items (λ = .6)

Table 4. Analysis of Variance (ANOVA) of Bias of Factor Correlation Estimation.

Variable	*F* (*df* [a])	η_p^2
EP	27.04 (3)**	.11
Size of λ	4.24 (1)*	.01
Asymmetry	6.89 (2)**	.02
Sample size	2.96 (4)*	.02
EP × λ	8.42 (3)**	.04
EP × asymmetry	1.47 (6)	.01
EP × sample size	5.75 (12)**	.09

Note: EP = estimation procedure; *F* (*df*) = Fischer–Snedecor *F* and degrees of freedom; η_p^2 = partial eta squared.
[a] Error degrees of freedom = 808.
$*p < .05$. $**p < .01$.

parameter estimates are stable (*SD* = 0.06). Finally, for samples equal to or lower than 100 subjects, a large instability of estimates is observed (*SD* = 0.15), and it tends 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 be measured. Hence, estimation procedures should be examined on this matter.

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 was within the range of −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, but the single relevant effect was a two-way interaction among the estimation procedures and sample size. Figure 5 illustrates that this effect was basically a slight bias for small sample sizes that decreases as sample size increases, where ML_{PE} tends to underestimate the correlation while WLS_{PE} tends to overestimate it and $DWLS_{PO}$ and ULS_{PO} are robust to small sample sizes.

SD of Correlations

Based on Table 5, we can determine that no statistically significant or meaningful difference was found between estimation procedures when treated as

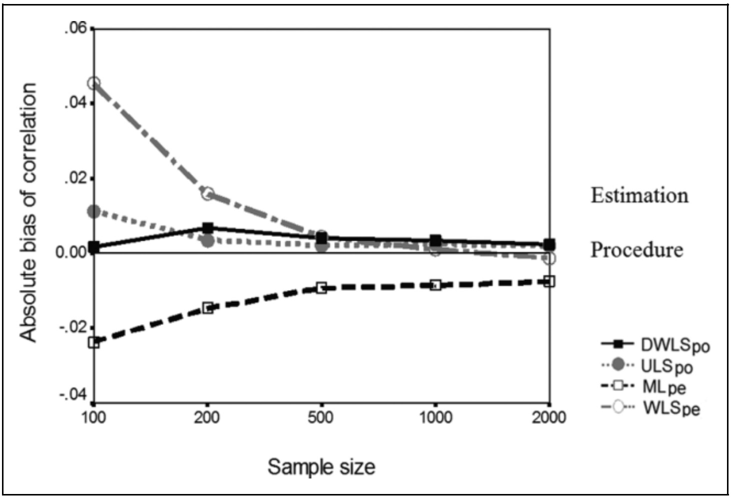

Figure 5. Absolute bias of correlation (ABC) estimate by sample size by estimation procedure.

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., $SD > 0.1$).

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

Only two statistically significant and relevant effects were found for SDC: the magnitude of λ parameters and the sample size. As shown in previous analyses, best results were found for items of good quality and poorer results for those of lower quality (e.g., when $\lambda = .3$ SDC $= 0.18$ and for $\lambda = .6$ SDC $= 0.08$), while heterogeneity of estimations was larger for smaller samples (e.g., when $n = 100$ SDC $= 0.23$ and for $n = 2{,}000$ SDC $= 0.06$).

Overall, results show that, to reach an acceptable level of heterogeneity (SDC < 0.1), samples of 2,000 subjects are required when the quality of the items is low ($\lambda = .3$), while a sample of 500 subjects could be enough if the quality of the items is medium ($\lambda = .6$).

Conclusions

This study aimed to determine the best procedure for analysing factorial models of four-point Likert type items on uni- and multidimensional scenarios.

Table 5. Analysis of Variance (ANOVA) of Standard Deviation (SD) of Factor Correlation Estimation.

Variable	*F* (*df* [a])	η_p^2
EP	.38 (3)	.00
Size of λ	1669.83 (1)**	.71
Asymmetry	30.46 (2)**	.08
Sample size	614.02 (4)**	.78
EP × λ	1.19 (3)	.01
EP × asymmetry	.18 (6)	.00
EP × sample size	.58 (12)	.01

Note: EP = estimation procedure; *F* (*df*) = Fischer–Snedecor *F* and degrees of freedom; η_p^2 = partial eta squared.
[a]Error degrees of freedom = 808.
$*p < .05$. $**p < .01$.

We compared IFA procedures with classical FA procedures and, overall, we found that IFA procedures outperformed the classical perspective.

According to our findings, although all procedures showed similar capacity for producing valid solutions and stable λ and correlation parameter estimates, ULS_{PO} and DWLS_{PO} yielded remarkably lower bias in both parameter estimates and were robust to the toughest scenarios: asymmetric item distributions, low item quality ($\lambda = .3$), and small sample sizes.

It has clearly been confirmed that employing classical estimation procedures in ordinal data with four response alternatives is inappropriate and counterproductive. This is consistent with previous research that reveals underestimation of key parameters in the model when classical FA procedures are employed (Beauducel and Herzberg 2006; DiStefano 2002; Dolan 1994; Holgado–Tello et al. 2010; Rhemtulla et al. 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 the results presented here using Pearson correlation matrices and considering its poor results on tetra and polychoric correlation matrices reported in previous research (Flora and Curran 2004) and (ii) second, that the poor performance of ML_{PE} could be due to the employment of product-moment Pearson correlations rather than to the ML estimation procedure itself, because several studies have shown that using ML estimation on tetra or polychoric correlation matrices yields fairly similar results to DWLS_{PO} and ULS_{PO}, especially in large samples (Dolan 1994; Rigdon and Ferguson 1991; Yang-Wallentin et al. 2010).

According to our findings, IFA should be considered the standard procedure for analyzing 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 ($\mathrm{DWLS}_{\mathrm{PO}}$ and $\mathrm{ULS}_{\mathrm{PO}}$), there are hardly any differences. In fact, although $\mathrm{ULS}_{\mathrm{PO}}$ seems better than $\mathrm{DWLS}_{\mathrm{PO}}$, 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 $\mathrm{ULS}_{\mathrm{PO}}$ 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 $\mathrm{ULS}_{\mathrm{PO}}$ or $\mathrm{DWLS}_{\mathrm{PO}}$ 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 programs such as Factor (Lorenzo-Seva and Ferrando 2006) that is used for exploratory IFA; LISREL (Jöreskog and Sörbom 2006), which is used for confirmatory IFA; and M-Plus (Muthén and Muthén 2011), which is used for both exploratory and confirmatory IFA.

In addition to our main research questions, our inquiry was concerned with the minimum requirements for employing IFA procedures on four-point Likert-type items. In this respect, our research allows us to maintain that if a researcher expects the quality of the items in the scale to be low ($\lambda = .3$), a sample of 500 subjects might be selected in order to ensure a large probability of achieving 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 ($\lambda = .6$), accurate estimations can be reached for small samples (200 or even 100 subjects) if item distributions are symmetric or mildly asymmetric.

To sum up, these research findings reveals 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 and Herzberg 2006; Dolan 1994; Rhemtulla et al. 2012).

Although these findings and guidelines are very interesting and promising for applied research, at least three important limitations to 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 whether these findings could be extended to exploratory models.

Second, we only considered four-point Likert-type items which, to some extent, cannot be completely extrapolated to higher or lower numbers of response categories. Given that, as the number of response categories increases, different procedures tend to yield better results and evidence similar performances (Beauducel and Herzberg 2006; Dolan 1994; Savalei and Rhemtulla 2013), careful research and analysis of three-point Likert scales scenario are still needed and could be well worthwhile considering that dichotomous cases have been widely investigated.

Finally, this research only considered highly "ideal" situations (e.g., homogeneous quality of the items, no cross-loadings, and 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.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Rodrigo Asún received financial support from the Chilean National Commission for Scientific and Technological Research (CONICYT) "Becas Chile" Doctoral Fellowship program (Grant no: 26081114FIC) Karina Rdz-Navarro received financial support from the Chilean National Commission for Scientific and Technological Research (CONICYT) "Becas Chile" Doctoral Fellowship program (Grant no: 72120061).

References

Andrich, David. 1978. "A Rating Formulation for Ordered Response Categories." *Psychometrika* 43:561-73.

Arbuckle, James L. 2010. *Amos (Version 19.0)* [Computer Program]. Chicago, IL: SPSS, an IBM Company.

Beauducel, André and Philipp Y. Herzberg. 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: 186-203.

Bernstein, Ira H. and Gary Teng. 1989. “Factoring Items and Factoring Scales are Different: Spurious Evidence for Multidimensionality Due to Item Categorization.” *Psychological Bulletin* 105:467-77.

Boote, Alfred S. 1981. “Reliability Testing of Psychographic Scales: Five-point or Seven-point? Anchored or Labeled?” *Journal of Advertising Research* 21:53-60.

Brown, Gene, Robert E. Widing, and Ronald L. Coulter. 1991. “Customer Evaluation of Retail Salespeople Using the SOCO Scale: A Replication, Extension, and Application.” *Journal of the Academy of Marketing Science* 9:347-51.

Browne, Michael W. 1984. “Asymptotic Distribution Free Methods in the Analysis of Covariance Structures.” *British Journal of Mathematical and Statistical Psychology* 37:127-41.

Carifio, James and Rocco J. Perla. 2007. “Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes.” *Journal of Social Sciences* 3:106-16.

Chang, Lei. 1994. “A Psychometric Evaluation of 4-point and 6-point Likert-type Scales in Relation to Reliability and Validity.” *Applied Psychological Measurement* 18:205-15.

Christoffersson, Anders. 1975. “Factor Analysis of Dichotomized Variables.” *Psychometrika* 40:5-32.

Christoffersson, Anders. 1977. “Two-step Weighted Least Squares Factor Analysis of Dichotomized Variables.” *Psychometrika* 42:433-38.

Cox III, Eli P. 1980. “The Optimal Number of Response Alternatives for a Scale: A Review.” *Journal of Marketing Research* 17:407-22.

DeVellis, Robert F. 1991. *Scale Development, Theory and Applications*. Vol. 26. Newbury Park, CA: Sage.

DiStefano, Christine. 2002. “The Impact of Categorization with Confirmatory Factor Analysis.” *Structural Equation Modeling: A Multidisciplinary Journal* 9:327-46.

Dolan, Conor 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* 47:309-26.

Fabrigar, Leandre R., Duane T. Wegener, Robert C. MacCallum, and Erin J. Strahan. 1999. “Evaluating the Use of Exploratory Factor Analysis in Psychological Research.” *Psychological Methods* 4:272-99.

Ferguson, Christopher J. 2009. “An Effect Size Primer: A Guide for Clinicians and Researchers.” *Professional Psychology: Research and Practice* 40:532-38.

Flora, David B. and Patrick J. Curran. 2004. "An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis with Ordinal Data." *Psychological Methods* 9:466-91.

Forero, Carlos G. and Alberto Maydeu-Olivares. 2009. "Estimation of IRT Graded Response Models: Limited versus full information methods." *Psychological Methods* 14:275-99.

Forero, Carlos G., Alberto Maydeu-Olivares, and David Gallardo-Pujol. 2009. "Factor Analysis with Ordinal Indicators: A Monte Carlo Study Comparing DWLS and ULS Estimation." *Structural Equation Modeling: A Multidisciplinary Journal* 16:625-41.

Gaito, John. 1980. "Measurement Scales and Statistics: Resurgence of an Old Misconception." *Psychological Bulletin* 87:564-67.

Garland, Ron. 1991. "The Mid-point on a Rating Scale: Is it Desirable?" *Marketing Bulletin* 2:66-70.

Garner, Wendell R. 1960. "Rating Scales, Discriminability and Information Transmission." *Psychological Review* 67:343-52.

González-Romá, Vicente and Begoña Espejo. 2003. "Testing the Middle Response Categories 'Not sure', 'In between' and '?' in Polytomous Items." *Psicothema* 15:278-84.

Hancock, Gregory R. and Alan J. Klockars. 1991. "The Effect of Scale Manipulations on Validity: Targeting Frequency Rating Scales for Anticipated Performance Levels." *Applied Ergonomics* 22:147-54.

Harwell, Michael, Clement A. Stone, Tse-Chi Hsu, and Levent Kirisci. 1996. "Montecarlo Studies in Item Response Theory." *Applied Psychological Measurement* 20:101-25.

Hau, Kit Tai and Herbert W. Marsh. 2004. "The Use of Items Parcels in Structural Equation Modelling: Non-normal Data and Small Sample Sizes." *British Journal of Mathematical Statistical Psychology* 57:327-51.

Holgado–Tello, Francisco Pablo, Salvador Chacón–Moscoso, Isabel Barbero–García, and Enrique Vila–Abad. 2010. "Polychoric Versus Pearson Correlations in Exploratory and Confirmatory Factor Analysis of Ordinal Variables." *Quality & Quantity* 44:153-66.

Hoogland, Jeffrey J. and Anne Boomsma. 1998. "Robustness Studies in Covariance Structural Modeling: An Overview and a Meta-analysis." *Sociological Methods & Research* 26:329-67.

Jamieson, Susan. 2004. "Likert Scales: How to (ab)Use Them." *Medical Education* 38:1212-18.

Jöreskog, Karl G. and Dag Sörbom. 2002. *PRELIS 2: User's Reference Guide*. Lincolnwood, IL: Scientific Software International, Inc.

Jöreskog, Karl G. and Dag Sörbom. 2006. *LISREL 8.8: User's Reference Guide*. Lincolnwood, IL: Scientific Software International, Inc.

Kulas, John T., Alicia A. Stachowski, and Brad A. Haynes. 2008. "Middle Response Functioning in Likert-responses to Personality Items." *Journal of Business and Psychology* 22:251-59.

Likert, Rensis. 1932. "A Technique for the Measurement of Attitudes." *Archives of Psychology* 22:44-55.

Likert, Rensis, Sydney Roslow, and Gardner Murphy. 1934. "A Simple and Reliable Method of Scoring Thurstone Attitudes Scales." *The Journal of Social Psychology* 5:228-38.

Loken, Barbara, Phyllis Pirie, K. A. Virnig, Ronald L. Hinkle, and Charles T. Salmon. 1987. "The Use of 0-10 Scales in Telephone Surveys." *Journal of the Market Research Society* 29:353-62.

Lord, Frederic M. 1953. "On the Statistical Treatment of Football Numbers." *American Psychologist* 8:750-51.

Lorenzo-Seva, Urbano and Pere J. Ferrando. 2006. "FACTOR: A Computer Program to Fit the Exploratory Factor Analysis Model." *Behavioral Research Methods, Instruments and Computers* 38:88-91.

Matell, Michael S. and Jacob Jacoby. 1971. "Is There an Optimal Number of Alternatives for Likert Scale Items? Study 1: Reliability and Validity." *Educational and Psychological Measurement* 31:657-74.

McDonald, Roderick P. 1982. "Linear Versus Nonlinear Models in Item Response Theory." *Applied Psychological Measurement* 6:379-96.

McDonald, Roderick P. 1999. *Test Theory: A Unified Approach*. Mahwah, NJ: Lawrence Erlbaum.

Michell, Joel. 2009. "The Psychometricians' Fallacy: Too Clever by Half?" *British Journal of Mathematical Statistical Psychology* 62:41-55.

Miller, George. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review* 63:81-97.

Muthén, Bengt. 1978. "Contributions to Factor Analysis of Dichotomous Variables." *Psychometrika* 43:551-60.

Muthén, Bengt. 1984. "A General Structural Equation Model with Dichotomous, Ordered Categorical, and Continuous Latent Variables Indicators." *Psychometrika* 49:115-32.

Muthén, Bengt. 1989. "Dichotomous Factor Analysis of Symptom Data." *Sociological Methods & Research* 18:19-65.

Muthén, Bengt. 1993. "Goodness of Fit with Categorical and Other Nonnormal Variables." Pp. 205-34 in *Testing Structural Equation Models*, edited by K. A. Bollen and J. S. Long. Newbury Park, CA: Sage.

Muthén, Bengt, Stephen H. C. du Toit, and Damir Spisic. 1997. "Robust Inference Using Weighted Least Squares and Quadratic Estimating Equations in Latent

Variable Modeling With Categorical and Continuous Outcomes.” Retrieved June 11, 2013 (http://pages.gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf).

Muthén, Linda K. and Bengt Muthén. 2011. *Mplus Version 6.11*. Los Angeles, CA: Muthen & Muthen.

Norman, Geoff. 2010. “Likert Scales, Levels of Measurement and the “Laws” of Statistics.” *Advances in Health Sciences Education* 15:625-32.

Nunnally, Jum C. 1978. *Psychometric Theory*. New York: McGraw-Hill.

Preston, Carolyn C. and Andrew M. Colman. 2000. “Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences.” *Acta Psychologica* 104:1-15.

Raaijmakers, Quinten A. W., J. T. C. van Hoof, T. F. M. A. Verbogt, and W. A. M. Vollebergh. 2000. “Adolescents’ Midpoint Response on Likert-type Scale Items: Neutral or Missing Values?” *International Journal of Public Opinion Research* 12:208-16.

Rhemtulla, Mijke, Patricia É. Brosseau-Liard, and Victoria Savalei. 2012. “When Can Categorical Variables Be Treated as Continuous? A Comparison of Robust Continuous and Categorical SEM Estimation Methods Under Suboptimal Conditions.” *Psychological Methods* 17:354-73.

Rigdon, Edward E. and Carl E. Ferguson, Jr. 1991. “The Performance of the Polychoric Correlation Coefficient and Selected Fitting Functions in Confirmatory Factor Analysis with Ordinal Data.” *Journal of Marketing Research* 28:491-97.

Savalei, Victoria and Mijke Rhemtulla. 2013. “The Performance of Robust Test Statistics with Categorical Data.” *British Journal of Mathematical and Statistical Psychology* 66:201-23.

Sijtsma, Klaas. 2009. “On the Use, the Misuse, and the Very Limited Usefulness of Cronbach’s Alpha.” *Psychometrika* 74:107-20.

Spector, Paul E. 1992. *Summating Rating Scale Construction: An Introduction*. Vol. 82. Newbury Park, CA: Sage.

Velleman, Paul F. and Leland Wilkinson. 1993. “Nominal, Ordinal, Interval, and Ratio Typologies are Misleading.” *American Statistician* 47:65-72.

Wirth, R. J. and Michael C. Edwards. 2007. “Item Factor Analysis: Current Approaches and Future Directions.” *Psychological Methods* 12:58-79.

Yang-Wallentin, Fan, Karl G. Jöreskog, and Hao Luo. 2010. “Confirmatory Factor Analysis of Ordinal Variables with Misspecified Models.” *Structural Equation Modeling: A Multidisciplinary Journal* 17:392-423.

Author Biographies

Rodrigo A. Asún is a tenured assistant professor of Quantitative Research and Statistics at the Department of Sociology, Faculty of Social Sciences, University of

Chile. His research is centered on social movements and latent variable modeling for categorical data.

Karina Rdz-Navarro is a lecturer of Statistics and Quantitative Methodology at the Faculty of Social Sciences, University of Chile. Her research is focused on nonlinear structural equation modeling and latent variable modeling for categorical and continuous data.

Jesús M. Alvarado is a tenured professor of Research Methodology and Quantitative Methods at Faculty of Psychology, Complutense University of Madrid. His research interests are the development and validation of psychometric tests and scales, factor analysis and structural equation modeling.