Qual Quant
DOI 10.1007/s11135-015-0282-1

Robust evaluation of fit indices to fake-good perturbation of ordinal data

Luigi Lombardi[1] · **Massimiliano Pastore**[2]

Abstract This study extended the findings of a former simulation study (Multivar Behav Res 47:519–546, 2012) to evaluate the sensitivity of a large set of SEM-based fit indices to fake-good ordinal data. In the new simulation study we manipulated a comprehensive set of factors (including 3 robust estimation procedures and 3 different faking good models) that could influence the performance of 8 widely used fit indices. The simulation study conditions were chosen to highlight the differences among the fit indices, as well as to cover a wide variety of conditions. Our results demonstrated empirically that the normed fit index (NFI) turned out to be the most reliable fit index with a high sensitivity to fake perturbations. This result was evident in all the simulation design conditions except for those characterized by slight faking levels of perturbations. Interestingly, unlike NFI, the comparative fit index seemed to be highly insensitive to fake data when robust estimation conditions were considered. On the basis of the results of the simulation study we proposed a simple qualitative criterion to evaluate the impact of faking on statistical results.

Keywords Goodness-of-fit indices · Sample generation by replacement (SGR) · Robust estimation · Ordinal data · Faking good

✉ Massimiliano Pastore
massimiliano.pastore@unipd.it

Luigi Lombardi
luigi.lombardi@unitn.it

[1] Department of Psychology and Cognitive Science, University of Trento, corso Bettini, 31, 38068 Rovereto, TN, Italy

[2] Department of Developmental and Social Psychology, University of Padova, via Venezia, 8, 35131 Padova, Italy

Published online: 20 November 2015

1 Introduction

In situations where a model is fitted on empirical data containing possible fake measurements, a SEM-based fit index that evaluates that model may not be very helpful in deciding whether or not it can be appropriate in representing the true relationships under study. Ideally, we would expect that a good fit index should approach its maximum under correct model specification and uncorrupted data but also degrade substantially under massive fake data. A variety of fit indices can be used to evaluate the overall fit of a structural equation model (e.g., Browne and Cudeck 1993; Hu and Bentler 1998; Jöreskog and Sörbom 1996a). However, because standard fit indices are designed to detect model misspecification, but they are not designed to detect the eventual presence of fake observations in the data, it is important to evaluate their behavior in faking scenarios. An open question is whether the results of a standard goodness-of-fit analysis can be integrated in order to provide useful information about the presence of fake perturbations in the data. In this paper, we will show that when a factorial model is fitted to Likert-type ordinal data using robust estimation procedures, the adoption of a simple qualitative criterion based on the performances of two well-known fit indices (NFI and CFI) can serve to yield predictions about statistical results corrupted by fake data.

Many self-report measures of attitudes, beliefs, personality, and pathology are constructed using items that may be easily manipulated by respondents. Several examples of data manipulation or data distortion can be found in areas like psychology (Hopwood et al. 2008), organizational and social science (Van der Geest and Sarkodie 1998), forensic medicine (Gray et al. 2003), and scientific frauds (Marshall 2000). In general, possible fake data confront the researcher with a crucial question: If data included fake data points, what would the chance be that the model is still a good one? Clearly, voluntarily perturbation of data constitutes biased information which certainly weakens the accuracy of statistical inferences.

There is now a vast psychometric literature in item response theory (IRT) and item factor-analytic (FA) modeling about the conceptualization of faking, the modeling of its components, and its interrelationships with individual differences. In particular, psychometric methods have been developed to identify and evaluate subjects responses for feigning (fake-bad, malingering) or defensiveness (fake-good, self-deception, social desirability) using factor analytic approaches (e.g., Ferrando 2005; Ferrando and Anguiano-Carrasco 2009, 2013; Fox and Meijer 2008; Holden and Book 2009; Leite and Cooper 2010; McFarland and Ryan 2000; Paulhus 1991; Ziegler and Buehner 2009), factor mixture models (e.g., Leite and Cooper 2010), IRT models (e.g., Ferrando and Anguiano-Carrasco 2009; Zickar and Drasgow 1996; Zickar and Robie 1999), mixed-models IRT (e.g., Zickar et al. 2004), case-diagnostic procedures (e.g., Pek and MacCallum 2011), and person-fit statistics (e.g., Zickar and Drasgow 1996; Zickar and Robie 1999). Notably, the majority of these methods are based on ad hoc empirical paradigms such as, for example, *coached faking* or *ad-lib faking* that require the administration of self-report questionnaires in a laboratory-type setting (e.g., honest motivating condition vs. faking motivating condition). However, one of the main problems with laboratory studies comparing situations with different types of instructions for self-representation is that they may suffer from the lack of ecological validity and may not provide sufficient support for their use in a large number of applied settings. In the real practice, self-reported responses are usually collected using a single administration and generally the statistical models are simply evaluated by using some type of goodness-of-fit statistic. Therefore, a systematic evaluation of

the pros and cons of standard goodness-of-fit indices on data collected in sensitive contexts can be of relevant interest for researchers working in psychology and social science fields.

Unfortunately, there is little knowledge about how the performance of a factorial model or a structural equation model will be in general affected from fake data perturbation. We are only aware of one study (Lombardi and Pastore 2012) which explored this issue in more depth using a systematic Monte Carlo simulation design. In particular, in this study the authors used a novel probabilistic procedure, called sample generation by replacement (SGR), to simulate artificial fake data and evaluate their impact of goodness-of-fit performances. The results showed that none of the fit indices considered in their simulation study really stood out as having ideal behavioral patterns: sensitive to fake perturbations but insensitive to other irrelevant factors (e.g., model types and sample size). However, important local differences were observed between the indices. In particular, some incremental fit indices (CFI, NNFI and NFI) were clearly more sensitive to fake perturbation than other absolute fit indices (GFI, AGFI, and ECVI), at least when the maximum likelihood (ML) estimation procedure was considered. Very surprisingly, only NFI turned out to be sensitive to fake data also under a weighted least square (WLS) estimation condition. Therefore, the authors concluded by recommending to include NFI in an ideal battery of model fit indices to evaluate the effect of potential fake observations in the data.

However, the SGR simulation study by Lombardi and Pastore (2012) was also characterized by some important limitations. First, because the focus of their study was on the impact of fake data under empirical investigations that are commonly encountered in applied research, they preferred to limit the SGR analysis to small sample sizes (100 and 200) only. Second, in the SGR simulation the fake perturbations were restricted to a simple uniform support fake-good distribution representing a purely random but polarized faking process. However, it is known that some empirical contexts may require different model assumptions about the faking process that cannot be captured by this simple uniform faking model. For example, different modulations of graded faking such as slight faking and extreme faking (e.g., Zickar et al. 2004; Zickar and Robie 1999) are clearly not consistent with this hypothesis. Finally, third, the estimation procedures were limited only to ML and WLS. However, it is known that the very popular ML may not have theoretical justification for use with ordinal variables and full WLS usually requires much larger sample sizes (maybe in the thousands) to fully avoid estimation biases (Ding et al. 1995; Flora and Curran 2004). A superior approach with ordinal data would have instead been to use polychoric correlations together with more robust estimation procedures (Beauducel and Herzberg 2006; Flora and Curran 2004; Ridgon and Ferguson 1991) such as, for example, unweighted least squares (ULS) or diagonally weighted least squares (DWLS). When used routinely with the polychoric correlations, these methods are known to provide consistent parameter estimates as well as correct standard errors (Forero et al. 2009; Yang-Wallentin et al. 2010).

To fill these gaps, in the current study we performed a new extensive SGR simulation study to investigate the sensitivity of 8 commonly used SEM-based fit indices (goodness of fit index, GFI; adjusted goodness of fit index, AGFI; expected cross validation index, ECVI; standardized root-mean-square residual index, SRMR; root-mean-square error of approximation, RMSEA; comparative fit index, CFI; nonnormed fit index, NNFI; and normed fit index, NFI) to fake perturbations of ordinal data in three different SEM models under a new set of simulation conditions. In particular, the new SGR simulation study was based on the following three important features:

1. it involved larger sample sizes to better evaluate the performances of the SEM-based fit indices
2. it included three additional robust estimation procedures, namely robust RULS, robust RDWLS, and robust maximum likelihood (RML) (see Yang-Wallentin et al. 2010)
3. it extended the data replacement procedure to mimicking also slight faking and extreme faking in the data perturbation process.

Finally, on the basis of the results of the new SGR simulation study, we proposed also a simple qualitative criterion to evaluate the results of a SEM-based analysis under faking corrupted data.

To provide a self-contained exposition, the first part of the paper briefly recapitulates the main aspects of the SGR approach to simulate fake data. Next the models of faking and the target SEM models used in this study are introduced. The second part describes the SGR simulation and reports results about the fit indices' performances. The third section illustrates the new qualitative criterion to evaluate the performance of a factorial model under faking scenarios. Finally, the article ends by presenting conclusions and some relevant comments about limitations, potential new applications and extensions of the SGR approach.

2 Sample generation by replacement (SGR)

SGR is a probabilistic resampling procedure that can be used to generate artificial fake discrete or ordinal data with a restricted number of values (Lombardi and Pastore 2012, 2014; Pastore and Lombardi 2014). SGR uses a two-stage sampling procedure based on two distinct generative models: the model defining the process that generates the data prior to any fake perturbation (*data generation process*) and the faking model which is used to perturb the data (*data replacement process*). By repeatedly sampling data from the SGR procedure we can generate the so called fake data sample (FDS) and eventually study the distribution of some relevant statistics computed on this simulated space. In SGR the first process is represented by some standard Monte Carlo procedures for ordinal data whereas the data replacement process is implemented using ad hoc probabilistic faking models. Overall, the entire procedure is split into two conceptually independent and possibly simpler components: data generation + data replacement.

More formally, in the SGR framework the original (fake-uncorrupted) data is represented by an $I \times J$ matrix $\mathbf{D}$, that is to say, I i.i.d. observations (hypothetical participants) each containing J elements (hypothetical participant's responses). We constraint entry d_{ij} of $\mathbf{D}$ ($i = 1, \ldots, I; j = 1, \ldots, J$) to take values on a small ordinal range $\mathcal{V}_Q = \{1, 2, \ldots, Q\}$ (e.g., $Q = 5$ for 5-point Likert items). In particular, let $\mathbf{d}_i$ be the $(1 \times J)$ array of $\mathbf{D}$ denoting the hypothetical pattern of responses of participant i. The array $\mathbf{d}_i$ is a multidimensional ordinal random variable with probability distribution $p(\mathbf{d}_i|\theta)$, where θ indicates the vector of parameters of the probabilistic model for the data generation process. Consequently, the data matrix $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_I]^T$ is drawn from the joint probability distribution

$$p(\mathbf{D}|\theta) = \prod_{i=1}^{I} p(\mathbf{d}_i|\theta). \tag{1}$$

which represents the original data generation process. The main idea underlying our replacement approach is to construct a new $I \times J$ ordinal data matrix $\mathbf{F}$, called the *fake data*

matrix of $\mathbf{D}$, by manipulating each element d_{ij} in $\mathbf{D}$ according to a replacement probability distribution (data replacement process). Let $\mathbf{f}_i$ be the $(1 \times J)$ array of $\mathbf{F}$ denoting the replaced pattern of fake responses of participant i. The fake response pattern $\mathbf{f}_i$ is a multidimensional ordinal random variable with conditional replacement probability distribution

$$p(\mathbf{f}_i|\mathbf{d}_i, \theta_F) = \prod_{j=1}^{J} p(f_{ij}|d_{ij}, \theta_F), \qquad i = 1, \ldots, I \tag{2}$$

where θ_F indicates the vector of parameters of the probabilistic faking model in the data replacement process. The main assumption of the conditional replacement distribution is that each fake response f_{ij} only depends on the corresponding data observation d_{ij} and the model parameter θ_F. Because the patterns of fake responses are also i.i.d. observations, the fake data matrix $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_I]^T$ is drawn from the joint probability distribution

$$p(\mathbf{F}|\mathbf{D}, \theta_F) = \prod_{i=1}^{I} \prod_{j=1}^{J} p(f_{ij}|d_{ij}, \theta_F) \tag{3}$$

Finally, the simulated joint data array $(\mathbf{D}, \mathbf{F})$ is described by the joint probability distribution

$$p(\mathbf{D}, \mathbf{F}|\theta, \theta_F) = \prod_{i=1}^{I} p(\mathbf{d}_i|\theta) p(\mathbf{f}_i|\mathbf{d}_i, \theta_F) \tag{4}$$

$$= \prod_{i=1}^{I} p(\mathbf{d}_i|\theta) \prod_{j=1}^{J} p(f_{ij}|d_{ij}, \theta_F) \tag{5}$$

Because SGR is a data simulation procedure to artificially generate fake data, the parameter array θ_F usually represents hypothetical a priori knowledge about the distribution of faking (e.g., the chance of observing a fake observation in the data) or empirically based knowledge about the process of faking (e.g., the direction of faking-fake-good vs. fake bad-).

Overall, SGR takes an interpretation perspective which incorporates in a global model all the available information (empirical or hypothetical) about the process of faking and the underlying true model representation. In particular, we stress that SGR is not a method for detecting faking at the individual level but a rational approach to evaluate statistical results under potential faking corrupted data. In addition, SGR has a statistical descriptive nature and tries to capture the phenomenological effect of faking according to an informational, data-oriented perspective based on a data replacement (information replacement) scheme. This makes SGR related in spirit to other statistical approaches such as, for example, uncertainty and sensitivity analysis (Helton et al. 2006) and prospective power analysis (Cohen 1988). All these approaches are characterized by an attempt to directly quantify uncertainty of general statistics computed on the data by means of specific hypothesis.

2.1 Representing SEM models with ordinal variables in SGR

Following the UVA framework (Muthén 1984; Lee et al. 1990; Jöreskog 1990) we assume that there exists a continuous data matrix $\mathbf{D}^*$ underlying the original ordinal data matrix $\mathbf{D}$.

More precisely, $\mathbf{D}^*$ is a random sample from the statistical population determined by the true population parameters θ_S of a target SEM model. In particular, θ_S defines the form of the SEM through the specifications of its means and intercepts, variances and covariances, regression parameters, and factor loadings. Let $\mathbf{d}_i^*$ be the $(1 \times J)$ array of $\mathbf{D}^*$ denoting the pattern of underlying continuous responses of participant i. Without loss of generality, it is convenient to let $\mathbf{d}_i^*$ have the multivariate normal distribution with density function $\phi(\mathbf{0}, \mathbf{R})$ where $\mathbf{0}$ and $\mathbf{R}$ denote the $(1 \times J)$ array of zeros representing the location vector of ϕ and the $(J \times J)$ correlation matrix $\mathbf{R} = \mathbf{R}(\theta_S)$ implied by the target SEM model, respectively. The connection between the ordinal variable d_{ij} and the underlying variable d_{ij}^* in $\mathbf{D}^*$ is given by

$$d_{ij} = q \qquad \text{iff} \qquad \alpha_{q-1} < d_{ij}^* < \alpha_q; \quad q = 1, \ldots, Q; i = 1, \ldots, I; j = 1, \ldots, J,$$

where

$$\alpha_0 = -\infty, \alpha_1 < \alpha_2 < \ldots < \alpha_{Q-1}, \alpha_Q = +\infty,$$

are threshold parameters. For each variable d_{ij} with Q categories, there are $Q - 1$ strictly increasing threshold parameters: $\alpha = (\alpha_1, \ldots, \alpha_{Q-1})$. Note that in our simplified context the vector of thresholds α is assumed to be the same for all the J underlying continuous variables. Therefore, the probability distribution for the multidimensional ordinal random variable $\mathbf{d}_i$ is given by

$$p(\mathbf{d}_i|\theta_M) = \int_{\alpha_{(d_{i1})-1}}^{\alpha_{d_{i1}}} \cdots \int_{\alpha_{(d_{iJ})-1}}^{\alpha_{d_{iJ}}} \phi(\mathbf{z}_i|\mathbf{0}, \mathbf{R}) d\mathbf{z}_i \tag{6}$$

with $\theta_M = (\alpha, \mathbf{R})$ and $\mathbf{z}_i = (z_{i1}, \ldots, z_{iJ})$ being the parameter vector of the original data generation model and the values for the continuous variables $\mathbf{d}_i^*$, respectively. Finally, the joint probability distribution is defined as

$$p(\mathbf{D}|\theta_M) = \prod_{i=1}^{I} \int_{\alpha_{(d_{i1})-1}}^{\alpha_{d_{i1}}} \cdots \int_{\alpha_{(d_{iJ})-1}}^{\alpha_{d_{iJ}}} \phi(\mathbf{z}_i|\mathbf{0}, \mathbf{R}) d\mathbf{z}_i \tag{7}$$

In the simulation study based on the SGR procedure we will first generate the continuous data $\mathbf{D}^*$ (according to the target SEM model) and subsequently transform it into its discrete counterpart $\mathbf{D}$ by using appropriate fixed threshold values α. Notice that this latter sampling procedure is equivalent to a direct sampling from the distribution defined in Eq. 7.

2.2 Representing models of faking in SGR

Faking good can be defined as a conscious attempt to present false information to create a favorable impression with the goal of influencing others (e.g., Furnham 1986; McFarland and Ryan 2000; Zickar and Robie 1999). More in general, there is a broad consensus that faking is an intentional response distortion aimed at achieving a personal gain (e.g., MacCann et al. 2011). For example, in personnel selection some job applicants may misrepresent themselves on a personality test hoping to increase the likelihood of being offered a job (e.g., Paulhus 1984; Zickar and Robie 1999; Donovan et al. 2014).

In SGR a fake-good manipulation represents a context in which the responses are exclusively subject to positive feigning:

$$f_{ij} \geq d_{ij} \qquad i = 1, \ldots, I; j = 1, \ldots, J. \tag{8}$$

In particular, the fake-good (as well as the fake-bad) scenario always entails a conditional replacement model in which the conditioning is a function of response polarity. In this study we used a flexible replacement distribution that mimicked the effect of faking good perturbations in ordered variables for different faking modulations (Pastore and Lombardi 2014; Lombardi and Pastore 2014). In particular, the replacement distribution is defined as follows (see also Fig. 1)

$$p(f_{ij} = q' | d_{ij} = q, \theta_F) = \begin{cases} 1, & q = q' = Q \\ \pi DG(q'; q+1, Q, \gamma, \delta), & 1 \leq q < q' \leq Q \\ 1 - \pi, & 1 \leq q' = q < Q \\ 0, & 1 \leq q' < q \leq Q \end{cases} \tag{9}$$

Eq. (9) denotes the conditional probability of replacing an original observed value q in entry (i, j) of $\mathbf{D}$ with the new value q'. In Eq. (7) the terms θ_F and DG are the parameter vector $\theta_F = (\gamma, \delta, \pi)$ of the faking model and the generalized beta distribution for discrete variables with bounds $a = q + 1$ and $b = Q$, respectively (for further details see Pastore and Lombardi 2014). Note that in the parameter vector, γ and δ are strictly positive shape parameters for the replacement distribution. Finally, the parameter π denotes the overall probability of faking good and acts as a weight to rescale DG.

Because of its flexibility, the faking model defined in Eq. (9) can represent both symmetric (Fig. 1, second column) and asymmetric (Fig. 1, first and third columns)

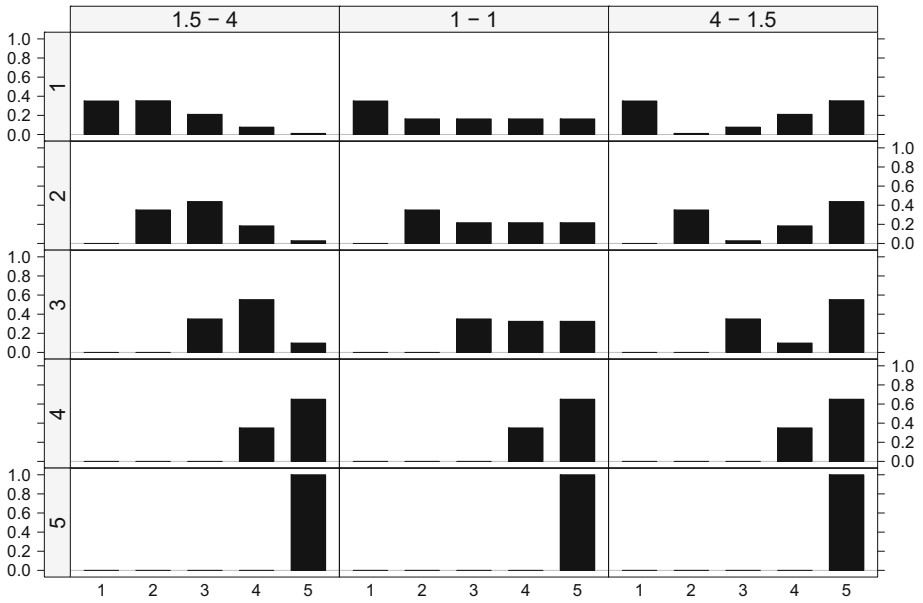

Fig. 1 Three models of conditional replacement distributions for a 5-point discrete r.v. *Each column* in the graphical representation corresponds to a different conditional replacement distribution with overall probability of replacement $\pi = 0.4$ and one of the three different assignments for the shape parameters ($\gamma = 1.5, \delta = 4$; $\gamma = \delta = 1$; and $\gamma = 4, \delta = 1.5$). *Each row* in the graphical representation corresponds to a different original 5-point discrete value h

replacement kernels. In particular, if $\gamma = \delta = 1$ (Fig. 1, second column), the model reduces to the uniform support fake-good distribution originally introduced in Lombardi and Pastore (2012). By contrast, if $1 \leq \gamma < \delta$ (resp. $1 \leq \delta < \gamma$), the model mimics asymmetric faking configurations corresponding to moderate positive shifts (resp. exaggerated positive shifts) in the value of the original response (Fig. 1, first and third columns). A relevant case is when $\pi = 0$. For this special condition the fake data matrix **F** reduces to the original data matrix **D**.

3 Simulation study

In the following two sections we present the specific faking models and the specific target SEM models that we used in the simulation study for evaluating the performances of the goodness-of-fit indices.

3.1 Three relevant models of faking

Several evidences have shown that individuals usually differ in the extent to which they fake (Zickar and Robie 1999; Zickar et al. 2004). For example, depending on the context, some individuals may distort their responses at a level that suggests extreme deception, whereas in other circumstances they can barely exaggerate their personality characteristics (Rosse et al. 1998). In general, the magnitude of faking differs both among individuals and sensitive contexts. For the modeling of the faking process we used three modulations of graded fake-good perturbations: uninformative/neutral faking, slight faking, and extreme faking.

The first representation (*uninformative model*: $\gamma = \delta = 1$) is characterized by the uniform support fake-good distribution (Lombardi and Pastore 2012). The idea is that in the absence of further knowledge all entries in the original data set **D** as well as all candidate replacement values are assumed to be equally likely in the process of replacement. Figure 1 (second column) shows four examples of uniform support fake-good distributions.

The second representation (*slight model*: $\gamma = 1.5; \delta = 4$) mimicked an asymmetric faking good scenario in which the observed self report measure corresponded to a moderate positive shift in the value of the original response. In this model the chance to replace an original value q with another greater value q' decreased as a function of the distance between q' and q which boiled down to a right skewed distribution for the replaced values. Figure 1 (first column) shows four examples of conditional replacement distributions for slight faking.

Finally, the third representation (*extreme model*: $\gamma = 4; \delta = 1.5$) described also an asymmetric faking good scenario in which the observed self report measure corresponded to an exaggerated positive shift in the value of the original response. Unlike the slight model, in the extreme model the chance to replace an original value q with another greater value q' increased as a function of the distance between q' and q. This reduced to a kind of left skewed distribution for the replaced values. Figure 1 (third column) shows four examples of conditional replacement distributions for extreme faking.

3.2 Target SEM models

To provide a comparison with the results of the original SGR simulation study we adopted the same three SEM models evaluated in Lombardi and Pastore (2012). The SEM models

were used for representing the processes that generated the underlying data $\mathbf{D}^*$ prior to any fake perturbation (and data discretization). The three factorial models depicted in (Fig. 2) are commonly encountered in applied research (Paxton et al. 2001; Curran et al. 2002). The first model, Model 1, had nine measured variables ($y_1, \ldots, y_9$) and three latent variables (η_1, η_2 and η_3). Each measured variable loaded on a single latent variable. Moreover, η_2 was regressed on η_1, and η_3 was regressed on η_2. The second model (Model 2) was similar to Model 1 but contained 15 measured variables ($y_1, \ldots, y_{15}$), with five indicators

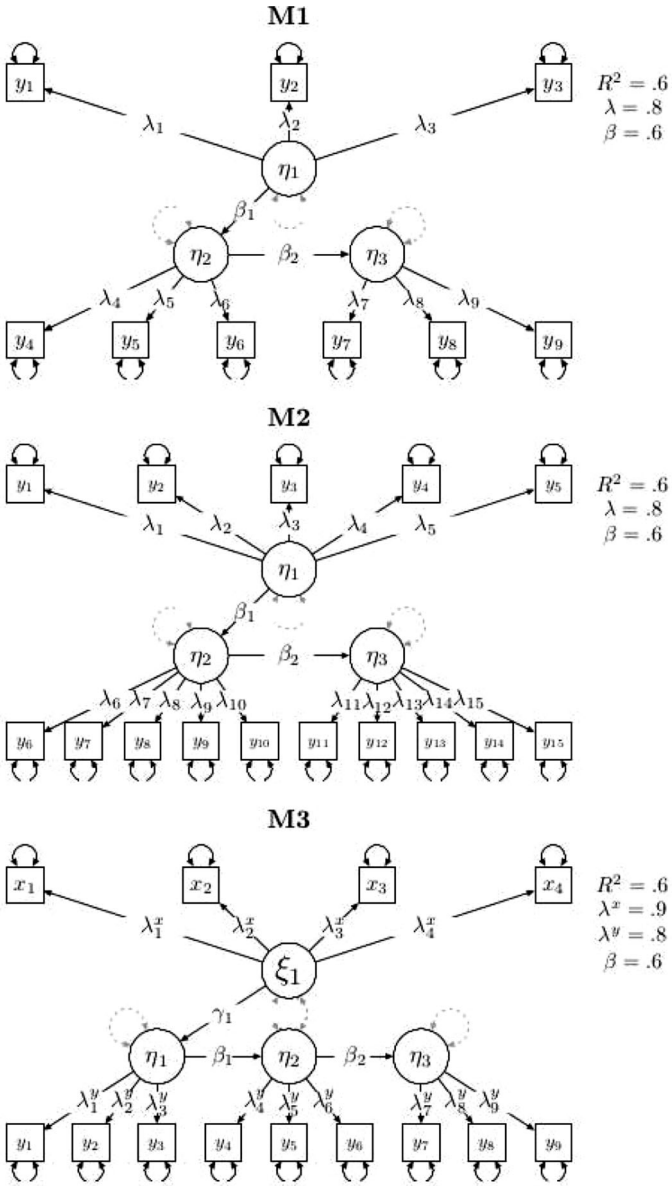

Fig. 2 Target SEM models. In M1 and M2 all error variances were set to .5 whereas in M3 they were all set to .09

per latent variable. Finally, Model 3 had 13 measured variables. The endogenous variables ($y_1, \ldots, y_9$) shared the same measurement structure as Model 1 (three indicators per latent variable), whereas the exogenous variables ($x_1, \ldots, x_4$) loaded on a new latent variable (ξ_1) which, in turn, affected η_1. More details about the structures of the models and the parameters assignments are reported in Fig. 2.

In our simulation study, the models also differed in terms of the way fake perturbations were propagated through the model's structure. In particular, in Model 1 and Model 2 the fake perturbation was propagated through all the observed variables (9 for Model 1 and 15 for Model 2). By contrast, in Model 3 the fake perturbation was propagated through the endogenous variables, $y_1, \ldots, y_9$, only; whereas the exogenous variables, $x_1, \ldots, x_4$, were considered fake independent. We expect that in general a fit index should be less sensitive to Model 3 as compared to Model 1, as the first contained proportionally less fake observations than the latter. By contrast, a good fit index should be equally sensitive to Model 1 and Model 2, as model size (defined as the number of observed variables in the model) should generally not affect the behavior of a fit index (e.g., Fan and Sivo 2007; Kenny and McCoach 2003). These two conditions will be separately evaluated in the simulation study.

3.3 Estimation methods

To estimate the target SEM models four alternative methods were considered in this study, namely robust RULS, robust RDWLS, robust maximum likelihood (RML), and standard maximum likelihood (ML). The three robust estimation procedures, RULS, RDWLS, and RML, are described in details in Yang-Wallentin et al. (2010). Overall, the robust methods are known to provide more accurate and less variable parameter estimates, as well as more precise standard errors and better coverage rates than standard ML and WLS (Forero et al. 2009; Yang-Wallentin et al. 2010). Robust estimation procedures are particularly relevant for our study as faking good perturbations usually entails skewed ordinal data and robust estimation procedures are known to be less sensitive to skewed data as compared to standard maximum likelihood approaches. Finally, WLS was not considered in this study as it is well known that it performs poorly unless sample size is very large and model size is modest (Dolan 1994; Flora and Curran 2004; Muthén and Kaplan 1992).

3.4 Types of fit indices

In this study we evaluated the performances of eight well-known fit indices: goodness of fit index (GFI), adjusted goodness of fit index (AGFI), expected cross validation index (ECVI), standardized root-mean-square residual index (SRMR), root-mean-square error of approximation (RMSEA), comparative fit index (CFI), nonnormed fit index (NNFI or TLI), and normed fit index (NFI). The basic characteristics of each of the eight fit indices are reported in Table 1. Notice that the detailed definitions and reviews of these fit indices are easily available in the SEM literature (e.g., Browne and Cudeck 1993; Hu and Bentler 1998; Fan and Wang 1998; Schermelleh-Engel et al. 2003).

3.5 Simulation design and data conditions

In our simulation study we used a 5-point ordinal scales as they are very common in many empirical investigations within the social and behavioral sciences. Five factors were systematically varied in a complete five-factorial design:

Table 1 Fit indices

Index	Reference	Direction	Range
GFI	Jöreskog and Sörbom (1984)	Large is good	≤ 1
AGFI	Jöreskog and Sörbom (1984)	Large is good	≤ 1
ECVI	Browne and Cudeck (1993)	Small is good	>0
SRMR	Jöreskog and Sörbom (1981) Bentler (1995)	Small is good	≥ 0
RMSEA	Steiger and Lind (1980)	Small is good	≥ 0
CFI	Bentler (1990)	Large is good	[0, 1]
NNFI (or TLI)	Bentler and Bonett (1980) Tucker and Lewis (1973)	Large is good	Can fall outside [0, 1]
NFI	Bentler and Bonett (1980)	Large is good	[0, 1]

GFI goodness of fit index, *AGFI* adjusted goodness of fit index, *ECVI* expected cross-validation index, *SRMR* standardized root-mean-square residual, *RMSEA* root-mean-square error of approximation, *CFI* comparative fit index, *NNFI* nonnormed fit index, *NFI* normed fit index

a. The model type (MT), at three levels: M_1, M_2, and M_3 (see Fig. 2);
b. The sample size (I), at three levels: 300, 600, and 1200;
c. The estimation procedure (E), at four levels: ML, RML, RULS, and RDWLS;
d. The faking good model type (FM), at three levels: uninformative faking, slight faking, and extreme faking;
e. The percentage of replacements (K) for the endogenous variables in the SEM model, at five levels: 0, 25, 50, 75, and 100 %.

Let t, n, e, f, and k be distinct levels of factors MT, I, E, FM and K, respectively. Moreover, let AS (number of Acceptable Solutions) be a counting variable used to control the flow chart of the simulation design. The following procedural steps were repeated for each of the 540 combinations of levels (t, n, e, f, k) of the simulation design:

1. Set $AS = 0$,
2. Generate a raw-data set $\mathbf{D}^*$ with size n according to model t. The data generation was performed using a standard MC procedure based on multivariate normal data (e.g., Fan et al. 2002). In particular, $\mathbf{d}_i^*$ $(i = 1, \ldots, n)$ is drawn from the multivariate normal distribution $\phi(\mathbf{0}, \mathbf{R}_t)$.
3. Discretize $\mathbf{D}^*$ on a 5-point scale using the method described by Jöreskog and Sörbom (1996b). In particular, the normal quantiles, -1.53, -0.49, 0.49, and 1.53, were used as corresponding threshold values, $\alpha_1, \alpha_2, \alpha_3, \alpha_4$. The four quantiles were computed using the inverse of the binomial CDF. Finally, the original continuous variable d_{ij}^* was discretized into a symmetrically distributed ordinal variable d_{ij} on the basis of the four threshold values. Note that the thresholds remained constant across all variables in $\mathbf{D}$.
4. Fit model t using the polychoric correlation matrix of $\mathbf{D}$; if the model yields an acceptable solution (according to the estimation procedure e), then proceed to Step 5; otherwise, go back to Step 2.
5. Sample a fake data matrix $\mathbf{F}$ using the conditional replacement probability distribution with replacement parameter $\theta_F = (\pi = \frac{k}{100}, \gamma_f, \delta_f)$ given $\mathbf{D}$ (see Eq. 9). The

coefficients γ_f and δ_f refer to the shaping parameters of the faking model f, whereas π denotes the overall probability of faking good.

6. Fit model t using the polychoric correlation matrix of the fake data set **F**; if the model yields an acceptable solution (according to the estimation procedure e), then increment AS (+1), save the fake matrix **F** for later analyses, and proceed to Step 7; otherwise, go back to Step 2.
7. Stop if variable AS counts 3000 acceptable solutions; else, go to Step 2.

This algorithm was used to generate 3000 distinct matrices **F** for each combination of levels (t, n, e, f, k) of the SGR simulation design. This number of replications should guarantee reasonable estimation stability in the tail regions of the fit indices. Finally, for each of the 3000 perturbed data matrices **F** the eight fit indices were separately evaluated and the results saved for later analyses. Overall, the whole procedure generated a total of $1{,}620{,}000 = 3000 \times 3 \times 3 \times 4 \times 3$ new fake matrices as well as an identical number of fit indices results.

3.6 Data source and statistical analyses

The simulated data were generated using the sgr package for sample generation by replacement (Lombardi and Pastore 2014) whereas model fitting and estimation were implemented through LISREL package (Jöreskog and Sörbom 1996a). Because the fit indices evaluated in this study were characterized by a nonzero level of skewness and kurtosis (see Table 2 for some descriptive statistics), we decided to model the dependent variables using a Gamma distribution (McCullagh and Nelder 1989; Dobson 2002; Wood 2006). In particular, Generalized linear models (GLM) based on the Gamma family with inverse link function were used as main statistical analysis to evaluate how the fit indices values were systematically influenced by the design factors. However, in order to guarantee a correct application of GLM models, we transformed all the indices that showed negative values or a negative skewness into new variables with correct ranges and skewnesses (see Table 2, last column). All the Gamma regression models included the main factor terms (I, MT, E, and K) and all the interaction terms as independent variables. Finally, for each fit index, we used an effect size measure defined as $\varphi = 100 \times (D_{source}/D_{null})$ with D_{source} and D_{null} being the deviance attributable to a target factor and the null deviance in the GLM model, respectively. The φ statistic can be understood as the

Table 2 Descriptive statistics of the fit indices

Index (y)	Mean	Sd	Min	Max	Skewness	Kurtosis	Recoded variable (y^*)
GFI	0.98	0.02	0.80	1.00	−1.84	4.62	$y^* = 1 - y + .1$
AGFI	0.97	0.03	0.72	1.00	−1.70	3.84	$y^* = 1 - y + .1$
ECVI	0.25	0.17	0.05	2.14	1.38	2.85	
SRMR	0.03	0.01	0.01	0.11	1.13	1.95	
RMSEA	0.02	0.02	0.00	0.15	1.32	2.25	$y^* = y + .1$
CFI	0.99	0.05	0.00	1.00	−7.68	74.09	$y^* = 1 - y + .1$
NNFI	0.99	0.06	−0.21	2.86	−5.80	56.59	$y^* = 1 - y + max(y)$
NFI	0.95	0.07	0.01	1.00	−3.73	19.35	$y^* = 1 - y + .1$

The last column reports the recoding equation for the negative skewed indices

percentage of deviance explained in a dependent variable attributable to a factor in the GLM model.

4 Results

Tables 3 and 4 report the measures of fit for the Gamma GLM models and the portion of deviance φ explained in a fit index attributable to the design factors and their interactions, respectively. We recall that, an ideal fit index should be sensitive to fake perturbations and faking good model type but should not be sensitive to other irrelevant factors, such as model types (in particular model size: M_1 vs. M_2) and sample size conditions. More specifically, we would expect that a large proportion of deviance φ in a fit index would be attributable to the relevant factors K and FM, and to the difference between M_1 versus M_3. By contrast, the proportion of deviance in a model fit index attributable to I and MT (in particular M_1 vs. M_2) should be minimal. Finally, we also expect that the fit indices would be sensitive to the estimation procedure particularly when a fake data set shows strong asymmetry and kurtosis (Flora and Curran 2004; Forero et al. 2009; Yang-Wallentin et al. 2010). In this latter scenario RML, RULS and RDWLS might represent more robust estimation procedures that are less sensitive to the fake perturbed data. Table 4 suggests that, for the conditions in this study, the fit indices exhibited different behaviors. Half of the fit indices (GFI, AGFI, ECVI, and SRMR) showed undesirable high sensitivity to sample size conditions. The other half (RMSEA, CFI, NNFI, and NFI) was clearly less sensitive to sample size with proportion of deviance attributable to factor I being about 12 % or lower. Overall this result indicates that the values of all the fit indices but NNFI are systematically affected (to different degrees) by sample size.

As shown in Table 4, for the model type conditions, three indices (GFI, ECVI, and NFI) showed some sensitivity to different SEM models with their proportion of deviance attributable to MT ranging from 7.83 % (for GFI) to 22.31 % (for ECVI), with NFI being in the middle (15.16 %). The other indices (AGFI, SRMR, RMSEA, CFI, and NNFI) were clearly less sensitive to this factor. To provide a better understanding of the fit indices' behaviors on the effects of different target models, we recalculated the effect size of factor MT on the basis of the two following conditions: (a) MT recoded with two levels: M_1 and M_2 (b) MT recoded with two levels: M_1 and M_3. Table 5 presents the results for the new recoded factors. As discussed in the previous section, an ideal fit index should show more sensitivity to the difference between M_1 and M_3 (fake proportion condition), and less sensitivity to the difference between M_1 and M_2 (model size condition). Table 5 also reports a descriptive statistic, (a – b), denoting the difference between the φ value of the recoded factor M_1 versus M_2 (a) and the φ value of the recoded factor M_1 versus M_3 (b). Note that, a good fit index should have a negative value for this statistic. A quick inspection of Table 5 immediately reveals that the behaviors of RMSEA, CFI, NNFI, and NFI are consistent with this expectation, whereas GFI, AGFI, ECVI, and SRMR are not. Notably, NFI showed the largest negative difference (−17.60), whereas ECVI was the index with the worst performance (9.03).

For the estimation procedure factor, all the indices but ECVI and SRMR were sensitive to different estimation methods (see Table 4) with RMSEA and NFI being among the sensitive indices the ones showing the highest (43.37 %) and the lowest (6.95 %) sensitivity to factor E, respectively. Finally, for the percentage of replacement conditions and the faking good model type, only SRMR, CFI, and NFI showed high sensitivity to increasing amount of faking and different modulations of graded fake-good perturbations.

Table 3 Measures of fit for the Gamma GLM models

Fit	GFI	AGFI	ECVI	SRMR	RMSEA	CFI	NNFI	NFI
D_{null}	3239.93	5618.14	62492.14	18569.46	2595.37	9478.95	102.44	18296.22
D_{res}	311.06	542.09	1048.44	2420.61	1073.41	1282.75	26.29	2820.91
Gen r^2	0.90	0.90	0.98	0.87	0.59	0.86	0.74	0.85
AIC	−1006753.26	−914278.48	−704534.56	−1087715.68	−846587.61	−836110.19	−551214.93	−667576.46

The *null deviance* (D_{null}) is the deviance for a model with just a constant term, while the *residual deviance* (D_{res}) is the deviance of the fitted model. These two statistics can be combined to give the *proportion of deviance explained*, a generalization of r^2, as follows: $(D_{null} - D_{res})/D_{null}$. AIC is the *Akaike Information Criteria* for the model

Table 4 Partitioning the deviance (φ) of goodness-of-fit indices

Source	GFI	AGFI	ECVI	SRMR	RMSEA	CFI	NNFI	NFI
K	2.77	3.02	0.04	14.46	0.12	8.11	2.05	16.24
I	24.89	27.75	66.36	54.38	7.04	3.21	0.63	12.12
MT	7.83	3.36	22.31	2.88	0.75	3.73	0.78	15.16
E	35.45	38.69	2.52	0.33	43.37	27.18	14.75	6.95
FM	2.86	3.07	0.08	9.03	0.20	10.93	2.66	23.45
K by I	0.26	0.18	0.02	1.22	0.02	0.56	0.20	0.07
K by MT	0.53	0.41	0.00	0.24	0.08	2.05	0.75	1.98
I by MT	0.92	0.27	5.20	0.17	0.03	0.23	0.12	0.12
K by E	0.59	0.81	0.15	0.12	1.65	8.67	8.33	0.93
I by E	4.30	3.49	0.83	0.02	1.42	1.78	2.39	0.02
MT by E	4.61	4.28	0.14	0.07	1.04	2.11	2.81	0.02
K by FM	2.20	2.24	0.09	2.75	0.23	4.46	2.36	2.57
I by FM	0.22	0.14	0.03	0.78	0.00	0.34	0.20	0.24
MT by FM	0.46	0.36	0.00	0.14	0.07	2.12	1.14	3.28
E by FM	0.47	0.58	0.14	0.04	1.23	6.46	11.26	0.49
K by I by MT	0.04	0.02	0.00	0.03	0.01	0.02	0.11	0.03
K by I by E	0.13	0.16	0.05	0.01	0.04	0.06	1.11	0.05
K by MT by E	0.22	0.21	0.01	0.00	0.05	0.84	2.55	0.01
I by MT by E	0.58	0.44	0.04	0.00	0.03	0.02	0.32	0.07
K by I by FM	0.14	0.07	0.03	0.24	0.00	0.05	0.15	0.19
K by MT by FM	0.22	0.14	0.00	0.01	0.03	0.39	0.89	0.29
I by MT by FM	0.03	0.02	0.00	0.02	0.02	0.05	0.20	0.05
K by E by FM	0.31	0.30	0.14	0.00	1.05	1.90	8.95	0.10
I by E by FM	0.08	0.09	0.05	0.00	0.03	0.07	1.32	0.03
MT by E by FM	0.17	0.15	0.01	0.00	0.06	0.80	3.56	0.01
K by I by MT by E	0.03	0.02	0.01	0.00	0.00	0.02	0.27	0.02
K by I by MT by FM	0.01	0.00	0.00	0.00	0.01	0.04	0.13	0.06
K by I by E by FM	0.03	0.02	0.04	0.00	0.02	0.05	0.88	0.01
K by MT by E by FM	0.05	0.04	0.01	0.00	0.04	0.17	2.82	0.01
I by MT by E by FM	0.02	0.02	0.00	0.00	0.00	0.01	0.36	0.01
K by I by MT by E by FM	0.00	0.00	0.00	0.00	0.01	0.05	0.25	0.01

Table 5 Deviance (φ) of goodness-of-fit indices for the MT recoded factors

Source MT	GFI	AGFI	ECVI	SRMR	RMSEA	CFI	NNFI	NFI
(a) M1 versus M2	13.19	5.86	28.73	3.13	0.21	0.22	0.12	0.87
(b) M1 versus M3	4.85	1.75	19.70	0.01	0.34	4.71	0.57	18.47
(a – b)	8.34	4.11	9.03	3.12	−0.13	−4.49	−0.45	−17.60

(a – b) indicates the difference between the φ value of the recoded factor M1 versus M2 and the φ value of the recoded factor M1 versus M3. A good fit index must show a negative value for (a – b)

In particular, in SRMR, CFI, and NFI the proportion of deviance attributable to K was about 14.46, 8.11, and 16.24 %, respectively. Similarly for the three indices (considered in same order as for factor K) the proportion of deviance attributable to FM was (considering the same order listed earlier) about 9.03, 10.93, and 23.45 %, respectively. The other five indices, GFI, AGFI, ECVI, RMSEA, and NNFI were less sensitive to data replacements and fake-good modulations. In particular, factor K and factor FM accounted for a very low amount of variation in ECVI (resp. only 0.04 and 0.08 %) and RMSEA (resp. only 0.12 and 0.20 %). CFI and NNFI showed also some interesting interaction effects. In particular, in CFI we observed two interaction effects with a proportion of deviance larger than 5 %, namely K by E (8.67 %) and E by FM (6.46 %). For NNFI there were three interactions above this threshold, namely K by E (8.33 %), E by FM (11.26 %), and the triple interaction K by E by FM (8.95 %).

On the basis of the results presented in Tables 4 and 5, we would tentatively consider CFI and NFI as having the more correct behavior expected from a model fit index. However, to better explore the behaviors of these two best performing fit indices, in what follows we will show their functional patterns and compare this new graphical analysis with the results reported in this section.

4.1 Graphical analysis

Because of space limitations, we preferred to show the results for the second model M2 only (see Fig. 2). However, this is not a serious loss as the information reported in Table 5 about factor MT is sufficient to discriminate between the performances of the two indices. Figure 3 shows the observed means of CFI as a function of factors K, MF, E, and I,

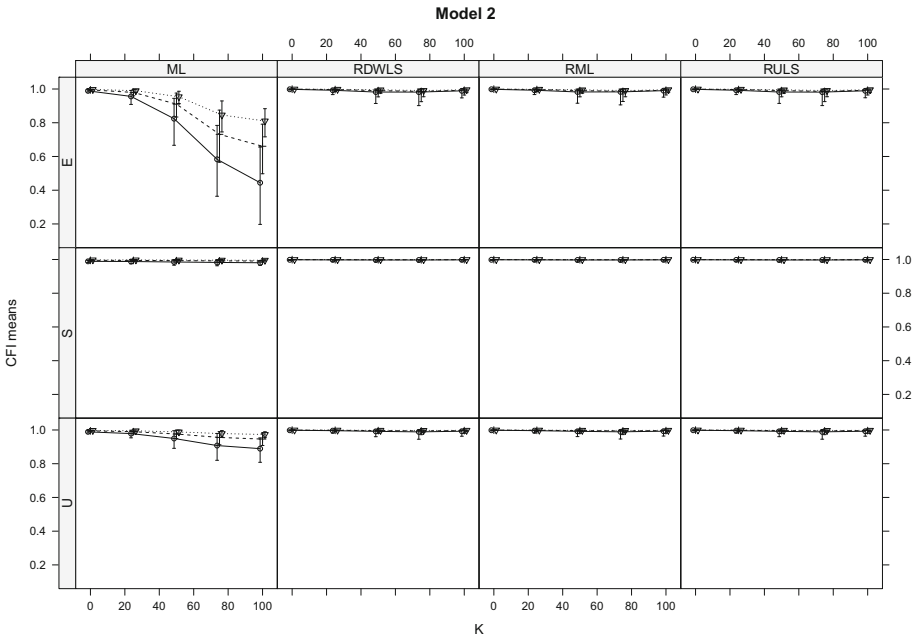

Fig. 3 Means of CFI as a function of percentage of replacements, models of faking, estimation method, and sample size. Segments represent 95 % interquantile intervals

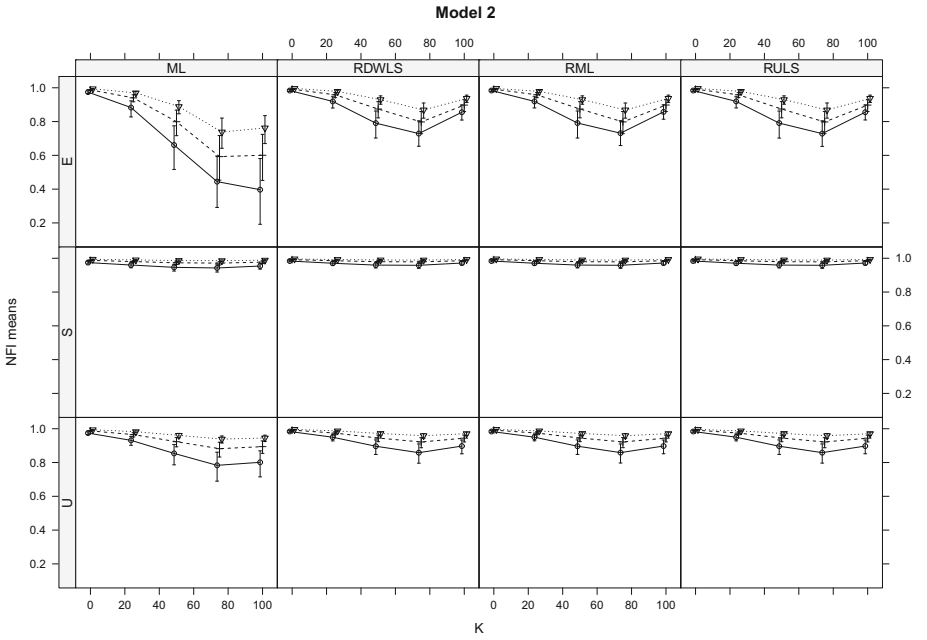

Fig. 4 Means of NFI as a function of percentage of replacements, models of faking, estimation method, and sample size. Segments represent 95 % interquantile intervals

respectively. Segments represent 95 % interquantile intervals. Somehow surprisingly, the CFI index resulted totally unsensitive to the design factors except for a minority of design conditions corresponding to the combinations (ML, extreme faking) and (ML, uniform faking). In particular, for these two conditions, CFI showed a sample size effect with the I = 1200 condition yielding an overall better performance for CFI. Moreover, in the ML condition the CFI mean decreased by increasing levels of replacements, that is to say, it degraded with larger amounts of fake perturbations (K). As expected, the effect was larger in the extreme faking condition, whereas it was practically absent in the slight faking condition. Finally, the uniform faking condition showed a moderate effect for factor K.

Figure 4 presents the patterns of NFI. We recall that this index showed the largest sensitivity to fake perturbation (see Table 4). This result was confirmed also in the graphical analysis. In particular, the NFI mean decreased by increasing levels of replacements (K), that is to say, it degraded with larger amounts of fake perturbations. This result was evident in all the design conditions except for those characterized by the slight faking perturbations. Notably, unlike CFI, the NFI index was sensitive to factors K and MF also in the robust estimation conditions (RML, RULS, and RDWLS). However, the largest impact of faking was observed for the ML estimation procedure which seemed to boost the sensitivity of NFI to fake perturbations. Like for CFI, also the NFI index showed a sample size effect with the I = 1200 reflecting the best performance for NFI.

4.2 Further graphical analysis

Under misspecified conditions in real data a fit index value simply reflects how well a model is able to reconstruct the relationships in the observed data, which are not

necessarily those that are actually represented in the true unknown population. This issue becomes even more relevant when considering perturbed data. The crucial question now becomes: if the data contained fake observations, would a model be able to correctly reconstruct the relations represented in the uncorrupted true population? In an attempt to answer this question, we decided to evaluate to what extend the four estimation procedures were able to reconstruct the true relationships in the population as measured by the population correlation matrix. This further analysis is important, as it will help us in better understanding the performances of the two best fit indices (CFI and NFI).

To reach this objective, we studied the difference between the original correlation matrix $\mathbf{R} = \mathbf{R}(\theta_S)$ implied by model M2 under the parameters assignment in the original generative phase (see the target models section described earlier in this article) and the reconstructed correlation matrix $\hat{\mathbf{R}}$ obtained by fitting the same model to the fake data sets **F**. We examined the average relative bias (ARB) across all study conditions. We used the following estimate of correlation bias:

$$ARB = \frac{200}{T[N(N-1)]} \sum_{t} \sum_{l=2}^{N} \sum_{s=1}^{l-1} \left(\frac{\hat{r}_{ls}^{t} - r_{ls}}{r_{ls}} \right) \tag{10}$$

with $\hat{r}_{ls}^{t}$ and r_{ls} being the (l, s) element of the reconstructed $(N \times N)$ correlation matrix $\hat{\mathbf{R}}^{t}$ in the tth sample replicate $(t = 1, \ldots, T)$ and the (l, s) element of the true $(N \times N)$ correlation matrix **R**, respectively. A large absolute value of ARB indicates a large discrepancy between the population correlation matrix and the reconstructed correlation matrix. Figure 5 shows ARB as a function of factors K, MF, E, and I, respectively. Clearly, ARB

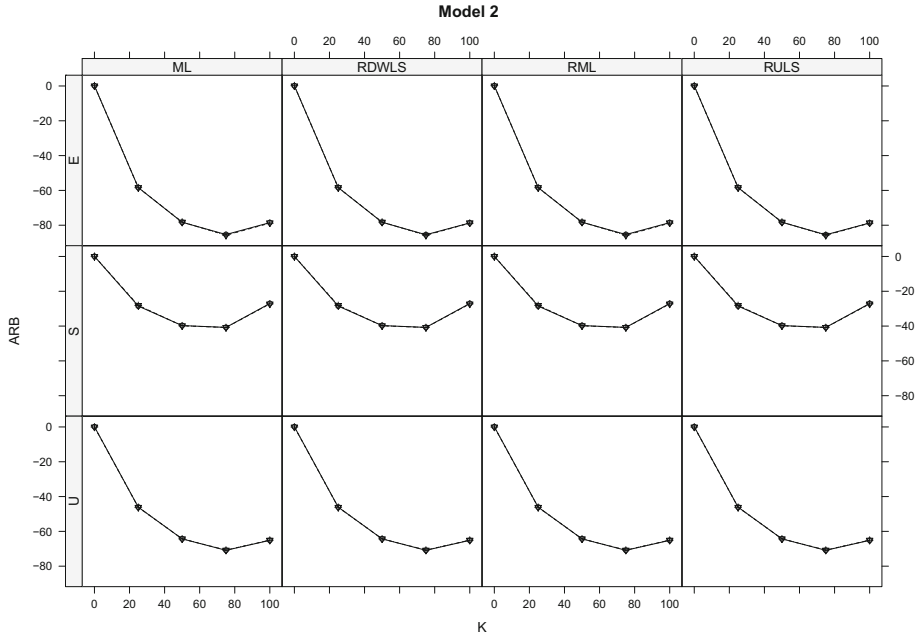

Fig. 5 Average relative bias (ARB) as a function of percentage of replacements, models of faking, estimation method, and sample size

was sensitive to both increasing levels of fake perturbations (K) and fake model type (FM). As expected, the largest discrepancy between the two correlation matrices was observed for the extreme faking condition. Notably, the behavior of ARB was not affected by factor E and factor I. In particular, for factor E all the four estimation procedures were equally affected by fake data in the reconstruction of the correlation matrix. The results were replicated also with the structural models M1 and M3. Therefore, the results seemed very consistent across different sample size and model type conditions.

Taken together the results of the three analyses lead us to conclude that NFI shows the more ideal behavior expected from a model fit index as it actually decreases its performance under fake data conditions irrespective of the estimation procedure adopted. However, as expected the fake effects on NFI were more evident for the standard ML estimation procedure. By contrast, CFI seemed to artificially boost its performance especially in the robust estimation conditions.

4.3 A qualitative criterion for fake data analysis

The main result of the new SGR simulation study can be summarized as follows. Under robust estimation conditions all the studied fit indices but NFI were substantially unsensitive to any level and type of fake perturbation on the original datasets. In particular, only NFI resulted sufficiently sensitive to fake perturbations in the data especially when uninformative and extreme faking conditions were considered. In what follows we describe a simple qualitative criterion, called the *faking criterion* (FC), that can be used to check if the results of a factorial analysis have been potentially corrupted by fake data. We assume that the model has been fitted using a robust estimation procedure and that the set of fit-indices used to evaluate the model includes also the NFI and CFI indices. The criterion acts by classifying the conjunctive combination of results of the two indices into two complementary classes: (a) positive evidence for faking (b) negative evidence for faking. The FC criterion is described in Fig. 6. The graphical representation illustrates the four distinct conditions resulting from the cross-combination of results between the two fit indices. Note that the second combination (first row, second column) indicates some evidence about the hypothesis of fake observations in the data: the most sensitive index (NFI) suggests model rejection whereas the second fit index (CFI), which is unsensitive to fake observations, endorses model acceptance. This is a diverging result which reflects a positive case for faking. By contrast, the remaining tree conditions indicate empirical results that show little evidence for faking. Figure 7 shows the criterion as a function of factors K, MF, E, and I,

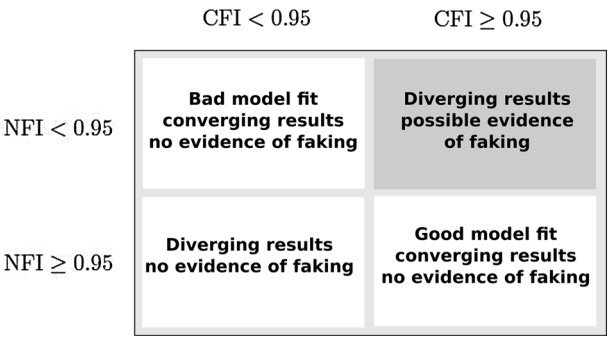

Fig. 6 FC criterion. The four conditions resulting from a cross-combination of NFI and CFI results as a function of the model acceptance threshold 0.95 (e.g., Hu and Bentler 1998)

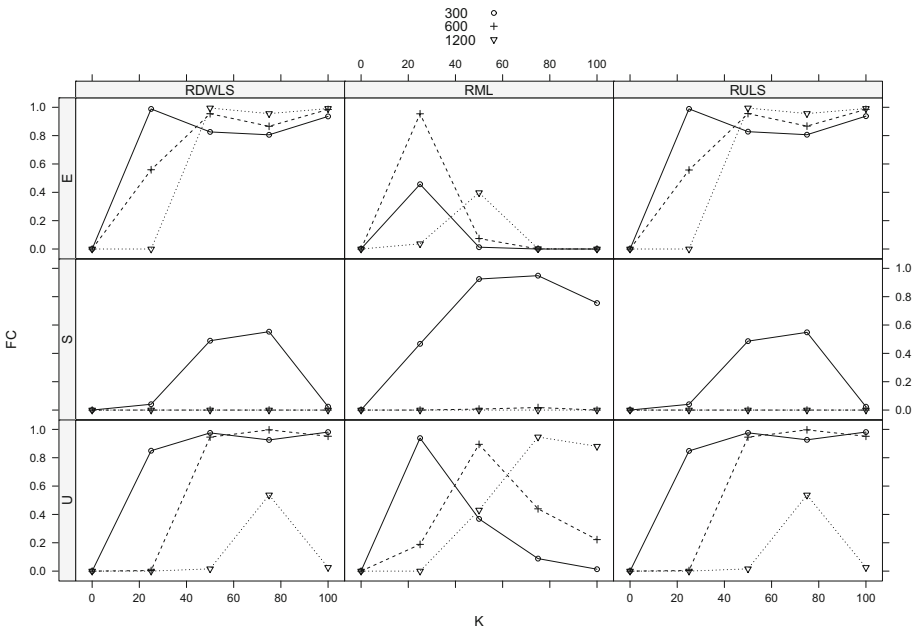

Fig. 7 Proportion of cases positively classified by the faking criterion (FC) as a function of percentage of replacements, models of faking, estimation method, and sample size (model M2 only)

respectively. Clearly, FC was sensitive to both increasing levels of fake perturbations (K) and fake model type (FM). As expected, the largest effect was observed for the extreme faking condition. It is worthwhile to note that when the criterion provides evidence for faking, then the difference between the observed value of CFI and the observed value of NFI can be understood as the strength of this evidence: the larger the difference the stronger is the evidence for faking.

5 Discussion

This study extended the findings of a recent SGR simulation study (Lombardi and Pastore 2012) to evaluate the sensitivity of fit indices to fake perturbed data. The new SGR simulation study sought to investigate the relative performance of 8 commonly used SEM-based fit indices to fake perturbations of ordinal data in three different SEM models using the SGR approach. We manipulated a comprehensive set of factors that could influence their performance, resulting in 540 different conditions. The conditions were chosen to highlight the differences among the fit indices, as well as to cover a wide variety of conditions. Our results demonstrate empirically that the incremental fit index NFI was clearly more sensitive to fake perturbations thus confirming the findings reported in a previous paper by Lombardi and Pastore (2012). This result was evident in all the simulation design conditions except for those characterized by slight faking levels of perturbations. Interestingly, unlike CFI, the NFI index showed a significant sensitivity to factors K and MF not only in the ML condition but also in the other robust estimation conditions (RML, RULS, and RDWLS).

Ideally, we expect that a fit index should be sensitive to the factors associated with fake perturbations. However, the behavior of CFI seemed not consistent with this expectation at least when robust estimation methods were considered in our simulation study. In particular, if a model is subjected to two different levels of fake perturbation in the data, CFI would lead us to the same conclusions about model fit, regardless of which quantity of perturbation is under consideration. Of course, this would still be a desirable property if the robust estimation procedure really reconstructed the original (fake-uncorrupted) true correlations in the population. Under this hypothesis, the lack of sensitivity of CFI to fake perturbations would simply reflect the robustness of RML, RULS, and RDWLS to potential large levels of skewness in the fake perturbed data (see Table 6). However, under misspecified conditions a fit index value simply reflects how well the model is able to reconstruct the relationships in the observed data, which are not necessarily those that are actually represented in the true population. The graphical analysis of the ARB statistic clearly showed that the correlation bias was largely affected by fake perturbations with the largest effect being observed for the extreme faking condition. By contrast, ARB was not affected by sample size and type of estimation conditions. Unlike the CFI index, NFI was in general more in line with the ARB patterns.

By taking into account the results of our simulation study we proposed a new qualitative criterion (FC) which acts as a safety warning for faking data. We recommend including this criterion in the ideal set of model fit indices to evaluate the effect of potential fake observations in the data. However, it is natural to ask why the FC criterion seemed to work when applied to fake data. In particular, why do NFI and CFI show different behaviors to fake data at least when robust estimation procedures are considered? In what follows, we provide a tentative answer to this relevant question. To begin, we first notice that both NFI and CFI are incremental fit indices that are based on a transformation of the χ^2 values of the target model and the baseline model (or null model), respectively. The baseline model is a model that specifies that all measured variables are uncorrelated. In particular, NFI indicates the improvement in fit realized by moving from the baseline model to the target model

$$\text{NFI} = \frac{\chi^2_{\text{base}} - \chi^2_{\text{target}}}{\chi^2_{\text{base}}}.$$

By contrast, CFI is an incremental fit index based on the non-centrality parameter (Bentler 1990; McDonald and Marsh 1990). The non-centrality parameter is calculated by substracting the *df* of the model from its χ^2 value. CFI takes the following form

$$\text{CFI} = \frac{(\chi^2_{\text{base}} - df_{\text{base}}) - (\chi^2_{\text{target}} - df_{\text{target}})}{\chi^2_{\text{base}} - df_{\text{base}}}.$$

Table 6 Expected skewness in the fake perturbed data as a function percentage of replacements (K) and faking models (FM)

K by FM (%)	Uninformative	Slight	Extreme
0	0.000	0.000	0.000
25	−0.141	−0.119	−0.188
50	−0.501	−0.346	−0.710
75	−0.956	−0.581	−1.508
100	−1.229	−0.592	−2.397

and allows for an adjustment that takes into account model parsimony. However, in models for which χ^2 is larger than the corresponding *df*, which is likely to represent the great majority of models tested in empirical researches, the NFI tends to be more conservative than CFI. Figure 8 exemplifies this situation. In the graphical representation we assume that the proportion between χ^2_{target} and χ^2_{base} is kept constant to a fixed value 0.2, that is to say, the Chi-square value of the baseline (null) model is five times larger than that of the target model. In this scenario, the NFI function takes a constant value (0.8), whereas CFI is inversely related to the original baseline χ^2_{base} value. Because the CFI function always dominates the NFI function, the NFI index turns out to be more conservative than CFI. In general, this dominance pattern does not change if we consider different proportions between the target and baseline models such that $\chi^2 > df$. This may implicitly explain why the FC criterion distinguished between fake data and honest data in our simulation study.

Another interesting issue regards the relationship between fake data and skewed data. Because faking good entails some level of skewness in the data, our results seemed clearly related to some findings about the performance of robust estimation procedures under skewed or kurtotic ordinal data. For example, Flora and Curran (2004) evaluated in a Monte Carlo simulation study the overall bias in parameter estimation for a factorial model (similar to our model M2) fitted on simulated five-point rating data as a function of sample size and level of skewness (low: 0.75 vs. moderate: 1.25). The authors showed that in all conditions the overall bias in parameter estimation was less than 5 % when a robust WLS was used to estimate the parameters. Similarly, in an extensive simulation study Yang-Wallentin et al. (2010) showed as the shape of the distribution categories for different kind of ordinal data did not seem to make any difference for the average relative bias in parameter estimates and that, in particular, all the estimation methods except for WLS had very similar good performances.

Apparently, these findings seem to be in contrast with what we observed in our study where correlation bias as measured by ARB was largely affected by fake corrupted skewed data. However, because the dependent variables analyzed in these studies were clearly different (parameter estimate relative bias vs. correlation bias) a direct comparison between the two sets of results is difficult and delicate. All the same, it is interesting to note that when more extreme levels of data skewness (>2) were considered in simulation studies (e.g., Forero et al. 2009), the overall bias was positive and larger than 10 % also for robust methods (e.g., DWLS). In general, these results demonstrated that the pattern of relative biases, both from the estimates and the standard errors, depended on a complex

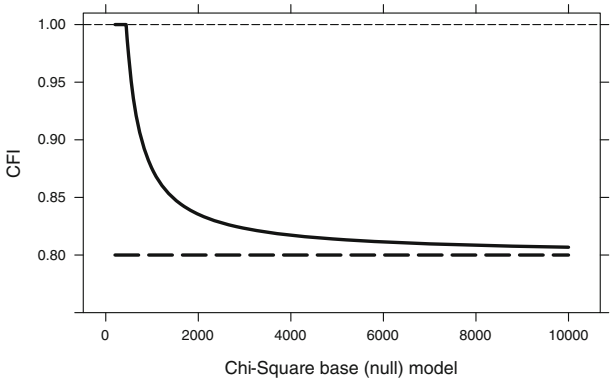

Fig. 8 CFI and NFI values as a function of χ^2_{base}. The proportion between χ^2_{target} and χ^2_{base} is kept constant to a fixed value 0.2. The degrees of freedom of the two models are the ones computed for Model 2 (target model: 88; baseline model: 105). Continuous (resp. dashed) line represents the CFI (resp. NFI) function

pattern of high-order interactions which had a slightly different effect on each estimation method. In sum, we believe that a deeper understanding of the interaction between the impact of fake data on parameter estimate relative bias or structural/correlational bias is necessary and worthy of further investigation in the study of factorial models for ordinal data.

5.1 Limitations and directions for future study

As with other Monte Carlo studies, our investigation involves simplifying decisions that result in lower external validity such as, for example, homogeneous loadings and the assumption that restricts the conditional replacement distribution to satisfy the conditional independence assumption. Unfortunately, this restriction clearly limits the range of empirical faking processes that can be mimicked by the current SGR simulation procedure. Therefore, although encouraging, the promise of this approach should be examined across more varied conditions. We acknowledge that more work still needs to be done.

References

Beauducel, A., Herzberg, P.Y.: On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. Struct. Equ. Model. **13**, 186–203 (2006)

Bentler, P.M.: Comparative fit indexes in structural models. Psychol. Bull. **107**, 238–246 (1990)

Bentler, P.M.: EQS Structural Equations Program Manual. Multivariate Software, Encino (1995)

Bentler, P.M., Bonett, D.G.: Significance tests and goodness of fit in the analysis of covariance structures. Psychol. Bull. **88**, 588–606 (1980)

Browne, M.W., Cudeck, R.: Alternative ways of assessing model fit. In: Bollen, K.A., Long, J.S. (eds.) Testing Structural Equation Models, pp. 136–162. Sage, Beverly Hills (1993)

Cohen, J.: Statistical Power Analysis for the Behavioral Sciences, 2nd edn. Lawrence Erlbaum Associates, Hillsdale (1988)

Curran, P.J., Bollen, K.A., Paxton, P., Kirby, J., Chen, F.: The noncentral Chi-square distribution in misspecified structural equation models: Finite sample results from a Monte Carlo simulation. Multivar. Behav. Res. **37**, 1–36 (2002)

Ding, L., Velicer, W.F., Harlow, L.L.: Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. Struct. Equ. Model.: A Multidiscip. J. **2**(2), 119–143 (1995)

Dobson, A.J.: An Introduction to Generalized Linear Models, 2nd edn. Chapman & Hall/CRC Press, Boca Raton (2002)

Dolan, C.V.: Factor analysis of variables with 2, 3, 5 and 7 response categories: a comparison of categorical variable estimators using simulated data. Br. J. Math. Stat. Psychol. **47**, 309–326 (1994)

Donovan, J.J., Dwight, S.A., Schneider, D.: The impact of applicant faking on selection measures, hiring decisions, and employee performance. J. Bus. Psychol. **29**, 1–15 (2014)

Fan, X., Felsovalyi, A., Sivo, S.A., Keenan, S.: SAS for Monte Carlo Studies: a Guide for Quantitative Researchers. SAS Institute Inc, Cary (2002)

Fan, X., Sivo, S.A.: Sensitivity of fit indices to model misspecification and model types. Multivar. Behav. Res. **42**, 509–529 (2007)

Fan, X., Wang, L.: Effects of potential confounding factors on fit indices and parameter estimates for true and misspecified SEM models. Educ. Psychol. Meas. **58**, 699–733 (1998)

Ferrando, P.J.: Factor analytic procedures for assessing social desirability in binary items. Multivar. Behav. Res. **40**, 331–349 (2005)

Ferrando, P.J., Anguiano-Carrasco, C.: Assessing the impact of faking on binary personality measures: an IRT-based multiple-group factor analytic procedure. Multivar. Behav. Res. **44**, 497–524 (2009)

Ferrando, P.J., Anguiano-Carrasco, C.: A structural modelbased optimal person-fit procedure for identifying faking. Educ. Psychol. Meas. **73**, 173–190 (2013)

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

Forero, C.G., Maydeu-Olivares, A., Gallardo-Pujol, D.: Factor analysis with ordinal indicators: a Monte Carlo study comparing DWLS and ULS estimation. Struct. Equ. Model. **16**, 625–641 (2009)
Fox, J.-P., Meijer, R.R.: Using item response theory to obtain individual information from randomized response data: an application using cheating data. Appl. Psychol. Meas. **32**, 595–610 (2008)
Furnham, A.: Response bias, social desirability and dissimulation. Personal. Individ. Differ. **7**, 385–400 (1986)
Gray, N.S., MacCulloch, M.J., Smith, J., Morris, M., Snowden, R.J.: Forensic psychology: violence viewed by psychopathic murderers. Nature **423**, 497–498 (2003)
Helton, J.C., Johnson, J.D., Salaberry, C.J., Storlie, C.B.: Survey of sampling based methods for uncertainty and sensitivity analysis. Reliab. Eng. Syst. Saf. **91**, 1175–1209 (2006)
Holden, R.R., Book, A.S.: Using hybrid Rasch-latent class modeling to improve the detection of fakers on a personality inventory. Personal. Individ. Differ. **47**, 185–190 (2009)
Hopwood, C.J., Talbert, C.A., Morey, L.C., Rogers, R.: Testing the incremental utility of the negative impression-positive impression differential in detecting simulated personality assessment inventory profiles. J. Clin. Psychol. **64**, 338–343 (2008)
Hu, L., Bentler, P.M.: Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. Psychol. Methods **3**, 424–453 (1998)
Jöreskog, K.: New developments in LISREL: analysis of ordinal variables using polychoric correlations and weighted least squares. Qual. & Quant. **24**, 387–404 (1990)
Jöreskog, K., Sörbom, D.: LISREL V: analysis of Linear Structural Relationships by the Method of Maximum Likelihood. National Educational Resources, Chicago (1981)
Jöreskog, K., Sörbom, D.: LISREL VI User's Guide, 3rd edn. Scientific Software, Mooresville (1984)
Jöreskog, K., Sörbom, D.: LISREL 8: user's Reference Guide. Scientific Software International Inc, Lincolnwood (1996a)
Jöreskog, K., Sörbom, D.: PRELIS 2: user's Reference Guide. Scientific Software International Inc, Lincolnwood (1996b)
Kenny, D.A., McCoach, D.B.: Effect of the number of variables on measures of fit in structural equation modeling. Struct. Equ. Model. **10**, 333–351 (2003)
Lee, S.-Y., Poon, W.-Y., Bentler, P.M.: Full maximum likelihood analysis of structural equation models with polytomous variables. Stat. Probab. Lett. **9**, 91–97 (1990)
Leite, W.L., Cooper, L.A.: Detecting social desiderability bias using factor mixture models. Multivar. Behav. Res. **45**, 271–293 (2010)
Lombardi, L., Pastore, M.: Sensitivity of fit indices to fake perturbation of ordinal data: a sample by replacement approach. Multivar. Behav. Res. **47**, 519–546 (2012)
Lombardi, L., Pastore, M.: sgr: a package for simulating conditional fake ordinal data. R. J. **6**, 164–177 (2014)
MacCann, C., Ziegler, M., Roberts, R.D.: Faking in personality assessment: Reflections and recommendations. *New Perspect. Faking Personal. Assess.*, 309–329 (2011)
Marshall, E.: Scientific misconduct. How prevalent is fraud? That's a million-dollar question. Science. **290**(5497), 1662–1663 (2000)
McCullagh, P., Nelder, J.A.: Generalized Linear Models. Chapman and Hall, London (1989)
McDonald, R.P., Marsh, H.W.: Choosing a multivariate model: Noncentrality and goodness of fit. Psychol. Bull. **107**, 247–255 (1990)
McFarland, L.A., Ryan, A.M.: Variance in faking across noncognitive measures. J. Appl. Psychol. **85**, 812–821 (2000)
Muthén, B.: A general structural equation model with dichotomous, ordered categorical and continuous latent variables indicators. Psychometrika **49**, 115–132 (1984)
Muthén, B., Kaplan, D.: A comparison of some methodologies for the factor analysis of non-normal Likert variables: a note on the size of the model. Br. J. Math. Stat. Psychol. **45**, 19–30 (1992)
Pastore, M., Lombardi, L.: The impact of faking on Cronbach's alpha for dichotomous and ordered rating scores. Qual. & Quant. **48**, 1191–1211 (2014)
Paulhus, D.L.: Two-component models of socially desirable responding. J. Personal. Soc. Psychol. **46**, 598–609 (1984)
Paulhus, D.L.: Measurement and control of response bias. In: Robinson, J.P., Shaver, P.R., Wrightsman, L.S. (eds.) Measures of Personality and Socialpsychological Attitudes, pp. 17–59. Academic Press, New York (1991)
Paxton, P., Curran, P.J., Bollen, K.A., Kirby, J., Chen, F.: Monte Carlo experiments: design and implementation. Struct. Equ. Model. **8**, 287–312 (2001)
Pek, J., MacCallum, R.C.: Sensitivity analysis in structural equation models: cases and their influence. Multivar. Behav. Res. **46**, 202–228 (2011)

Ridgon, E.E., Ferguson, C.E.: The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. J. Mark. Res. **28**, 491–497 (1991)
Rosse, J.G., Stecher, M.D., Miller, J.L., Levin, R.A.: The impact of response distortion on preemployment personality testing and hiring decisions. J. Appl. Psychol. **83**(4), 634–644 (1998)
Schermelleh-Engel, K., Moosbrugger, H., Mller, H.: Evaluating the fit of structural equation models: tests of significance and descriptive goodness- of-fit measures. Methods Psychol. Res. Online **8**(2), 23–74 (2003)
Steiger, J.H., Lind, J.C.: *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA (1980, May)
Tucker, L.R., Lewis, C.: A reliability coefficient for maximum likelihood factor analysis. Psychometrika **38**, 1–10 (1973)
Van der Geest, S., Sarkodie, S.: The fake patient: a research experiment in a Ghanaian hospital. Soc. Sci. & Med. **47**, 1373–1381 (1998)
Wood, S.N.: Generalized Additive Models. Taylor and Francis Group, Boca Raton (2006)
Yang-Wallentin, F., Jöreskog, K., Luo, H.: Confirmatory factor analysis of ordinal variables with misspecified models. Struct. Equ. Model. **17**, 392–423 (2010)
Zickar, M.J., Drasgow, F.: Detecting faking on a personality instrument using appropriateness measurement. Appl. Psychol. Meas. **20**, 71–87 (1996)
Zickar, M.J., Robie, C.: Modeling faking good on personality items: an item-level analysis. J. Appl. Psychol. **84**, 551–563 (1999)
Zickar, M.J., Gibby, R.E., Robie, C.: Uncovering faking samples in applicant, incumbent, and experimental data sets: an application of mixed-model item response theory. Organ. Res. Methods **7**, 168–190 (2004)
Ziegler, M., Buehner, M.: Modeling socially desirable responding and its effects. Educ. & Psychol. Meas. **69**, 548–565 (2009)