Multivariate Behavioral Research

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hmbr20

Sensitivity of Fit Indices to Fake Perturbation of Ordinal Data: A Sample by Replacement Approach

Luigi Lombardi [a] & Massimiliano Pastore [b]

[a] Department of Cognitive Science and Education, University of Trento

[b] Department of Developmental and Social Psychology, University of Padova

Version of record first published: 15 Aug 2012

To cite this article: Luigi Lombardi & Massimiliano Pastore (2012): Sensitivity of Fit Indices to Fake Perturbation of Ordinal Data: A Sample by Replacement Approach, Multivariate Behavioral Research, 47:4, 519-546

To link to this article: http://dx.doi.org/10.1080/00273171.2012.692616

Multivariate Behavioral Research, 47:519–546, 2012

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2012.692616

Psychology Press
Taylor & Francis Group

Sensitivity of Fit Indices to Fake Perturbation of Ordinal Data: A Sample by Replacement Approach

Luigi Lombardi
Department of Cognitive Science and Education, University of Trento

Massimiliano Pastore
Department of Developmental and Social Psychology, University of Padova

In many psychological questionnaires the need to analyze empirical data raises the fundamental problem of possible fake or fraudulent observations in the data. This aspect is particularly relevant for researchers working on sensitive topics such as, for example, risky sexual behaviors and drug addictions. Our contribution presents a new probabilistic approach, called Sample Generation by Replacement (SGR), to address the problem of evaluating 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-good ordinal data. We used SGR to perform a simulation study involving 3 different SEM models, 2 sample size conditions, and 2 estimation methods: maximum likelihood (ML) and weighted least squares (WLS). Our results show that the incremental fit indices (CFI, NNFI, and NFI) are clearly more sensitive to fake perturbation than the absolute fit indices (GFI, AGFI, and ECVI). Overall, NFI turned out to be the best and most reliable fit index. We also applied SGR to real behavioral data on (non)compliance in liver transplant patients.

Correspondence concerning this article should be addressed to Luigi Lombardi, Department of Cognitive Science and Education, University of Trento, Corso Bettini 31, I-38068 Rovereto (TN), Italy. E-mail: luigi.lombardi@unitn.it

A major problem in psychological measurements is that in some circumstances there is no basis to assume that participants are responding honestly. In real-life contexts, some individuals tend to distort their behaviors or actions in order to reach specific goals. 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 (Anderson, Warner, & Spector, 1984). Similarly, in the administration of diagnostic tests individuals often attempt to malinger posttraumatic stress disorder in order to secure financial gain and/or treatment or to avoid being charged with a crime (Hall & Hall, 2007).

Researchers interested in the study of human behavior in contexts like psychology (Furedy & Ben-Shakhar, 1991; Hopwood, Talbert, Morey, & Rogers, 2008; Lykken, 1960; Sartori, Agosta, Zogmaister, Ferrara, & Castiello, 2008), organizational and social science (Van der Geest & Sarkodie, 1998), psychiatry (Beaber, Marston, Michelli, & Mills, 1985), forensic medicine (Gray, MacCulloch, Smith, Morris, & Snowden, 2003; Mossman & Hart, 1996), scientific frauds (Marshall, 2000), and economics (Crawford, 2003; Sobel, 1985) face similar problems when analyzing and interpreting empirical data. In particular, possible fake data confront the researcher with a crucial question: If data included fake data points, would the answer to the research question be different from what it actually is? Even in the easiest case—that is, randomly fake data—the answer is not necessarily obvious as even the random perturbation of data constitutes biased information, which weakens the accuracy of scientific inferences.

A case of particular empirical interest in multivariate analysis of behavioral data is the situation in which a researcher wants to evaluate the uncertainty associated to the acceptability of a given structural equation model (SEM) as a result of propagation through the model of fake observations in input data. In this case the crucial question can be rewritten as the following: If the data contained $q\%$ fake observations, what would the chance be that the model is still a good one? A variety of fit indices can be used to evaluate the overall fit of a structural equation model (e.g., Browne & Cudeck, 1993; Hu & Bentler, 1998; Jöreskog & Sörbom, 1996a). Because fit indices are usually designed to detect model misspecification, but they are not designed to detect perturbation in the data, it is certainly legitimate to wonder whether fit indices are reliably sensitive also to fake observations. In particular, 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 data perturbation (i.e., presence of fake observations in the data set).

The great majority of past research on structural equation modeling has focused on several aspects related to the use and interpretation of model fit indices. Most studies based on Monte Carlo simulations examined the behaviors

of the fit indices under different data and model conditions, such as, for example, sample size, continuous versus ordinal data, estimation methods, model misspecification, and model types (e.g., Enders & Finley, 2003; Fan & Sivo, 2005, 2007; Fan, Thompson, & Wang, 1999; Fan & Wang, 1998; Gerbing & Anderson, 1993; Hu & Bentler, 1998, 1999; Marsh, Hau, & Wen, 2004; Wu & West, 2010; Yu & Muthén, 2002). However, there is little information about how fit indices will be sensitive to fake perturbations or how fake data may interact with other relevant factors to affect model fit. In this article we try to fill this gap by proposing a new probabilistic approach, called Sample Generation by Replacements (SGR), to deal with possibly fake data in SEM models. In particular, we examined in an SGR simulation study the sensitivity of eight 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-good ordinal data in three different SEM models, two sample size conditions, and two different estimation methods. In the simulation design, SGR was used to generate different levels of ordinal data perturbations based on a simple model of mimicking faking good behaviors (deception).

The article is organized as follows: The first part of the article outlines the SGR approach and introduces the models of faking and the target SEM models used in this study. The second part describes the SGR simulation and reports results about the fit indices' performances. The third part illustrates our method with an application to real data about (non)compliance in transplant patients. Finally, the fourth part presents conclusions and some relevant comments about limitations, potential new applications, and extensions of the SGR approach.

SAMPLE GENERATION BY REPLACEMENT (SGR)

SGR is a probabilistic resampling procedure that can be applied to discrete/ordinal data with a restricted number of values (e.g., Likert-type scales) and generalizes a previous deterministic procedure to simulate data replacement in discrete data arrays (Lombardi, Pastore, & Nucci, 2004). Variables characterized by an ordinal level of measurement are common in many empirical investigations within the social and behavioral sciences. This type of variables are also used to assess many psychological constructs based on self-reported measures. So, ordinal variables would seem to be a natural choice to study the effect of fake responses in empirical data.

With regard to the fake-data problem in general, we think of the original data as being represented by an $I \times J$ matrix $\mathbf{D}$, that is to say, I observations

(participants) each containing J elements (participant's responses). We assume that entry d_{ij} of $\mathbf{D}$ $(i = 1, \ldots, I; j = 1, \ldots, J)$ takes 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 pattern of responses of participant i. The response pattern $\mathbf{d}_i$ is a multidimensional ordinal random variable with probability distribution $p(\mathbf{d}_i|\boldsymbol{\theta}_M)$, where $\boldsymbol{\theta}_M$ indicates the vector of parameters of the probabilistic model of the data. Moreover, we assume that the response patterns are independent and identically distributed (i.i.d.) observations. Therefore, the data matrix $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_I]^T$ is drawn from the joint probability distribution

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

In the multivariate latent variable framework there are two main approaches for modeling ordinal variables according to Equation (1). The first is the Underlying Variable Approach (UVA) developed within the structural equation modeling framework (Jöreskog & Sörbom, 1996b; Muthén, 1984). This approach assumes that the observed ordinal variables are treated as metric through assumed underlying normal variables. In the UVA context the vector of parameters $\boldsymbol{\theta}_M$ represents the true population parameters of an SEM model. The second approach is Item Response Theory (IRT) where the probabilistic model of the data is characterized by the true latent parameters $\boldsymbol{\theta}_M$ of a graded response IRT model (Moustaki & Knott, 2000; Samejima, 1969). Because in this contribution we are interested in evaluating SEM-based fit indices, we limit our attention to the first approach.

The main idea of 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. Let $\mathbf{f}_i$ be the $(1 \times J)$ array of $\mathbf{F}$ denoting the 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

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

where $\boldsymbol{\theta}_F$ indicates the vector of parameters of the probabilistic faking model. In the conditional replacement probability distributions we assume that each fake response f_{ij} only depends on the corresponding data observation d_{ij} and the model parameter $\boldsymbol{\theta}_F$. 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

$$p(\mathbf{F}|\mathbf{D}, \boldsymbol{\theta}_F) = \prod_{i=1}^{I} p(\mathbf{f}_i|\mathbf{d}_i, \boldsymbol{\theta}_F) \tag{3}$$

$$= \prod_{i=1}^{I} \prod_{j=1}^{J} p(f_{ij}|d_{ij}, \boldsymbol{\theta}_F). \tag{4}$$

It is important to note that the faking model integrates together two different kinds of information: (a) the observed data **D** representing variables' features and relations generated according to Equation (1) and (b) the model parameter, $\boldsymbol{\theta}_F$, which characterizes some relevant properties of the faking model. In general, $\boldsymbol{\theta}_F$ 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). In sum, SGR is characterized by 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 and the model representing the faking process to perturb the data. By repeatedly sampling data from Equations (1–4) we can generate the so called *fake data sample* (FDS). We can then study the distribution of some relevant statistics computed on this FDS.

SIMULATION STUDY

The SGR approach can be easily reformulated to study the effect of possible fake data on the performances of SEM-based fit indices. In this special case, we assume that the original data **D** has been generated by a target SEM model (UVA framework). More precisely, **D** is a random sample from the statistical population determined by the true population parameters $\boldsymbol{\theta}_M$ of the SEM model. In this context, an SGR analysis allows us to evaluate SEM-based fit indices under the corresponding FDS. The perturbation is carried out by means of a joint replacement probability distribution that mimics the faking process of interest. Finally, the distribution of results for the fit indices is evaluated and eventually compared with the results observed for the original data sets prior to any fake perturbation.

In the following two sections we introduce the model of faking and the target SEM models that we used in this study for representing the faking process and the original data generation processes, respectively.

A Model of Faking

Fake data may alter a large variety of self-report measures. This problem is particularly relevant for researchers working on sensitive topics such as, for example, rash driving, risky sexual behavior, drug addictions, tax evasion, political preferences, and personnel selection. In this article we limit our attention to a simple, but important scenario of faking: the so called fake-good (McFarland & Ryan, 2000; Paulhus, 1984). Note that the fake-good (as well as the fake-bad) scenario entails a conditional replacement model in which the conditioning is a function of response polarity.

We used a simple parametrized replacement distribution to model the fake-good scenario described earlier. The model is called the *uniform support fake-good distribution* and represents a context in which responses are exclusively subject to positive feigning: $f_{ij} \geq d_{ij}$ $(i = 1, \ldots, I; j = 1, \ldots, J)$. In particular,

$$p_g(f_{ij} = q' | d_{ij} = q, \theta_F) = \begin{cases} 1, & q = q' = Q \\ \dfrac{\theta_F}{Q - q}, & 1 \leq q < q' \leq Q \\ 1 - \theta_F, & 1 \leq q = q' < Q \\ 0, & 1 \leq q' < q \leq Q \end{cases} \tag{5}$$

with θ_F being the overall probability of replacement. Equation (5) denotes the conditional probability of replacing an original observed value q in entry (i, j) of **D** with the new value q'. Note that this model does not allow us to substitute the original observed value with lower ones. A particular case is when $\theta_F = 0$. For this special condition the fake data matrix **F** reduces to the original data matrix **D** (see Equation (5)). Moreover, notice that the replacement model is characterized by a uniform probabilistic kernel. More precisely, in the fake-good model all the values $q' > q$ are assumed to be equally likely in the process of replacement. In sum, the model represents a *purely random* but *polarized* malingering *process* (PRPP).

Some important comments are in order concerning the rationale behind the usage of PRPP to simulate fake data. First, PRPP is based on two basic properties: (a) the *principle of indifference* and (b) the *asymmetry of the faking behavior*. The first property reflects that in the absence of further knowledge all entries in **D** as well as all candidate replacement values are assumed to be equally likely in the process of replacement. In other words, PRPP assumes a kind of random world model that can be used whenever we deal with randomly fake data. The principle of indifference requires the simplest quantitative representation for the replacement process. The second property suggests that fake responses are

mainly due to asymmetrical processes. For example, fake-good (resp. fake-bad) responses can be characterized by a positive (resp. negative) polarity with respect to the original fake-uncorrupted responses.

The great majority of past and current research on faking has focused on the asymmetrical/qualitative properties of faking behavior (e.g., McFarland & Ryan, 2000; Paulhus, 1984). However, there is little knowledge about the distributional/quantitative characteristics of ordinal fake responses. In our opinion, PRPP can represent a good compromise between the lack of information about the distributional properties of faking and the well-known asymmetrical qualitative properties typical of behaviors such as malingering, defensiveness, and self-deception. Finally, although some empirical contexts may require different model assumptions as well as different fake distribution conditions, we wanted to understand the impact of fake data under the most simple and less invasive distributional conditions first.

Target SEM Models

We selected three target SEM models (see Figure 1) that are commonly encountered in applied research (Curran, Bollen, Paxton, Kirby, & Chen, 2002; Paxton, Curran, Bollen, Kirby, & Chen, 2001) to representing the process that generates the data prior to any fake perturbation. The first model, Model 1, contained 9 measured variables ($y_1, \ldots, y_9$) and 3 latent variables (η_1, η_2, and η_3). Each measured variable loaded on a single latent variable. Further, η_2 was regressed on η_1, and η_3 was regressed on η_2. The second model, Model 2, had the same basic structure as Model 1 but contained 15 measured variables ($y_1, \ldots, y_{15}$) with five indicators per latent variable. Finally, Model 3 contained 13 measured variables. The endogenous variables ($y_1, \ldots, y_9$) had 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.

We defined the values of population parameters to be homogeneous across all three model specifications. For Model 1, all factor loadings (λ^y) were set to .7 and error variances (θ^ϵ) were set to .5. The regression parameters among the latent variables (β) were set to a value of .6. For Model 2, all the values were exactly the same as those of Model 1 except for the addition of two measured variables per latent variable. Finally, for Model 3, we included four exogenous variables and set relative factor loadings (λ^x) to .95 and error variance (θ^δ) to .09. The regression parameters among the latent variables (β and γ) were set to a value of .6.

The models also differed in terms of propagation of fake perturbation in the data. 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,

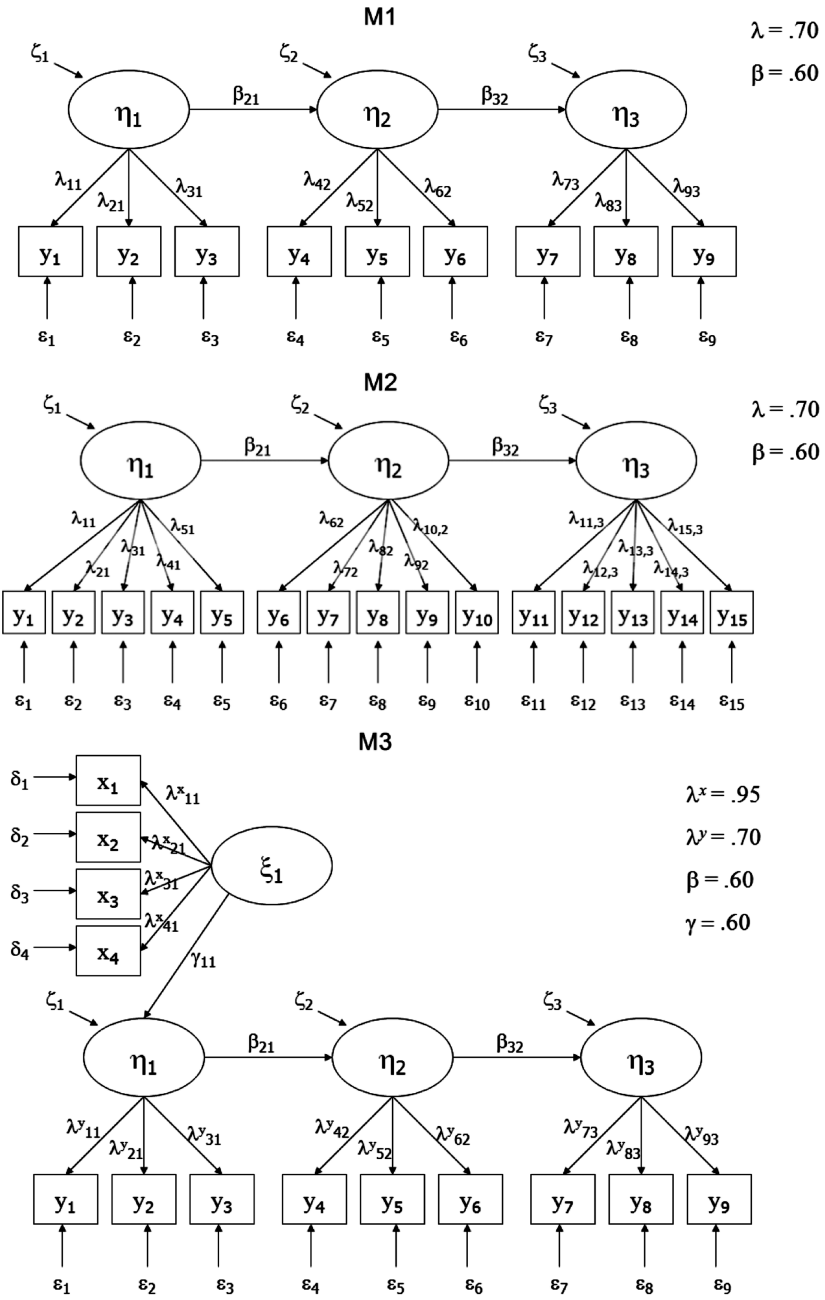

FIGURE 1 Target SEM models.

$y_1, \ldots, y_9$, only, whereas the exogenous variables, $x_1, \ldots, x_4$, were considered fake independent. In sum, for a fixed probability of replacement θ_F and a fixed sample size I, two relevant conditions were observed:

1. Model 1 and Model 2 were affected by the same overall probability θ_F of fake perturbation, but they differed in terms of total amount of fake data: $\theta_F \times (I \times 9)$ for Model 1 and $\theta_F \times (I \times 15)$ for Model 2.
2. Model 1 and Model 3 were affected by a different overall probability of fake perturbation (θ_F for Model 1 and $\theta_F \times \frac{I \times 9}{I \times (9+4)}$ for Model 3), but they did not differ in terms of total amount of fake data ($\theta_F \times (I \times 9)$ for both).

These two conditions will be separately evaluated in our simulation study. In particular, we expect that a good fit index should be less sensitive to Model 3 compared with 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 because model size (defined as the number of observed variables in the model) should not affect the performance of a fit index (e.g., Fan & Sivo, 2007; Kenny & McCoach, 2003).

Types of Fit Indices

Eight fit indices were examined in this study: 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). This group represents a collection of widely known and commonly used stand-alone fit indices in the SEM literature. The basic properties of each of the eight fit indices are summarized in Table 1. The definitions and reviews of these fit indices are easily available in the SEM literature (e.g., Browne & Cudeck, 1993; Fan & Wang, 1998; Hu & Bentler, 1998; Sun, 2005; Yuan, 2005).

Simulation Design and Data Conditions

Although other studies may typically involve a wider spectrum of SEM models, in this study we wanted to understand the impact of fake data on the fit indices under empirical scenarios that are commonly encountered in applied research (Curran et al., 2002; Paxton et al., 2001). This also means fitting the models directly on Likert-type data instead of standard continuous variables. In our simulation study we opted for a 5-point ordinal scale, which is very common in many empirical investigations within the social and behavioral sciences.

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 (1984) 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	$[0, 1]$

Note. 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.

Now we are in the position to provide all details of our simulation design. Four factors were systematically varied in a complete four-factorial design:

1. The model type (MT), at three levels: M_1, M_2, and M_3 (see Figure 1);
2. The sample size (I), at two levels: 100 and 200;
3. The estimation procedure (E), at two levels: maximum likelihood (ML) and full weighted least squares (WLS), also known as the asymptotically distribution free estimator; and
4. The percentage of replacements (K) for the endogenous variables in the SEM model at 11 levels: $0\%, 10\%, \ldots, 100\%$.

Let t, i, e, and k be distinct levels of factors MT, I, E, 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 132 combinations of levels (t, i, e, k) of the simulation design:

1. Set $AS = 0$.
2. Generate a raw-data set **D** with size i according to model t. The data generation was performed using a standard Monte Carlo (MC) procedure based on multivariate normal data (Fan, Felsovalyi, Sivo, & Keenan, 2002; Kaiser & Dickman, 1962).
3. Discretize **D** on a 5-point scale using the method described by Jöreskog and Sörbom (1996b).

4. Fit model t using the polychoric correlation matrix of the discretized data **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. Construct a fake data matrix **F** of the discretized data **D** using the conditional replacement probability with probability of replacement $\theta_F = \frac{k}{100}$ (see Equation (5)).
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 4,000 acceptable solutions; otherwise, go to Step 2.

This algorithm was used to generate 4,000 distinct matrices **F** for each combination of levels (t, i, e, k) of the SGR simulation design. This number of replications was chosen to achieve reasonable estimation stability in the tail regions of the fit indices. Finally, for each of the 4,000 perturbed data matrices **F** the eight fit indices were evaluated and the results saved for later analyses. Note that in Step 5, if $k = 0$, then the fake matrix **F** reduces to the data matrix **D** (Step 3) as the probability of replacement θ_F boils down to zero (see Equation (5)). The whole procedure generated a total of $528{,}000 = 4{,}000 \times 3 \times 2 \times 2 \times 11$ new fake matrices as well as an equivalent number of fit indices results.

Some important comments are in order concerning the discretization procedure adopted in Step 3. After sampling continuous data from the distribution described in Step 2, we transformed these samples into five-category ordinal data by applying a set of thresholds that remained constant across all data **D**. In our simulation design the discretization procedure was based on the maximum likelihood (ML) assumption (Flora & Curran, 2004). Because a five-category ordinal variable has four distinct thresholds, $-\infty < \alpha_1 < \alpha_2 < \alpha_3 < \alpha_4 < +\infty$, the normal quantiles, -1.53, -0.49, 0.49, and 1.53, were used as corresponding threshold values. The four quantiles were computed using the inverse of the binomial cumulative distribution function (CDF; Jöreskog & Sörbom, 1996b). Finally, the original continuous data set **D** was discretized into symmetrically distributed ordinal variables (Step 3) on the basis of the four threshold values.

Data Source and Statistical Analyses

Data were simulated using a combination of R scripts (R Development Core Team, 2010). Model fitting and estimation were implemented through LISREL package (Jöreskog & Sörbom, 1996a). For each replication, the relevant fit indices' results were saved for later analyses. Because the fit indices considered in this study showed a nonzero level of skewness and kurtosis (see Table 2 for

TABLE 2
Descriptive Statistics of the Fit Indices

Index (y)	*M*	*SD*	*Min*	*Max*	*Skewness*	*Kurtosis*	*Recoded Variable (y^*)*
GFI	0.90	0.04	0.71	0.99	−0.56	−0.48	$y^* = 1 - y$
AGFI	0.86	0.06	0.58	0.98	−0.60	−0.31	$y^* = 1 - y$
ECVI	0.98	0.49	0.33	3.38	0.64	−0.53	
SRMR	0.06	0.02	0.02	0.16	0.65	0.08	
RMSEA	0.04	0.03	0.00	0.18	0.13	−0.32	$y^* = y + .1$
CFI	0.97	0.07	0.00	1.00	−4.01	20.79	$y^* = 1 - y + .1$
NNFI	0.96	0.10	−25.99	9.83	−23.69	6,281.28	$y^* = 1 - y + max(y)$
NFI	0.90	0.10	−0.16	0.99	−2.13	5.15	$y^* = 1 - y + .1$

Note. The last column reports the recoding equation for the negative skewed indices. M = mean; SD = standard deviation. Fit indices described in the table note of Table 1.

some descriptive statistics), a natural choice to model this data would be the Gamma distribution (McCullagh & Nelder, 1989; Dobson, 2002; Wood, 2006). Therefore, generalized linear models (GLM) were used as primary statistical analysis to evaluate how the fit indices values were systematically influenced by the design factors. In particular, all the GLM models used in our analyses were based on the Gamma family with inverse link function. However, because a correct application of the Gamma family requires nonnegative data and symmetrically or positively skewed data distributions, we transformed all the indices that showed negative values or a negative skewness into new variables with correct ranges and skewnesses. The transformation was performed according to the recoding equations described in Table 2. Finally, all the Gamma regression models included the main factor terms (I, MT, E, and K) and all the interaction terms as independent variables.

The GLM analysis allows us to partition the deviance of a model fit index into different components contributed to the design factors. For each fit index, the deviance attributable to a factor (D_{source}) and the null deviance (D_{null}), that is, the deviance for the GLM model with just a constant term, were used to obtain the effect size for the GLM factor:

$$\varphi = 100 \times \frac{D_{source}}{D_{null}}.$$

The φ statistic can be understood as that percentage of deviance explained in a dependent variable attributable to a factor in the GLM model. In other words, the φ value of a dependent variable represents the sensitivity of that dependent variable to different design factors. Note that this statistic is invariant to the recoding scheme reported in Table 2 as the transformed variables only affect the sign of the parameter estimates of the GLM models.

RESULTS

Table 3 reports the measures of fit for the Gamma GLM models used to study what design factors influence the variation of the fit indices values in the simulation design. Table 4 presents the portion of deviance φ explained in a fit index attributable to the design factors and their interactions. An ideal fit index should be sensitive to fake perturbations and 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 (I = 100, 200). More specifically, we would expect that a large proportion of deviance φ in a fit index would be attributable to the relevant factor K and to the difference between M_1 and M_3. However, 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 fake data sets show a strong level of asymmetries in the marginal distributions of the ordinal values due to the asymmetric perturbation process. In this latter scenario WLS might still represent a more robust estimation procedure, although some problems may occur when it is used with small sample sizes (Flora & Curran, 2004). 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 considered in our simulation study are systematically affected (to different degrees) by sample size.

As shown in Table 4, for the model type conditions, four indices (GFI, AGFI, ECVI, and NFI) showed high sensitivity to different SEM models with their proportion of deviance attributable to MT being about 24% or higher, whereas SRMR, RMSEA, CFI, and NNFI were clearly less sensitive to this factor. To better understand the behaviors of the fit indices 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 and (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 shoud 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 shows a simple 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). A good fit index should have a negative value for this statistic. A quick inspection of Table 5 immediately shows that the behaviors of AGFI, RMSEA, NNFI, and NFI are consistent with

TABLE 3
Measures of Fit for the Gamma GLM Models

Measures of Fit	*GFI*	*AGFI*	*ECVI*	*SRMR*	*RMSEA*	*CFI*	*NNFI*	*NFI*
D_{null}	126,593.30	97,286.95	141,784.50	32,251.37	24,527.62	101,701.98	83.33	95,729.18
D_{res}	20,721.73	20,721.73	4,234.85	9,420.89	12,018.10	26,668.92	37.82	20,575.38
Gen r^2	0.84	0.79	0.97	0.71	0.51	0.74	0.55	0.79
AIC	−2,808,253	−2,358,206	−1,199,501	−3,505,775	−2,602,077	−2,146,891	−1,116,193	−1,750,072

Note. The *null deviance* (D_{null}) is the deviance for a model with just a constant term, whereas 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. M = mean; SD = standard deviation. Fit indices described in the table note of Table 1.

TABLE 4
Partitioning the Deviance (φ) of Goodness-of-Fit Indices

Source	*GFI*	*AGFI*	*ECVI*	*SRMR*	*RMSEA*	*CFI*	*NNFI*	*NFI*
K	0.89	1.13	0.21	16.01	0.29	18.14	6.49	30.00
I	38.10	49.67	42.55	45.95	3.31	6.44	2.20	12.38
MT	40.46	24.30	46.62	7.43	2.36	7.12	2.46	23.67
E	0.53	0.72	2.83	0.20	40.67	36.00	23.25	9.88
K by I	0.06	0.07	0.01	0.42	0.04	0.09	0.56	0.49
K by MT	0.03	0.02	0.00	0.36	0.06	1.32	1.73	0.96
I by MT	3.02	2.11	3.90	0.10	0.05	0.20	0.56	0.35
K by E	0.29	0.41	0.25	0.25	3.01	3.11	8.89	0.12
I by E	0.00	0.00	0.45	0.00	0.10	0.61	2.68	0.00
MT by E	0.18	0.21	0.13	0.03	0.94	0.57	2.31	0.42
K by I by MT	0.01	0.01	0.00	0.02	0.01	0.05	0.29	0.14
K by I by E	0.00	0.00	0.01	0.00	0.12	0.00	0.96	0.01
K by MT by E	0.06	0.05	0.01	0.02	0.05	0.08	1.78	0.07
I by MT by E	0.00	0.00	0.03	0.00	0.00	0.02	0.30	0.02
K by I by MT by E	0.00	0.00	0.00	0.00	0.00	0.04	0.17	0.00

Note. K = percentage of replacements; I = sample size; MT = model type; E = estimation procedure. Fit indices described in table note of Table 1.

this expectation, whereas GFI, ECVI, SRMR, and CFI are not. In particular, NFI showed the largest negative difference (−19.07), whereas ECVI was the index with the worst performance (11.13).

Finally, for the percentage of replacement conditions, SRMR, CFI, NNFI, and NFI showed high sensitivity to increasing amount of faking with their proportion of deviance attributable to K being about 16%, 18%, 6%, and 30%, respectively. The other four indices, GFI, AGFI, ECVI, and RMSEA were less sensitive to data replacements. In particular, factor K accounted for a very low amount of variation in ECVI and RMSEA (resp. only 0.21% and 0.29%). Interestingly, CFI, NNFI, and NFI were also sensitive to estimation procedure. However, the

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 vs. M2	43.52	23.26	54.91	9.43	0.30	2.21	1.10	3.29
(b) M1 vs. M3	42.94	28.35	43.78	0.42	3.10	3.77	0.51	22.36
(a-b)	0.58	−5.09	11.13	9.01	−2.80	−1.56	0.59	−19.07

Note. (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). Fit indices described in the table note of Table 1.

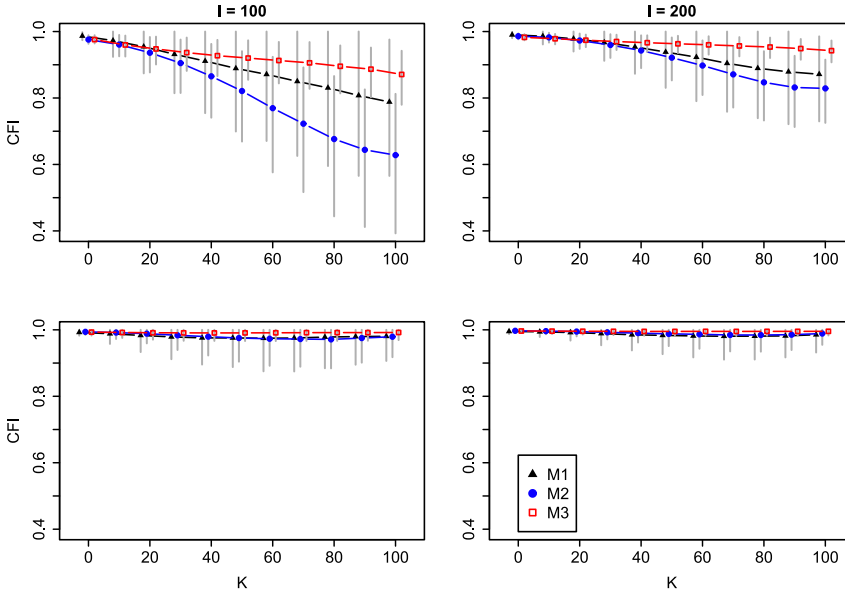

FIGURE 2 Means of Comparative Fit Index (CFI) as a function of percentage of replacements, models of faking, and sample size. Segments represent 95% interquantile intervals. Top panel: maximum likelihood estimation procedure. Bottom panel: weighted least square estimation procedure (color figure available online).

fit index with the highest sensitivity to estimation procedure was RMSEA with its proportion of deviance attributable to E being about 41%.

Based on the preliminary results presented in Tables 4 and 5 we would tentatively consider CFI, NNFI, and NFI as having the more ideal behaviors expected from a model fit index. However, to further explore the behaviors of these best performing fit indices, in the next section we graphically display their functional patterns and compare this new analysis with the results of this section.

Graphical Analysis

Figure 2 shows the observed means of CFI as a function of factors K, I, and E, respectively. Segments represent 95% interquantile intervals. For the model size condition, M_3 yielded on average better fits than the smaller model M_1 thus confirming the results reported in Table 5. For the sample size effect, the I = 200 condition yielded better performances in all three SEM models only when the ML estimation procedure was used (Figure 2, top panel). By contrast, the effects of K, I, and MT disappeared in the WLS condition (Figure 2, bottom panel). Very similar results can be observed for the NNFI index (Figure 3).

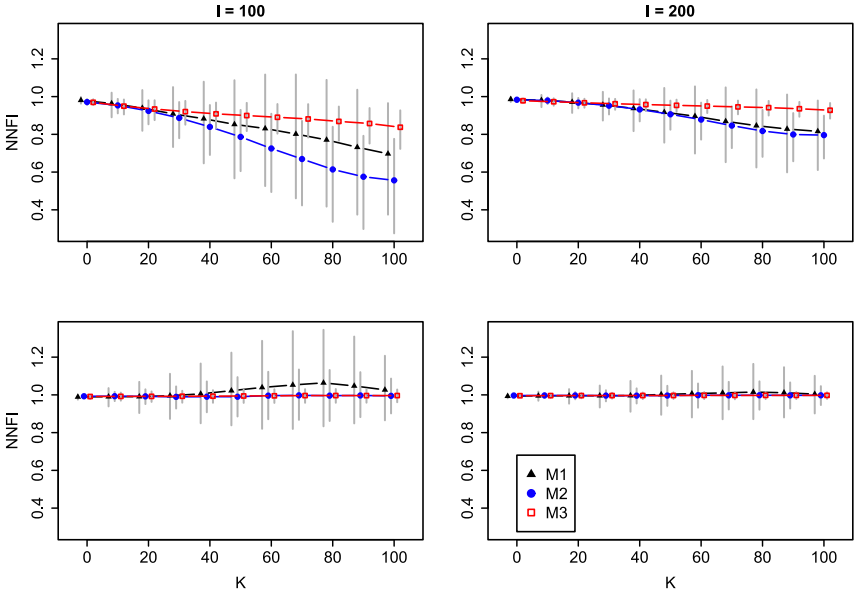

FIGURE 3 Means of Nonnormed Fit Index (NNFI) as a function of percentage of replacements, models of faking, and sample size. Segments represent 95% interquantile intervals. Top panel: maximum likelihood estimation procedure. Bottom panel: weighted least square estimation procedure (color figure available online).

Figure 4 presents the patterns of NFI. This index showed the largest sensitivity to fake perturbation. In particular, the NFI mean decreased by increasing levels of replacements (K), that is to say, it degraded with larger amounts of fake perturbations. NFI was also less sensitive to sample size and model type conditions. Finally, in the WLS condition the effect of replacements was clearly still present as well as the effect attributable to M_1 versus M_3 compared with that of M_1 versus M_2. Overall, NFI resulted as the best and most reliable fit index.

EMPIRICAL APPLICATION

The SGR approach is illustrated using data from a questionnaire about (non)compliance in liver transplant patients. The current section is divided into two subsections: the first introduces the empirical data set and the result of a simple SEM model fitted to the data; the second discusses how we can use SGR to compare the performances of two distinct generative models with respect to a faking-good scenario.

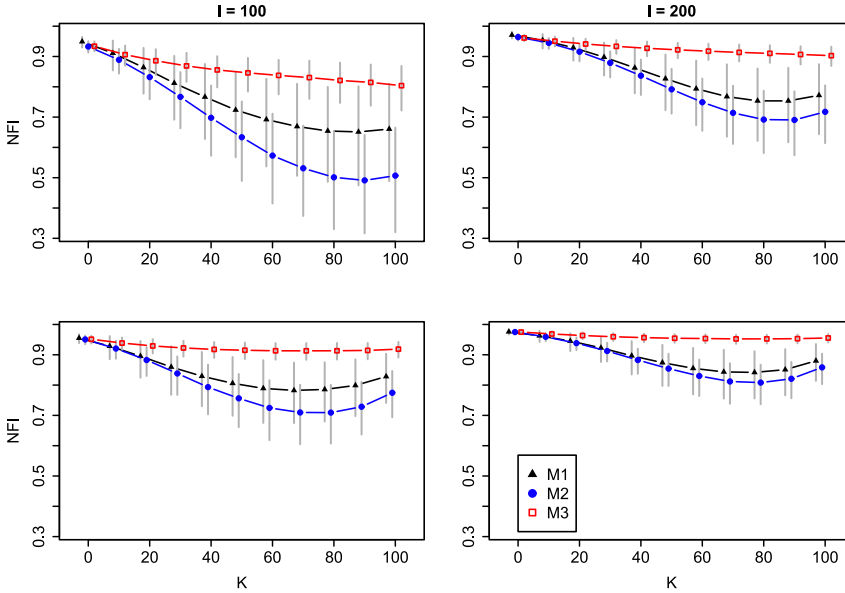

FIGURE 4 Means of Normed Fit Index (NFI) as a function of percentage of replacements, models of faking, and sample size. Segments represent 95% interquantile intervals. Top panel: maximum likelihood estimation procedure. Bottom panel: weighted least square estimation procedure (color figure available online).

Original Data Set and SEM Model

Progress in the techniques of transplantation has increased significantly life expectancy of many patients affected by serious diseases. Nevertheless transplantation invariably includes the acceptance of a lifelong pharmaceutical regimen in the absence of which every surgical effort is invalidated. The lack of adherence to therapies is one of the principal risk factors for patients after surgery (Matinlauri, Nurminen, Hockerstedt, & Isoniemi, 2005). Moreover, risk behaviors such as smoking and drinking alcohol while following the therapeutical regimen can seriously affect results of medical treatments (Cuadrado, Fabrega, Casafont, & Pons-Romero, 2005). In this context, social desirability factors may drastically limit the validity of self-report measures about risk behaviors. So, for example, a patient diagnosed with alcohol dependence who follows a pharmaceutical regimen after the liver transplant would deliberately answer fraudulently a question about drinking alcohol due to abstinence from ethanol (Foster et al., 1997).

In this application we studied the potential impact of fake-good responses on the relationships between compliance indicators and safe behaviors indicators in a group of 134 patients (30 women, 104 men) recruited in the local

transplantation center of Veneto district (Northeast Italy) during a period of 3 years (2003–2006). Ages ranged from 29 to 65, with a mean of 53.3 and a standard deviation of 7.6. Moreover, a relevant proportion of patients (84.4%) was suffering from alcoholic cirrhosis or Hepatitis C virus (HCV) cirrhosis at the time of the first visit. Data was collected using six self-report ordinal measures selected from a larger survey about (non)compliance in liver transplant patients:

- Safe behaviors items
  —I smoke often (0), sometimes (1), never (2).
  —I drink alcohol often (0), sometimes (1), never (2).
  —I use drugs often (0), sometimes (1), never (2).
- Compliance items about pharmacological treatments, medical visits, and patient-physician communication
  —I forget the treatment often (0), sometimes (1), never (2).
  —I forget the visit often (0), sometimes (1), never (2).
  —I forget communications often (0), sometimes (1), never (2).

The resulting (134×6) data matrix was subjected to an SEM model with two latent variables (one for each group of measures). More precisely, each measured variable loaded on its corresponding single latent variable (see Figure 5). Further, the compliance factor was regressed on the safe behaviors factor. The expected result is that a patient showing an high level for safe behaviors (resp. an high level for risk behaviors) should be characterized by a compliance (resp. noncompliance) profile.

SGR Analysis

In fitting the SEM models we opted for ML as the alternative full WLS is usually better suited for estimations based on larger sample sizes (Flora & Curran, 2004). Moreover, although it does not have theoretical justification for use with ordinal variables, ML still performed well in our SGR simulation study (see also Yang-Wallentin, Joreskog, & Luo, 2010). Specifically ML seemed to boost the sensitivity of fit indices to fake perturbations.

The result of the original SEM model was poor (CFI = .919, NFI = .897, NNFI = .847) and showed a nonsignificant value for the regression parameter denoting a lack of association between the two factors in the model. However, the observed result may have been affected by fake observations. This hypothesis was supported by the strong ceiling effects observed in the data (see Table 6).

An SGR analysis was used to evaluate the impact of eventual fake-good data on the SEM model's performance. In the first step, we defined two generative SEM models (M_0 and M_1) representing two alternative hypothetical true models for the data. The two generative models had the same structure of the original

TABLE 6
Descriptive Statistics for the Six Observed Ordinal Variables

	Often %	*Sometimes* %	*Never* %
Safe behaviors variables			
Smoking	14.93	28.36	56.72
Drinking alcohol	0	29.10	70.90
Drug consumption	0	21.64	78.36
Marginal %	*4.97*	*26.37*	*68.66*
Compliance variables			
Forgetting the treatment	0	49.25	50.75
Forgetting the visit	0	42.54	57.46
Forgetting communications	0	38.06	61.94
Marginal %	*0*	*43.28*	*56.72*

SEM model, but they differed with respect to the value of the structural parameter between the two latent variables. In particular, in model M_0 the structural parameter was set to 0.05 denoting a *weak* association between the safe behaviors latent dimension and the compliance latent dimension, whereas in model M_1 the structural parameter was set to 0.95 denoting a *strong* association between the two latent dimensions. All the other components (parameters' values) were set to identical values in M_0 and M_1 (see Figure 5). In particular, the loadings for the safe behaviors factor were all set to the hypothetical true value 0.6, whereas the

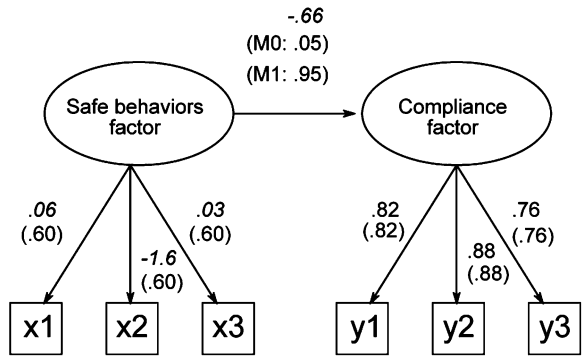

FIGURE 5 SEM model for the six ordinal variables: smoking (x_1), alcohol drinking (x_2), drug consumption (x_3), pharmacological treatments (y_1), medical visits (y_2), and patient-physician communication (y_3). The values reported in italic type represent the nonsignificant parameters of the original SEM model fitted on the observed data. The values in parentheses are the true parameters values of the two generative models M_0 and M_1.

loadings for the compliance factor were set to the original estimated values: 0.82 (pharmacological treatments), 0.88 (medical visits), and 0.76 (patient-physician communication), respectively. The two generative models were used to simulate new data without any ceiling effect for the safe behaviors variables.

In the second step we implemented the fake-good model (Equation (5)). In particular, the percentage of replacements (K) for the safe behaviors variables were set at 37 distinct levels ranging from 10% to 80% (by a step of 2%). In sum, we assumed fake-good perturbations only for the risk behaviors items. The compliance variables were, instead, assumed not to be fake dependent. Finally, for each generative model and each level of K, we generated 2,000 distinct fake data matrices and saved the resulting NFI values.

The results of the SGR analysis are shown in Figure 6. The figure represents the difference Δ_{NFI} between the estimated density values corresponding to the original NFI result (0.897) under the NFI distributions of M_1 and M_0 (respectively) as a function of percentage of replacements K. The results showed that the weak dependency model M_0 was preferred to the strong dependency model M_1 (negative Δ_{NFI}) up to a level of faking of approximately 60%. By contrast, the SGR results were on average more in favor of M_1 (positive Δ_{NFI}) when larger amounts of faking were considered. In other words, if we assume that a large proportion of patients ($\geq$ 60%) manipulated their answers on the risk behaviors items, then the observed goodness-of-fit result (NFI = 0.877) will be more consistent with a true SEM model representing a *strong* dependence

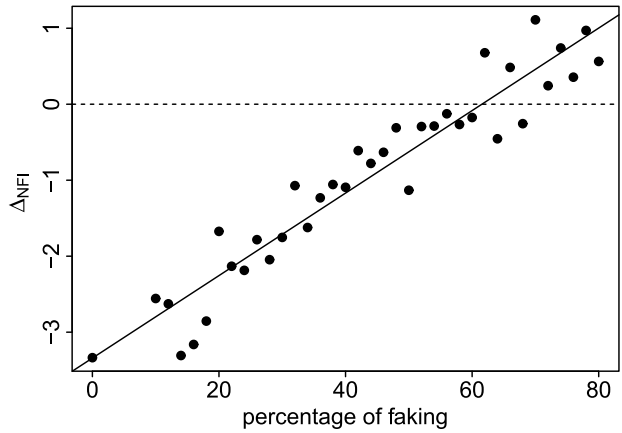

FIGURE 6 Difference Δ_{NFI} between the estimated density values corresponding to the original Normed Fit Index (NFI) result (0.897) under the NFI distributions of M_1 and M_0 as a function of percentage of replacements K. For each level k of replacement and each generative model, the simulated NFI distribution was approximated by using a kernel density estimator with Gaussian kernel smoothing (Sheather & Jones, 1991).

between the two latent dimensions. In this hypothetical scenario, a patient taking behavioral risks will also be a noncompliant patient.

DISCUSSION AND FINAL CONCLUSIONS

In situations where a model is fitted on empirical data containing possible fake measurements, a fit index that evaluates that model may not be very helpful in deciding whether or not it can be appropriate in representing the relationships under study. In particular, 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. 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. Previous research in this area has not adequately examined the issue of fit indices' sensitivity to fake data perturbation. As far as we know, this is the first time that the effect of fake data on fit indices has been systematically evaluated in a simulation study.

The results of our SGR simulation study lead us to believe that none of the fit indices considered in this article really stood out as having ideal behavioral patterns: sensitive to fake perturbations and/or faking model type but insensitive to model types and sample size. However, important local differences were observed between the indices. More specifically, three indices (CFI, NNFI, and NFI) showed considerable sensitivity to fake perturbation in the ML estimation condition, but only NFI resulted sensitive to fake data also in the WLS estimation condition. By contrast, GFI, AGFI, RMSEA, and ECVI were among the indices less sensitive to data replacement but largely sensitive to sample size and model size (e.g., M1 vs. M2). In particular, for the model size condition, the results show that ECVI would penalize larger models and reward smaller models even when the models were equated in terms of fake perturbation intensity. This result is not difficult to understand and immediately derives from the formal definition of this index. A quick inspection of the ECVI equation reveals that this absolute index includes a penalty term that is an increasing function of the number of estimated parameters, and this explains the observed ECVI pattern as described earlier. These results are also consistent with those discussed in other research about the sensitivity of absolute fit indices to sample size and model size (Breivik & Olsson, 2001; Fan & Sivo, 2007; Hu & Bentler, 1998; Kenny & McCoach, 2003).

SRMR yielded an intermediate performance. More precisely, this index showed a desirable sensitivity to fake perturbations but also a relevant sensitivity to sample size and model size (M_1 vs. M_2; see Table 5). In particular, SRMR had the tendency to yield better fit values (smaller value of SRMR) as the number

of observed variables increased in the model. Therefore, unlike ECVI, SRMR would penalize less complex smaller models.

Our results demonstrate empirically that the incremental fit indices used in our study (CFI, NNFI, and NFI) were clearly more sensitive to fake perturbation than the absolute fit indices (GFI, AGFI, and ECVI), at least when the ML estimation procedure was considered. Therefore, it is natural to ask why such a difference exists. In particular, what in the incremental fit indices disposes them to be affected more by increasing levels of replacements in the observed data? In what follows, we propose a tentative answer to this important question. Because all the fit indices are based on a transformation of the minimum statistic T_T of the maximum likelihood function $F_{ML}(\mathbf{R_F}, \hat{\Sigma}_\mathbf{F})$, we studied how the polychoric correlation matrix $\mathbf{R_F}$ was affected by increasing levels of fake replacements in the data. The main result was that the average correlations decreased by increasing the amount of fake data perturbations as fake observations usually tend to weaken the original relationships in the data. One obvious consequence is that the correlation matrices also tend toward the identity matrix **I**. In SEM literature **I** denotes an independent or null model that assumes zero population correlations among the observed variables. Note that absolute fit indices evaluate the model fit of the hypothesized model without a comparison with a baseline model, whereas relative fit indices measure the specific improvement in model fit of the hypothesized model relative to a baseline model, such as, for example, the null model (Bollen & Curran, 2006). In particular, comparative fit indices assign larger values (better fits) to those target models that have larger distances from the baseline model. Therefore, when the correlation matrix derived from the perturbed data tends toward matrix **I**, a comparative fit index will penalize the target model, thus reflecting a larger sensitivity to fake perturbation compared with absolute fit indices. However, the difference between some of the incremental fit indices (CFI and NNFI) and the absolute fit indices disappears when the WLS estimation procedure is considered. It is difficult to explain why in the WLS condition the two incremental fit indices were unsensitive to fake perturbation. A tentative answer might call for asymmetric representations in the data, which are typical of fake-good perturbations. In this condition WLS could have masked the effects of data replacements due to a sort of absorbtion effect. Moreover, it is also remarkable that in our simulation study full WLS outperformed ML in terms of number of acceptable solutions (approximately 30% more) in fitting the SEM models.

Nonetheless, we still believe that a deeper understanding of the interaction between fake observations and estimation methods is necessary and worthy of further investigation in the study of SEM-based fit indices. For example, a new simulation study could also involve additional estimation methods (e.g., robust WLS) as well as larger sample sizes to better evaluate some estimation methods (e.g., full WLS) that apparently require larger sample sizes

(maybe in the thousands) to fully avoid estimation biases (see Flora & Curran, 2004).

Implications for Applied Research and Future Directions

There are several specific implications of our findings with respect to using model fit indices in practice. First, our findings suggest that NFI outperforms all the other fit indices considered in our simulation study. Therefore, we recommend including NFI in the ideal battery of model fit indices to evaluate the effect of eventual fake observations in the data. Second, it is highly probable that fit indices are normally not able to distinguish between fake-good and fake-bad data. In a previous version of the article we also tested the fit indices to different faking conditions. In particular, in a preliminary simulation study all the fit indices were fully insensitive to the difference between fake-good and fake-bad scenarios. This result indicated that the fit indices are totally unable to account for any kind of polarity in the fake perturbation of the data. In practice, however, this might not be a serious concern as in an empirical context is usually not difficult to guess what kind of faking process may affect participants' responses. An example regarding the effect of faking good in self-report measures of compliance in liver transplant patients has been discussed in this article. In general, if data have been collected for studying some stigmatizing characteristics (e.g., risky sexual behavior or drug addictions), then a fake-good manipulation appears to be the natural faking polarity for this kind of data.

Finally, there are also some important practical implications that characterize SGR at a more general level. SGR is defined by a two-stage sampling procedure based on two distinct and well-separated generative models: the model representing the process that generates the data prior to any fake perturbation and the model representing the faking process to perturb the data. Therefore, the overall generative problem is split into two conceptually independent and possibly simpler components (*divide et impera* approach). This makes SGR different from other statistical models, which, instead, try to model the fake problem directly in the original statistical model by using ad hoc empirical paradigms such as *ad lib faking* or *coached faking* to collect data and simulate fake reports (Zickar, Gibby, & Robie, 2004; Zickar & Robie, 1999). SGR is also different from person-fit analysis (e.g., Meijer & Sijtsma, 2001), which is used to investigate the validity of item-score vectors in the IRT framework. Within these distinctions SGR seems more related in spirit to uncertainty analysis (Morgan, Henrion, & Small, 1990) and careless responding analysis (Woods, 2006), which are characterized by an attempt to directly quantify uncertainty of general statistics computed on the data. Finally, new relevant SGR developments may indeed lie in applying it to diverse problems beyond those considered here (i.e., for different types of data and/or with different probabilities of faking for

statistical units and different conditional distributions of faking not necessarily based on uniform kernels).

In sum, the SGR approach has several advantages for analyzing possible fake data. SGR offers a conceptually new approach, which is general, flexible, and works well in practice. Overall, the essential characteristic of SGR is its explicit use of mathematical models and appropriate probability distributions for quantifying uncertainty in inferences based on possible fake data. Moreover, SGR involves the derivation of new statistical results as well as the evaluation of the implications of such new results: Are the substantive conclusions reasonable? How sensitive are the results to the modeling assumptions about the process of faking? Because of its simplicity, general applicability, and originality, we strongly believe SGR will prove to be of great value in all those psychological fields that face the problem of studying sensitive topics.

ACKNOWLEDGMENTS

We thank two anonymous reviewers and the action editor for helpful comments on an earlier draft of the article. We are also grateful to Luigi Burigana for his critical comments about the formal framework of SGR.

REFERENCES

Anderson, C. D., Warner, J. L., & Spector, C. E. (1984). Inflation bias in self-assessment examination: Implications for valid employee selection. *Journal of Applied Psychology, 69,* 574–580.

Beaber, R. J., Marston, A., Michelli, J., & Mills, M. J. (1985). A brief test for measuring malingering in schizophrenic individuals. *American Journal of Psychiatry, 142,* 1478–1481.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

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

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation approach*. Wiley Series on Probability and Mathematical Statistics. Hoboken, NJ: Wiley.

Breivik, E., & Olsson, U. H. (2001). Adding variables to improve fit: The effect of model size on fit assessment in LISREL. In R. Cudeck, S. Du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future. A Festchrift in honour of Karl Jöreskog* (pp. 169–194). Chicago, IL: Scientific Software International.

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

Crawford, V. P. (2003). Lying for strategic advantage: Rational and boundedly rational misrepresentation of intentions. *The American Economic Review, 93,* 133–149.

Cuadrado A., Fabrega, E., Casafont, F., & Pons-Romero, F. (2005). Alcohol recidivism impairs long-term patient survival after orthotopic viver transplantation for alcoholic liver disease. *Liver Transplantation, 11,* 420–426.

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

Dobson, A. J. (2002). *An introduction to generalized linear models* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC Press.

Enders, C., & Finley, S. (2003, April). *SEM fit index criteria re-examined: An investigation of ML and robust fit indices in complex models*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Fan, X., Felsovalyi, A., Sivo, S. A., & Keenan, S. (2002). *SAS for Monte Carlo studies: A guide for quantitative researchers*. Cary, NC: SAS Institute.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling, 12,* 343–367.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research, 42,* 509–529.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling, 6,* 56–83.

Fan, X., & Wang, L. (1998). Effects of potential confounding factors on fit indices and parameter estimates for true and misspecified SEM models. *Educational and Psychological Measurement, 58,* 699–733.

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

Foster, P. F., Fabrega, E., Karademir, S., Sankary, N. H., Mital, D., & Williams, J. W. (1997). Prediction of abstinence from ethanol in alcoholic recipients following liver transplantation. *Hepatology, 25,* 1469–1477.

Furedy, J. J., & Ben-Shakhar, G. (1991). The roles of deception, intention to deceive, and motivation to avoid detection in the psychophysiological detection of guilty knowledge. *Psychophysiology, 28,* 163–171.

Gerbing, D. W., & Anderson, J.C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation modeling* (pp. 40–65). Newbury Park, CA: Sage.

Gray, N. S., MacCulloch, M. J., Smith, J., Morris, M., & Snowden, R. J. (2003). Forensic psychology: Violence viewed by psychopathic murderers. *Nature, 423,* 497–498.

Hall, R. C., & Hall, R. C. (2007). Detection of malingered PTSD: An overview of clinical, psychometric, and physiological assessment. Where do we stand? *Journal of Forensic Science, 52,* 717–725.

Hopwood, C. J., Talbert, C. A., Morey, L. C., & Rogers, R. (2008). Testing the incremental utility of the negative impression-positive impression differential in detecting simulated personality assessment inventory profiles. *Journal of Clinical Psychology, 64,* 338–343.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to under-parameterized model misspecification. *Psychological Methods, 3,* 424–453.

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

Jöreskog, K., & Sörbom, D. (1984). *LISREL VI user's guide* (3rd ed.). Mooresville, IN: Scientific Software International.

Jöreskog, K., & Sörbom, D. (1996a). *LISREL 8: User's reference guide.* Lincolnwood, IL: Scientific Software International.

Jöreskog, K., & Sörbom, D. (1996b). *PRELIS 2: User's reference guide.* Lincolnwood, IL: Scientific Software International.

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika, 27,* 179–182.

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling, 10,* 333–351.

Lombardi, L., Pastore, M., & Nucci, M. (2004). Evaluating uncertainty of model acceptance in empirical applications: A replacement approach. In K. van Montfort, H. Oud, & A. Satorra (Eds.), *Recent developments on structural equation models: Theory and applications* (pp. 69–82). Dordecht, The Netherlands: Kluwer Academic.

Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology, 44,* 258–262.

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

Marshall, E. (2000). How prevalent is fraud? That's a million-dollar question. *Science, 290,* 1662–1663.

Matinlauri, I. H., Nurminen, M. M., Hockerstedt, K. A., & Isoniemi, H. M. (2005). Risk factors predicting survival of liver transplantation. *Transplant Proceedings, 5,* 1155–1160.

McCullagh, P., & Nelder J. A. (1989). *Generalized linear models*. London, UK: Chapman & Hall.

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85,* 812–821.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25,* 107–135.

Morgan, M. G., Henrion, M., & Small, M. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. New York, NY: Cambridge University Press.

Mossman, D., & Hart, K. J. (1996). Presenting evidence of malingering to courts: Insights from decision theory. *Behavioral Sciences and the Law, 14,* 271–291.

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

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

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46,* 598–609.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8,* 287–312.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. R foundation for statistical computing, Vienna, Austria. Retrieved from http://www.R-project.org

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement No. 17*.

Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to accurately detect autobiographical events. *Psychological Science, 19,* 772–780.

Sheather, S. J., & Jones, M. C. (1991). A realible data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, (B), 53,* 683–690.

Sobel, J. (1985). A theory of credibility. *The review of economic studies, 52,* 557–573.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Sun, J. (2005). Assessing goodness of fit in confirmatory factor analysis. *Measurement and Evaluation in Counseling and Development, 37,* 240–256.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.

Van der Geest, S., & Sarkodie, S. (1998). The fake patient: A research experiment in a Ghanaian hospital. *Social Science & Medicine, 47,* 1373–1381.

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28,* 189–194.

Wood, S. N. (2006). *Generalized additive models*. Boca Raton, FL: Taylor & Francis.

Wu, W., & West, S. G. (2010). Sensitivity of fit indices to misspecification in growth curve models. *Multivariate Behavioral Research, 45,* 420–452.

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

Yu, C. Y., & Muthén, B. (2002, April). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Yuan, K. H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40,* 115–148.

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods, 7,* 168–190.

Zickar, M. J., & Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology, 84,* 551–563.