The XIII International Conference
"Applied Stochastic Models and Data Analysis" (ASMDA-2009)
June 30-July 3, 2009, Vilnius, LITHUANIA

ISBN 978-9955-28-463-5
L. Sakalauskas, C. Skiadas and E. K. Zavadskas (Eds.): ASMDA-2009
Selected papers. Vilnius, 2009, pp. 27–32


# A PROBABILISTIC APPROACH FOR EVALUATING THE SENSITIVITY TO FAKE DATA IN STRUCTURAL EQUATION MODELING

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

[1]*Department of Cognitive Sciences and Education University of Trento, Corso Bettini n. 31, I-38068 Rovereto (TN), Italy*
[2]*Department of Developmental and Social Psychology University of Padova, Via Venezia, 8 I- 35131 Padova, Italy*
*E-mail:* [1]*luigi.lombardi@ unitn.it;* [2]*massimiliano.pastore@unipd.it*

**Abstract:** In this paper we address the issue of evaluating the sensitivity of goodness-of-fit indices in structural equation modeling when fake data are considered in three different factorial models with varying sample sizes (n= 50, 100 and 200). The sensitivity evaluation is carried out by means of a simulation procedure which combines a standard Monte Carlo approach and a new probabilistic version of a recent data perturbation procedure called Sample Generation by Replacements (SGR, Lombardi, Pastore and Nucci, 2004). Probabilistic SGR (PSGR) will be used to generate data perturbations based on three different models of faking: fake-uniform, fake-good (deception) and fake-bad (malingering). For each scenario of faking the performance of four very popular goodness-of-fit indices (two absolute indices: GFI, and AGFI; and two incremental indices: CFI and NNFI) will be evaluated.

**Keywords:** SEM, Fake Data Analysis, Goodness-of-Fit indices

## 1. Introduction

In this contribution we propose a simple procedure to evaluate the impact of fake data in structural equation modeling. The paper is organized as follows. Section 2 outlines a probabilistic framework, called PSGR (Probabilistic Sample Generation by Replacements), to generate new data from an observed data matrix. Section 3 describes the simulation study for evaluating the sensitivity of four goodness-of-fit (GOF) indices to three relevant perturbed data scenarios. Finally, in Section 4 we discuss main results and report some concluding remarks.

## 2. Probabilistic SGR

The full data set is represented by an $n \times m$ matrix $\mathbf{D}$, that is to say, $n$ observations each containing $m$ elements (subject's responses). We assume that entry $d_{ij}$ of $\mathbf{D}$ ($\forall i = 1,\ldots, n; \forall j = 1,\ldots, m$) takes values on a small set $Q=\{1,2,\ldots,q\}$. Finally, let $\mathbf{d}_j$ be the $j^{\text{th}}$ column of $\mathbf{D}$. The main idea of our replacement approach is to construct a new data vector $\mathbf{f}_j$, called the *fake data vector*, from the original data vector $\mathbf{d}_j$ by manipulating each entry in $\mathbf{d}_j$ according to a $q \times q$ replacement matrix $\mathbf{R}$. Finally, $\mathbf{F}=[\mathbf{f}_1, \mathbf{f}_2,\ldots, \mathbf{f}_m]$ is the fake data matrix. In particular, entry $r_{hk}$ of $\mathbf{R}$ ($\forall h, \forall k = 1,\ldots, q$) denotes the conditional probability $p(f_{ij}=k|d_{ij}=h)$ of replacing the original observed value $h$ with the new value $k$:

$$\mathbf{R} = \begin{pmatrix} r_{11} & \cdots & r_{1q} \\ \vdots & \ddots & \vdots \\ r_{q1} & \cdots & r_{qq} \end{pmatrix}$$

Three different replacement matrices, $\mathbf{R}_1$, $\mathbf{R}_2$, and $\mathbf{R}_3$, were considered in our study. Each matrix $\mathbf{R}_s$ corresponds to a different scenario of faking. $\mathbf{R}_1$ is called the uniform replacement matrix and represents a context in which responses are subject to random faking. In $\mathbf{R}_1$, each entry in diag($\mathbf{R}_1$) is set equal to $\alpha$ which, in turn, denotes the probability of non-replacement, whereas the off-diagonal entries are set to

$$r_{hk} = \frac{1-\alpha}{q-1}, \qquad \forall h \neq k$$

$\mathbf{R}_2$ is called the high replacement matrix and represents a fake-good scenario in which $f_{ij} \geq d_{ij}$ ($\forall i$ = 1,…, $n$;$\forall j$ = 1,…, $m$). In particular,

$$r_{hk} = \frac{1-\alpha}{q-h}, \qquad \forall h < k$$

Finally $\mathbf{R}_3$ is called the low replacement matrix and represents a fake-bad scenario in which $f_{ij} \leq d_{ij}$ ($\forall i$ = 1,…, $n$;$\forall j$ = 1,…, $m$). In particular,

$$r_{hk} = \frac{1-\alpha}{h-1}, \qquad \forall k < h$$

The structures of $\mathbf{R}_2$ and $\mathbf{R}_3$ are reported below

$$\mathbf{R}_2 = \begin{pmatrix} \alpha & \cdots & r_{1q} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha \end{pmatrix} \qquad \mathbf{R}_3 = \begin{pmatrix} \alpha & \cdots & 0 \\ \vdots & \ddots & \vdots \\ r_{q1} & \cdots & \alpha \end{pmatrix}$$

## 3. Simulation study

In this study, four fit-indices were examined with respect to structured perturbation of data. We considered two absolute GOF indices (Goodness of Fit Index, GFI, and Adjusted Goodness of fit Index, AGFI; Jöreskog & Sörbom, 1994) and two incremental GOF indices (Comparative Fit Index, CFI; Bentler, 1990, and Nonnormed Fit Index, NNFI; Bentler & Bonnett, 1980; Tucker & Lewis, 1973). In this evaluation, three different types of target models were involved. We selected three prototype models commonly encountered in applied research (Paxton, Curran, Bollen, Kirby and Chen, 2001; see Figure 1). The following procedural steps were repeated for each target model $M_s$ ($s$ =1, 2, 3):

1. According to $M_s$, 5000 raw-data sets $\mathbf{D}_t$ with $n$ = 50, 100 and 200 observations were generated. Next, each $\mathbf{D}_t$ ($t$ = 1, . . . , 5000) was discretized on a 5-point scale using the method described by Jöreskog and Sörbom (1996).
2. For each discretized matrix $\underline{\mathbf{D}}_t$ we constructed a collection of fake matrices ${}_z\mathbf{F}_{t,q}$ by using the replacement matrix $\mathbf{R}_z$ ($z$=1,2,3) with non-replacement proportion $\alpha$=1-($q$/100) and $q$ = 10, 20, ..., 100. In particular, for $M_3$ we perturbed only the endogenous variables. The exogenous variables were considered fake independent.
3. Each perturbed data matrix ${}_z\mathbf{F}_{t,q}$ was subjected to model $M_s$ and the four GOF indices were finally evaluated. The whole procedure generated a total of 150000 new perturbed data matrices for each target model.

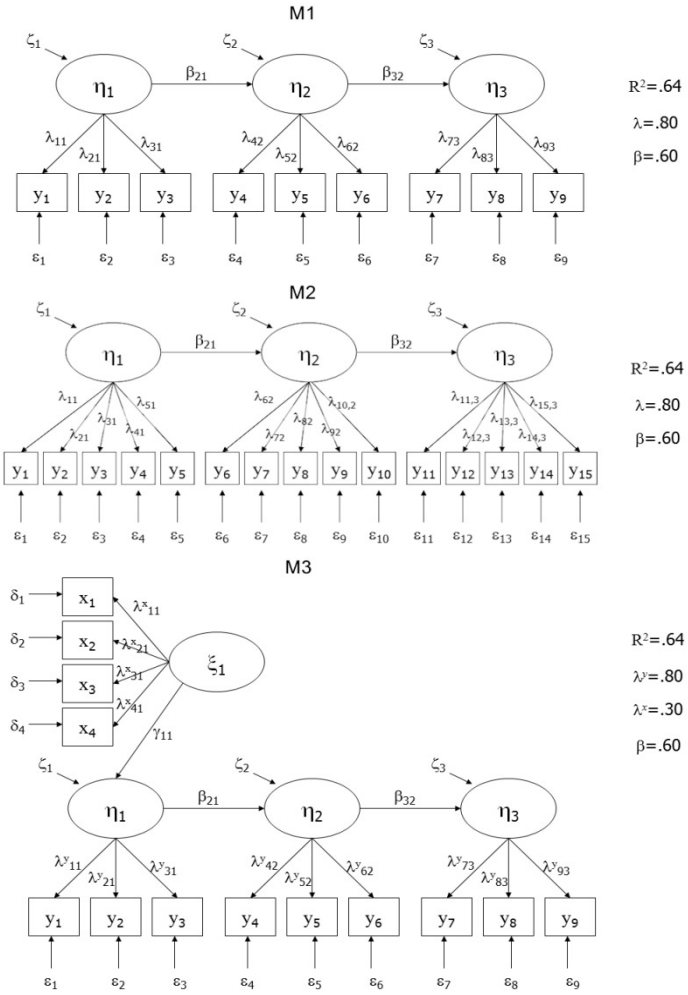


**Fig. 1.** Target models

## 4. Results and conclusions

For sake of simplicity and space, we will limit the description of our results to the $n$ = 200 case only.

Figure 2 shows the medians of GFI and AGFI fits for the three target models and the three scenarios of faking (Fake Uniform [FU], Fake-Good [FH], and Fake-Bad [FL]). Segments are the 95% interquantile intervals, whereas dashed lines represent the cutoff optimal value (.95). The GFI (resp. AGFI) median appeared not to be affected by increasing levels of replacements. In particular, in the FU scenario the medians of GFI and AGFI increased with larger percentage of replaced elements. The latter was a very unexpected result, indeed. Note that a good index should approach its maximum under correct model specification, but also degrade under massive data perturbation.

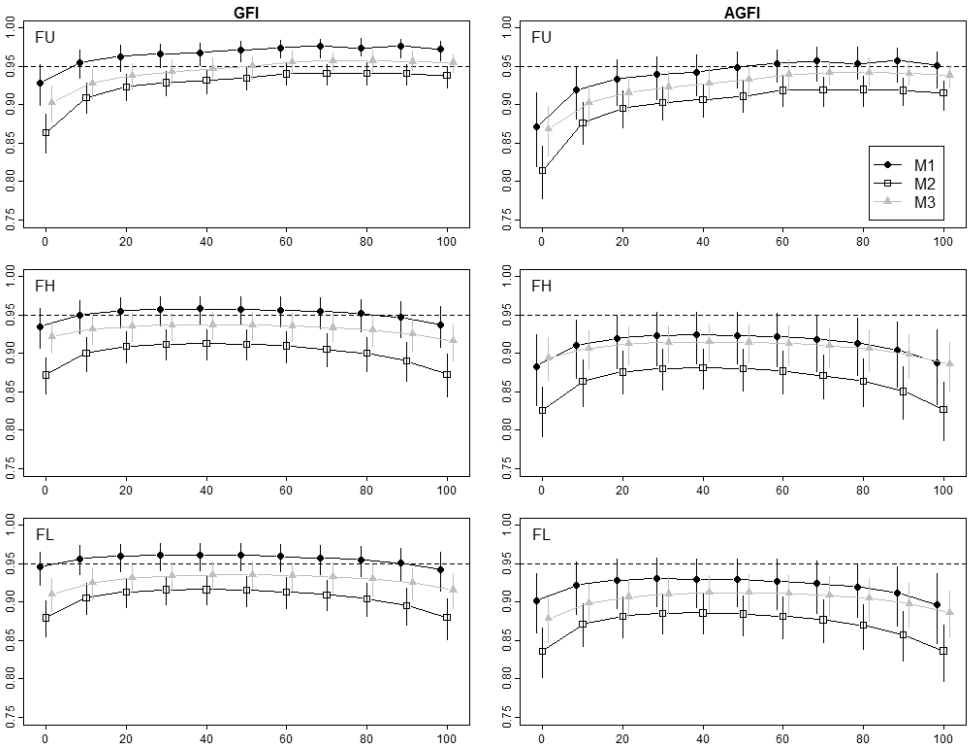


**Fig. 2.** Medians of GFI and AGFI as a function of percentage of replacements and models of faking (FU, FH, and FL). Segments represent 95% interquantile intervals

Figure 3 shows the results for CFI. By increasing the percentage of replacements, CFI medians decreased and, in general, variability increased. The FU scenario revealed more variability as compared to FH and FL scenarios.

Figure 4 shows the results for NNFI. The NNFI pattern is very similar to that of CFI and, in general, by increasing the percentage of replacements, the NNFI medians decreased and variability increased.

Overall our results indicate that the two incremental fit-indices, CFI and NNFI, were more sensitive to fake data. In particular, the effect of perturbed data was very smooth and regular for the fake-good (FH) and fake-bad (FL) scenarios, whereas resulted more extreme and irregular for the fake uniform (FU) scenario. Since the absolute fit-indices, GFI and AGFI, seemed to be unaffected by fake perturbation, we strongly recommend to choose CFI or NNFI to evaluate the goodness-of-fit of a factorial model. This is particularly relevant whenever we suspect that the subjects' responses may have been corrupted by faking (for example, in personnel selection some job applicants may misrepresent themselves on a personality test hoping to increase the likelihood of obtaining a job offer.)

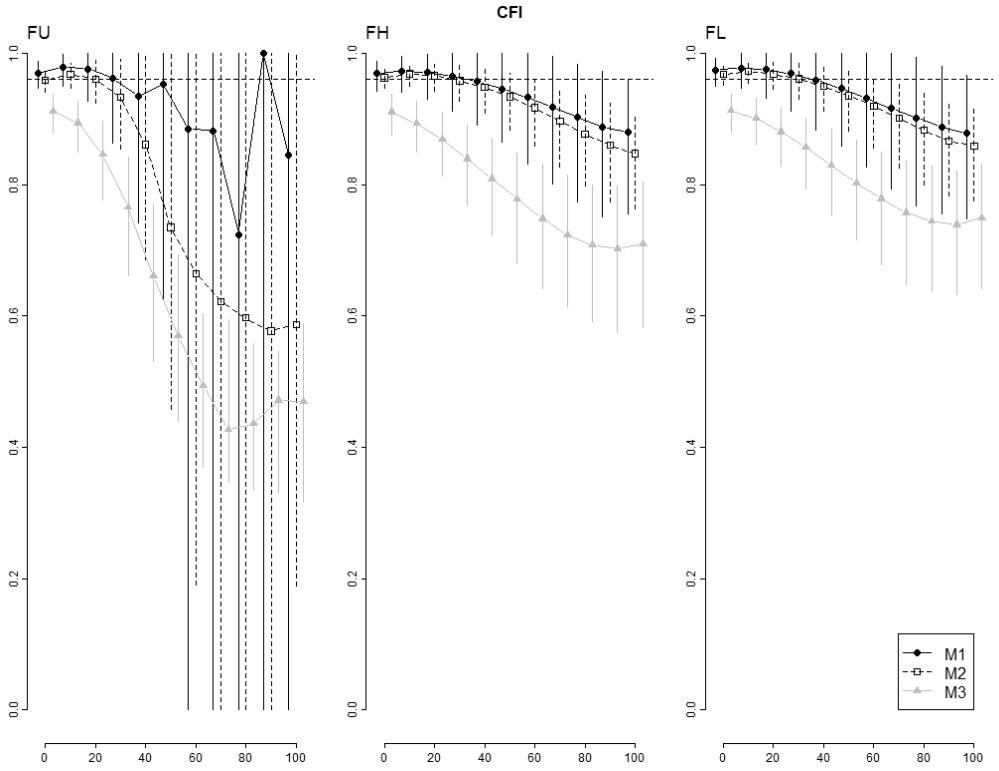


**Fig. 3.** Medians of CFI as a function of percentage of replacements and models of faking (FU, FH, and FL). Segments represent 95% interquantile intervals

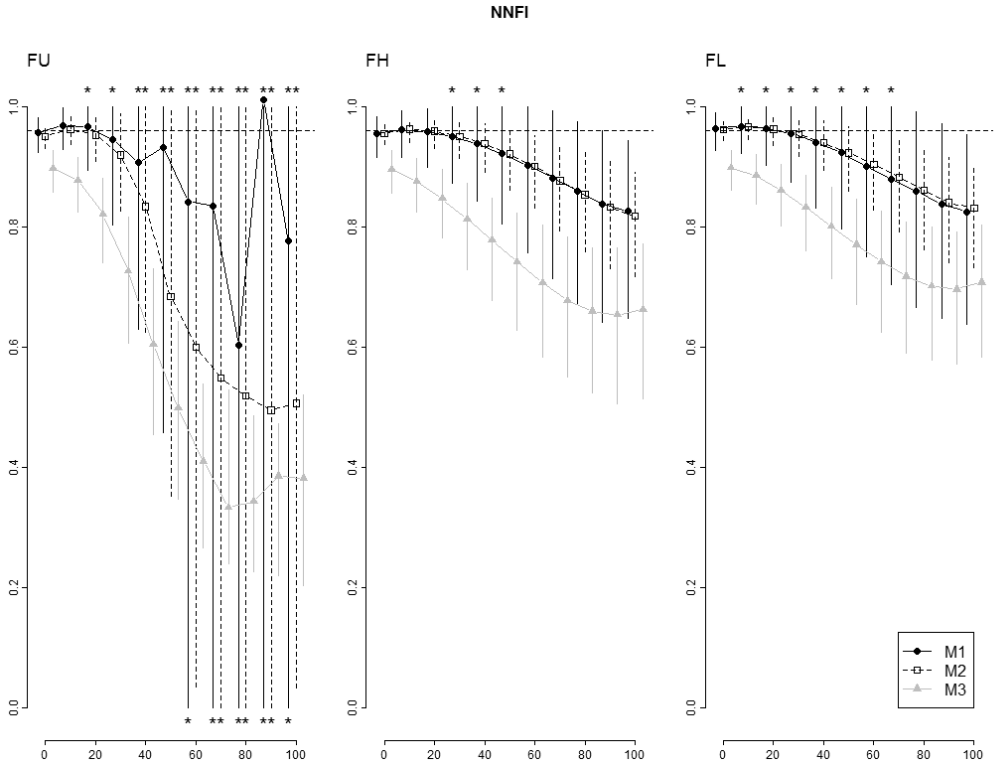


**Fig. 4.** Medians of NNFI as a function of percentage of replacements and models of faking (FU, FH, and FL). Segments represent 95% interquantile intervals. (*) indicates that a percentile falls outside the range [0,1]

## References

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

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

Jöreskog, K. G., and Sörbom, D. 1994. *LISREL VI user's guide (3rd ed.)*. Scientific Software, Mooresville, IN, 1994.

Jöreskog, K. G., and Sörbom, D. 1996. *PRELIS 2: User's reference guide*. Scientific Software, Chicago, IL.

Lombardi, L.; Pastore, M., and Nucci, M. 2004. Evaluating uncertainty of model acceptability in empirical applications: a replacement approach, in Monfort K., Oude H., Satorra A. (Ed.). *Recent developments in structural equation modeling: theory and applications*. Amsterdam, Kluwer.

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

Tucker, L. R., and Lewis, C. 1973. A Reliability Coefficient for Maximum Likelihood Factor Analysis. *Psychometrika* 38: 1–10.