Qual Quant (2014) 48:1191–1211
DOI 10.1007/s11135-013-9829-1

The impact of faking on Cronbach's alpha for dichotomous and ordered rating scores

Massimiliano Pastore · Luigi Lombardi

Published online: 22 February 2013

Abstract In many psychological inventories (i.e., personnel selection surveys and diagnostic tests) the collected samples often include fraudulent records. This confronts the researcher with the crucial problem of biases yielded by the usage of standard statistical models. In this paper we applied a recent probabilistic perturbation procedure, called sample generation by replacement (SGR)—(Lombardi and Pastore, Multivar. Behav. Res 47:519–546, 2012), to study the sensitivity of Cronbach's alpha index to fake perturbations in dichotomous and ordered data, respectively. We used SGR to perform two distinct SGR simulation studies involving two sample size conditions, three item set sizes, and twenty levels of faking perturbations. Moreover, in the second SGR simulation study we also evaluated an additional factor, type of faking model, to study sample reliability under different modulations of graded faking (uniform faking, average faking, slight faking, and extreme faking). To simulate these more complex faking models we proposed a novel extension of the SGR perturbation procedure based on a discrete version of the generalized beta density distribution. We also applied the new procedure to real behavioral data on emotional instability.

Keywords Sample generation by replacement · Cronbach's alpha · Faking good · Monte Carlo

1 Introduction

In many self-report questionnaires the need to analyze empirical data raises the major problem of possible fake observations in the data. Fake data come up from different causes. For

M. Pastore (✉)
Department of Developmental and Social Psychology, University of Padova, Via Venezia, 8, 35131Padua, Italy
e-mail: massimiliano.pastore@unipd.it

L. Lombardi
Department of Psychology and Cognitive Science, University of Trento, Corso Bettini, 31, 38068 Rovereto, TN, Italy
e-mail: luigi.lombardi@unitn.it

example, an individual may deliberately attempt to manipulate or distort responses to personality inventories and attitude tests to create positive impressions (e.g., Furnham 1986; McFarland and Ryan 2000; Paulhus 1991; Zickar and Robie 1999) and match the so called ideal candidate profile (Paulhus 1991). In other circumstances, individuals may tend to malinger responses on a symptom checklist to simulate grossly exaggerated physical or psychological symptoms in order to reach specific goals such as, for example, obtaining financial compensation, avoiding being charged with a crime, avoiding military duty, or obtaining drugs (e.g., Hall and Hall 2007; Mittenberg et al. 2002).

Many self-report measures of attitudes, beliefs, personality, and pathology include items that can be easily manipulated as respondents may know what the correct, socially desirable, or convenient answer is even if that answer does not match their own true, but hidden intentions. This problem is common in areas like psychology (Hopwood et al. 2008), organizational and social science (Van der Geest and Sarkodie 1988), forensic medicine (Gray et al. 2003), and scientific frauds (Marshall 2000).

In general, self-report measures must confront the researcher with a crucial question (see Lombardi and Pastore 2012): *If data included fake data observations, would the answer to the research question be different from what it actually is?* We call this general question the fake effect question (FEQ). A case of particular empirical interest in the analysis of behavioral data is the situation in which a researcher needs to estimate the reliability of composite test scores in possible fake data samples. In this context the FEQ can be rewritten as: *If data contained fake self-report measures, what would the true reliability test score be?* And, in particular, *how can we reconstruct the true reliability score from the observed reliability score computed on the fake-corrupted data?* (reconstruction problem question, RPQ). Of course, FEQ and RPQ are strictly related queries. In this paper we propose a rational/statistical approach in an attempt to answer these important questions.

The reliability measure of a multi-item composite score is one of the most relevant concepts in classical testing theory (Allen and Yen 1979; Nunnally and Bernstein 1994). It is defined as the product–moment correlation between the composite score and the scores on a test parallel to it (Novick 1966; Novick and Lewis 1967; Lord and Novick 1968). The reliability of a test score is higher if the parallel test forms correlate higher. However, since in practical situations the administration of two parallel versions of a test may be unattainable, many alternatives have been proposed (e.g., Guttman 1945; Cronbach 1951; Revelle 1979; Bentler and Woodward 1980; McDonald 1985) that use the data available from a single test administration to compute reliability. In particular, when the items in the composite are assumed to be essentially τ-equivalent (Lord and Novick 1968), many of these coefficients reduce all to the same psychometric construct and the corresponding sample versions are known to be equivalent and consistent estimates of reliability (e.g., Sijtsma 2009; Yuan and Bentler 2002; Zinbarg et al. 2005). Therefore, the essentially τ-equivalent condition represents an ideal framework to study and evaluate the pure impact of fake responses on sample reliability. In general, we would expect that the sample reliability should approach its maximum with high correlated items and uncorrupted data, but also degrade substantially under massive fake data perturbation.

There is now a vast literature about reliability that covers many different aspects, theoretical and applicative ones. For example, MC simulations and theoretical studies have been conducted to evaluate the robustness of reliability measures under violation of symmetry and presence of outliers (e.g., Yuan and Bentler 2002; Maydeu-Olivares et al. 2007; Li Gwet 2008; Liu et al. 2010), as well as the impact of sample size, number of response categories, and number of items (e.g., Lissitz and Green 1975; Cortina 1993; Weng 2004; Duhachek et al. 2005). Moreover, the issue of faking data in multivariate data analysis has been investigated

using ad hoc empirical paradigms such as *ad lib* faking or coached faking to collect data and simulate fake reports (Zickar et al. 2004; Zickar and Robie 1999). In particular, in recent years some authors have proposed rational methods for assessing fake data in social desirability contexts or faking-motivating situations by using factor analytic approaches (Ferrando 2005; Ferrando and Anguiano-Carrasco 2009, 2011) as well as factor mixture models (Leite and Cooper 2010).

Here we take a different approach, called *sample generation by replacement* (SGR; Lombardi and Pastore 2012), to analyze possible fake data. 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 the approaches described earlier, which, instead, try to model the fake problem directly in the original statistical model by using sophisticated or complex sampling procedures (e.g.,ad lib faking and coached faking) to collect data and simulate fake reports. In general, SGR is more related in spirit to uncertainty analysis (Morgan et al. 1990) and careless responding analysis (Woods 2006), which are characterized by an attempt to directly quantify hypothetical uncertainty of general statistics computed on the data.

In this contribution, we examined, in two distinct SGR simulation studies, the sensitivity (under the τ-equivalent assumption) of Cronbach's alpha, to fake perturbations in dichotomous and ordered data, respectively. Each of the two SGR simulation studies involved two sample size conditions, three item set sizes, and twenty levels of faking perturbations. Moreover, in the second SGR simulation study we also added an additional factor, type of faking model, to evaluate the sensitivity of the coefficient alpha under different modulations of graded faking. In order to simulate these more complex faking models we introduced a novel extension of the SGR resampling distribution based on a discrete version of the generalized beta density distribution.

The remainder of the paper is organized as follows: The first section starts with a brief recapitulation of the main components of the SGR approach as originally introduced by Lombardi and Pastore (2012). Next, it continues by illustrating the problem of representing reliability in SGR. Finally, it ends by introducing the new models of faking. The second section presents the two simulation studies and reports results about the evaluation of the sample coefficient alpha under fake data perturbations. The third section illustrates our method with an application to real data about emotional instability. In this application SGR is used to reconstruct approximate 95 % confidence intervals (CIs) for true reliability coefficient values (see RPQ). Finally, the fourth and last section presents conclusions and some relevant comments about limitations, potential new applications and extensions of the SGR approach.

2 The SGR approach

2.1 The fake-data representation problem in general

We think of two distinct data constructs: the original data set and the fake data set. The original data set contains hidden sensitive responses and, therefore, is not a directly observable object. By contrast, the fake data set contains observable self-report responses. In particular, we assume that an observable self-report response corresponds to an original response after being corrupted by some faking process. We now provide a more formal description of these important data constructs.

The *original data set* is represented by an $I \times J$ array $\mathbf{D}$, with row $\mathbf{d}_i$ $(i = 1, \ldots, I)$ containing the hidden response profile of the ith individual to the J items. We assume that entry d_{ij} of $\mathbf{D}$ $(i = 1, \ldots, I; j = 1, \ldots, J)$ takes values on a small ordinal range $\mathcal{V}_q^Q = \{q, q+1, \ldots, q+t = Q\}$ with $q \in \mathbb{N} \cup \{0\}$ and $t \in \mathbb{N}$. The hidden response pattern $\mathbf{d}_i$ is a multidimensional ordinal random variable with probability distribution $p(\mathbf{d}_i|\theta_M)$, where θ_M indicates the vector of parameters of the probabilistic model of the original data. Moreover, we assume that the original response patterns are independent and identically distributed (i.i.d.) observations and consequently

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

The model defined in Eq. (1) is called the *true original data model* and represents the true data generation process. In the multivariate latent variable framework a common approach for modeling ordinal variables according to Eq. (1) is the underlying variable approach (Muthén 1984; Jöreskog and Sörbom 1996a). This approach assumes that the observed ordinal variables are treated as metric through assumed underlying normal variables. For example, under the simplified essentially τ-equivalent condition, the vector of parameters θ_M represents the true population parameters of a single factorial model with equal loadings for the J items.

The *fake data set* is represented by an $I \times J$ array $\mathbf{F}$ containing the observed subjects' responses. Likewise for the entries of $\mathbf{D}$, also the entries of $\mathbf{F}$ take values on the same discrete set $\mathcal{V}_q^Q$. We assume that the fake array $\mathbf{F}$ can be constructed by manipulating each element d_{ij} in $\mathbf{D}$ according to a *replacement probability distribution*. In particular, let f_{ij} be the element of $\mathbf{F}$ denoting the observed response of subject i to item j. The observed response f_{ij} is a discrete random value from the conditional replacement probability distribution $p(f_{ij}|d_{ij}, \theta_F)$ where θ_F indicates the parameter vector for the replacement model. In the conditional replacement probability distribution we assume that each observed response f_{ij} only depends on the corresponding original observation d_{ij} and the model parameter θ_F. Therefore, the fake data array is drawn from the joint probability distribution

$$p(\mathbf{F}|\mathbf{D}, \theta_F) = \prod_{i=1}^{I} \prod_{j=1}^{J} p(f_{ij}|d_{ij}, \theta_F) \tag{2}$$

The model defined in Eq. (2) is called the *faking model* of the data and represents the faking generation process.

It is important to note that the faking model integrates together two different kinds of information: (a) the observed data $\mathbf{D}$ representing variables' features and relations generated according to Eq. (1) (b) the model parameter, θ_F, which characterizes some relevant properties of the faking model. In general, θ_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 Eqs. (1–2) we can generate the so called *fake data sample* (FDS). We can then study the distribution of some relevant statistics computed on this FDS.

In the following two sections we introduce the original data model and the faking models that we used for representing the sampling processes in this study.

2.2 Representing the alpha index in SGR

We can easily reformulate the general definitions of the SGR approach to study the effect of fake data perturbation on the reliability of composite scores under the essentially τ-equivalent condition. Note that in this simplified condition many important reliability coefficients such as, for example, the β index (Revelle 1979), the ω index (McDonald 1985), the λ_3 index (Guttman 1945), and the *glb* index (Bentler and Woodward 1980) reduce all to the coefficient alpha (Cronbach 1951)

$$\alpha = \frac{J}{J-1}\left(1 - \mathrm{tr}(\Sigma)/\sum_{jj'}\sigma_{jj'}\right),$$

and the corresponding sample coefficient

$$\widehat{\alpha} = \frac{J}{J-1}\left(1 - \mathrm{tr}(\mathbf{S})/\sum_{jj'}s_{jj'}\right)$$

is known to be a consistent estimate of reliability (e.g., Sijtsma 2009; Yuan and Bentler 2002; Zinbarg et al. 2005). In the above equations $\mathbf{S}$ and Σ denote the sample covariance matrix and its population counterpart, respectively. In this paper we will study the performance of the sample coefficient alpha under different fake perturbation scenarios.

The τ-equivalent condition reveals also a natural way to represent the *true original data model* in the SGR framework. We assume that the original data $\mathbf{D}$ has been generated by a single factor with unit variance and J equal factor loadings. More precisely, the underlying normal data set representing $\mathbf{D}$ is a random sample from the statistical population determined by the true population parameters θ_M of the single factorial model and $\mathbf{D}$ represents its discretization. In particular, since all the J item loadings are equal, $\lambda_1 = \lambda_2 = \ldots = \lambda_J = \lambda$, then in the true original data model the population covariance of any two items, j and j', boils down to $\sigma_{jj'} = \lambda^2$.

Finally, it is clear that many empirical contexts may require different and more complex true original data model assumptions (e.g., multidimensional factor, unequal general factor loadings). However, in this first study we wanted to understand the impact of fake data under the simplest and purest representation of sample reliability first.

2.3 Representing models of faking in SGR

The previous section described a simple and elegant way to characterize the true original data model in the SGR framework. In this section, we present a novel family of faking models that extends the uniform support fake–good distribution originally introduced in Lombardi and Pastore (2012). The new proposal shows two relevant features. First, it can represent faking models for dichotomous data as well as ordinal rating data. Second, unlike the uniform support fake–good distribution, it can model both symmetric and asymmetric faking good processes. This latter property is particularly important for mimicking different modulations of graded faking such as, for example, uninformative/neutral faking, slight faking, and extreme faking.

2.3.1 The uniform support fake–good distribution

Lombardi and Pastore (2012) used a simple parametrized distribution to model the faking good scenario (Paulhus 1984; McFarland and Ryan 2000). In general, faking good can be

conceptualized as an individual's deliberate attempt to manipulate or distort responses to create a positive impression (Paulhus 1984; Zickar and Robie 1999; McFarland and Ryan 2000). Notice that, the faking good (as well as the faking bad) scenario always entails a conditional replacement model in which the conditioning is a function of response polarity. This model represents a perturbation 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, the replacement probability distribution is defined as follows

$$p(f_{ij} = k | d_{ij} = h, \theta_F) = \begin{cases} 1, & h = k = Q \\ \frac{\pi}{Q-h}, & q \leq h < k \leq Q \\ 1 - \pi, & q \leq h = k < Q \\ 0, & q \leq k < h \leq Q \end{cases} \tag{3}$$

The uniform support fake–good distribution has only one parameter, π, denoting the overall probability of replacement in the model. Eq. (3) defines the conditional probability of replacing an original observed value h in entry (i, j) of $\mathbf{D}$ with the new value k. Note that this model does not allow to substitute the original observed value with lower ones. Moreover, this simple replacement model is characterized by a uniform probabilistic kernel. More precisely, in the faking good model all the values $k > h$ are assumed to be equally likely in the process of replacement. In sum, the model represents a *purely random*, but *polarized* malingering *process*: PRPP (see Lombardi and Pastore 2012, for additional details about the PRPP assumption).

However, some empirical contexts may require different model assumptions as well as different fake distribution conditions that cannot be captured by this simple faking model. In particular, 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. In sum we can recognize at least two different forms of asymmetric faking: slight faking and extreme faking. Thus, it seems worthwhile to generalize the SGR framework to investigate also these relevant typologies of faking.

2.3.2 Generalized beta distributions

In this paper we propose a novel characterization of the conditional replacement distribution that extends the uniform support fake–good model to more complex asymmetric faking scenarios. Because our new proposal is based on a discrete version of the generalized beta density, in what follows we present some relevant properties of this important distribution function before introducing the new faking models.

Generalized beta distribution. This distribution extends the classical beta distribution to continuous random variables beyond the normalized range [0, 1] (Whitby 1971). This distribution is defined as follows

$$G(x; a, b, \gamma, \delta) = \frac{1}{B(\gamma, \delta)(b-a)^{\gamma+\delta-1}} (x-a)^{\gamma-1} (b-x)^{\delta-1}, \tag{4}$$

where $B(\gamma, \delta)$ is the beta function. The parameters $a \in \mathbb{R}$ and $b \in \mathbb{R}$ (with $a < b$) are the left and right end points respectively, and $\gamma > 0$ and $\delta > 0$ are the governing shape parameters for a and b respectively. For all the values of the r.v. X that fall outside the interval $[a, b]$, G simply takes value 0. The generalized beta distribution reduces to the beta distribution when

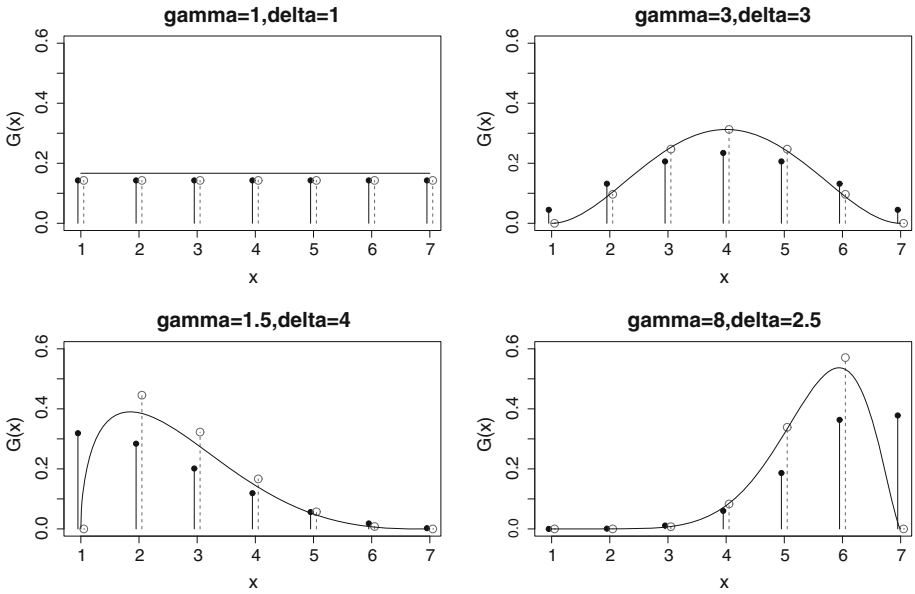

Fig. 1 Four examples for the generalized beta distribution G with a continuous variable with bounds $a = 1$ and $b = 7$ together with the corresponding pmf of its discrete counterpart DG (continuous segments with filled circles) for a 7-point discrete r.v. The graphical representation also shows a discrete version of the generalized beta distribution (*dashed segments with unfilled circles*) which is not based on the correction terms implemented in Eq. (6)

$a = 0$ and $b = 1$. Figure 1 shows four examples of generalized beta distributions. We now describe a useful way to derive a discrete representation for G.

Generalized beta distribution for discrete variables. Let X be a discrete random variable with range

$$R_X = \{a, a+1, a+2, \ldots, a+t-1, a+t = b\}$$

and where $a \in \mathbb{N} \cup \{0\}$ and $t \in \mathbb{N}$. We propose the following *generalized discrete beta distribution* for the r.v. X

$$DG(x; a, b, \gamma, \delta) = \begin{cases} \frac{G^*(x;a,b,\gamma,\delta)}{\sum_{x' \in R_X} G^*(x';a,b,\gamma,\delta)}, & x \in R_X \\ 0, & x \notin R_X \end{cases} \tag{5}$$

where G^* is a modified version of the generalized beta distribution G defined as

$$G^*(x; a, b, \gamma, \delta) = \frac{1}{B(\gamma, \delta)(b-a+2)^{\gamma+\delta-1}} (x-a+1)^{\gamma-1} (b-x+1)^{\delta-1}, \tag{6}$$

Note that the denominator of Eq. (5) acts as a normalizing factor for the mass probability distribution. Moreover, in G^* we also added two positive constants ["2" in $b - a(+2)$; and "1" in $x - a(+1)$ and $b - x(+1)$] acting as correction terms to cut the tails of the generalized beta distribution. Figure 1 also shows a comparison between the pmf of DG and the corresponding pmf obtained by not implementing the correction terms in the generalized discrete beta distribution. This modification of G plays a key role in modeling the new conditional replacement distribution.

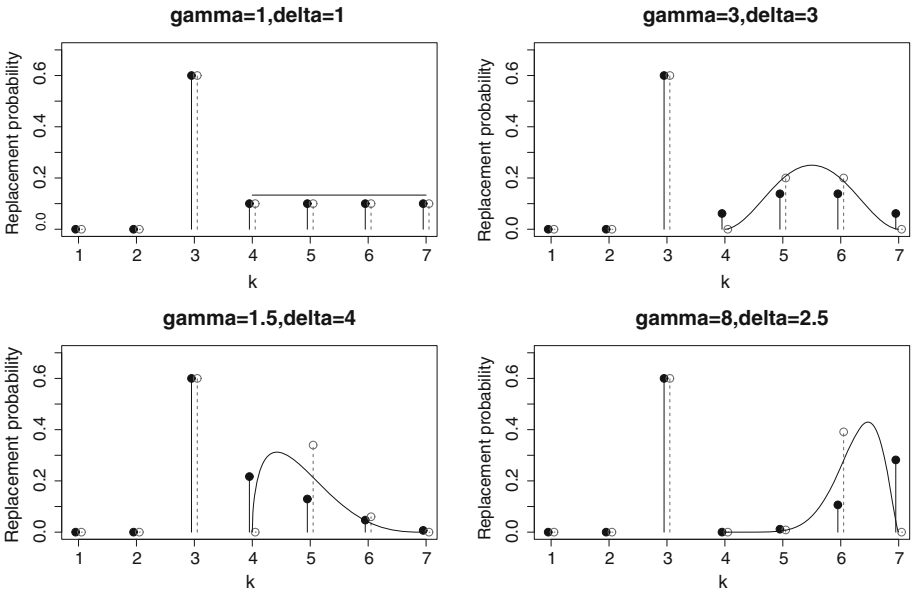

Fig. 2 Four examples of replacement distributions (continuous segments with filled circles) for a 7-point discrete r.v. with original value $h = 3$ and $1 - \pi = 0.6$. The graphical representation also shows the effect of not considering the correction terms in the implementation of the replacement distribution (*dashed segments with unfilled circles*). Note that, without the correction terms, smooth transitions in the moderate (*resp. extreme*) positive shift are not guaranteed in the replacement distribution (*bottom panels*)

2.3.3 The new replacement distribution.

We used the distribution defined in Eq. (5) as a basis for modeling faking good scenarios in ordinal rating items with values in $\mathcal{V}_q^Q$. More precisely, the new conditional replacement distribution can be described according to the following equation

$$p(f_{ij} = k | d_{ij} = h, \theta_F) = \begin{cases} 1, & h = k = Q \\ \pi\, DG(k; h+1, Q, \gamma, \delta), & q \le h < k \le Q \\ 1 - \pi, & q \le k = h < Q \\ 0, & q \le k < h \le Q \end{cases} \tag{7}$$

with θ_F and DG being the parameter vector $\theta_F = (\gamma, \delta, \pi)$ and the generalized beta distribution for discrete variables (Eq. 5) with bounds $a = h + 1$ and $b = Q$, respectively. Finally, the parameter π denotes the overall probability of replacement and acts as a weight to rescale DG. Four examples of replacements distributions are shown in Fig. 2.

This new faking model can easily represent both symmetric (Fig. 2, top panels) and asymmetric replacement kernels (Fig. 2, bottom panels). In particular, if $\gamma = \delta = 1$, the model reduces to the uniform support fake–good distribution [see Eq. (3) and Fig. 2 top-left panel]. By contrast, if $1 \le \gamma < \delta$ (resp. $1 \le \delta < \gamma$), the model mimics asymmetric faking configurations corresponding to moderate positive shifts (resp. exaggerated positive shifts) in the value of the original response (Fig. 2, bottom panels). A relevant case is when $\pi = 0$. For this special condition the fake data matrix $\mathbf{F}$ reduces to the original data matrix $\mathbf{D}$ [see Eq. (7)]. Finally, if $\gamma = \delta = 1$ and $a = 0$, $b = 1$, then Eq. (7) boils down to a simple faking

good model for dichotomous items ($\mathcal{V}_0^1 = \{0, 1\}$). In particular, Eq. (7) reduces to

$$p(f_{ij} = k | d_{ij} = h, \theta_F) = \begin{cases} 1, & h = 1, k = 1 \\ \pi, & h = 0, k = 1 \\ 1 - \pi, & k = h = 0 \\ 0, & h = 1, k = 0 \end{cases} \tag{8}$$

with π being the overall probability of replacing a zero with a one in the self report response.

3 SGR simulation study 1 (dichotomous items)

3.1 Faking model

For the faking good model we used the simple replacement distribution defined in Eq. (8). In this first simulation study we modelled a context in which binary responses were exclusively subject to positive feigning. In sum, the model described the simplest data replacement process and, in the absence of further knowledge about the process of faking, all entries in the original data set were assumed to be equally likely in the process of replacement (PRPP assumption).

3.2 Simulation design and data conditions

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

a. the sample size (I), at two levels: 30 and 100[1];
b. the number of dichotomous items (J) in the composite score, at three levels: 6, 12, and 18;
c. the equal factor loadings (L) of the single factorial model, at 19 levels: .05, .10, . . . , .95 (by a.05 step);
d. the percentage of fake–good replacements (K) in the original data, at 20 levels: 0, 5, . . . , 95 % (by a 5 % step).

Let n_i, n_j, λ, and k be distinct levels of factors I, J, L, and K, respectively. The following procedural steps were repeated 4,000 times for each of the $2 \times 3 \times 19 \times 20 = 2,280$ combinations of levels (n_i, n_j, λ, k) of the simulation design:

1. Generate a raw-data set **D** with size n_i according to the single factorial model with n_j variables and equivalent factor loadings λ. The data generation was performed using a standard MC procedure based on multivariate normal data (Fan et al. 2002; Kaiser and Dickman 1962).
2. Dichotomize **D** using the method described by Jöreskog and Sörbom (1996b).
3. Sample a fake data matrix **F** using the conditional replacement probability distribution with replacement parameter $\theta_F = \pi = \frac{k}{100}$ (see Eq. 8).
4. Compute the sample alpha coefficient on the covariance matrix of **F** as well as on the covariance matrix of the original dichotomized data **D** and save their values for later analyses.

[1] Recommendations of sample size requirements for reliability studies vary widely in the literature. For example Fleiss (1981) suggests that a sample size of 15–20 observations is enough, whereas Nunnally and Bernstein (1994) recommend a sample size of at least 300 observations. In this study, we consider $I = 30$ as an example of very small sample size (lower bound condition) and $I = 100$ as an example of medium/large sample size.

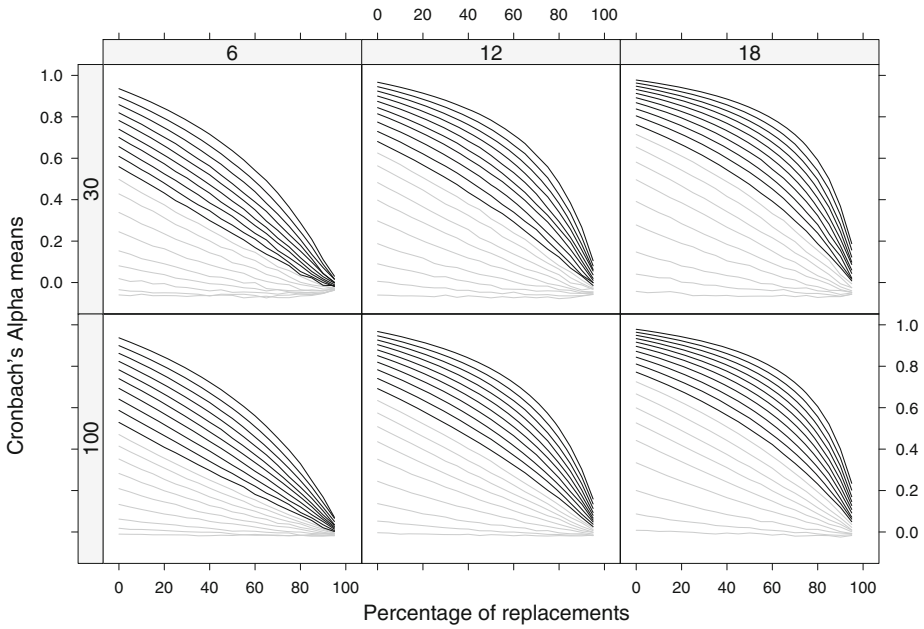

Fig. 3 Means of coefficient alpha as a function of percentage of replacements, factor loading value, number of dichotomous items, and sample size. *Top panel* small sample condition (I = 30). *Bottom panel* larger sample condition (I = 100). Performance curves corresponding to loadings ≥ 0.50 (resp. < 0.50) are represented with *black* (resp. *ligth grey*) *lines*

This algorithm was used to generate 4,000 distinct matrices **F** and associated alpha coefficient estimates for each combination of levels (n_i, n_j, λ, k) of the SGR simulation design. This number of replications was chosen to achieve reasonable estimation stability in the tail regions of the alpha estimates. Note that in step 3, if $k = 0$, then the fake matrix **F** reduces to the data matrix **D** as the probability of replacement π boils down to zero (see Eq. 8). The whole procedure generated a total of $9,120,000 = 4,000 \times 2 \times 3 \times 19 \times 20$ new fake matrices as well as an equivalent number of alpha coefficient estimates. Moreover, by varying the 19 values of λ we were able to generate original data that produced alpha coefficient estimates spanning the entire range of the statistic.

Some additional details are necessary concerning the discretization procedure adopted in step 2. After sampling continuous data from the distribution described in step 1, we transformed these samples into dichotomous (0/1) data by applying a single threshold that remained constant across all data **D**. Since, a dichotomous variable has only one distinct threshold, $-\infty \leq \upsilon_1 < +\infty$, the normal quantile 0 was used as the corresponding threshold value. Finally, the original continuous data d_{ij} was transformed into 1 if $d_{ij} > 0$; otherwise the new value was set to 0.

In the current and in the next simulation study all the data were simulated using a combination of R scripts (R Core Team 2012).

3.3 Results and discussion

Figure 3 (*faking effect chart*) shows the mean of the sample alpha coefficient as a function of factors I, J, and K, respectively. For the sake of convenience, in the faking effect chart we

Table 1 Partial $\eta^2 \times 100$ and ANOVA results for the first simulation study

	Partial $\eta^2 \times 100$	Sum sq	Df	F value	p
I	0	0.46	1	9.79	0.0018
J	9	10.36	2	109.39	0.0000
L	47	95.02	1	2, 007.42	0.0000
I by J	0	0.00	2	0.03	0.9704
I by L	0	0.09	1	1.84	0.1752
J by L	1	0.94	2	9.98	0.0000
I by J by L	0	0.00	2	0.01	0.9921
Residuals		107.36	2, 268		

The three-way ANOVA was computed on the collection of 2,280 sample coefficient means obtained by pooling all the means across the levels of factor *K*

represented all the performance curves generated by lambda values greater than 0.5 with black lines; whereas the performance curves with lambda values not greater than 0.5 were depicted using light grey lines. Note that in this chart the original uncorrupted alpha values always correspond to the non-replacement condition ($k = 0$). We can easily identify a dominance relation in the faking effect chart. In particular, let α_1 and α_2 be two original alpha values such that $\alpha_2 > \alpha_1$, then the performance curve of α_2 always dominates that of α_1. This dominance relation was observed across all the six conditions represented in Fig. 3. In sum, the alpha coefficient resulted sensitive to fake perturbation as the sample alpha mean decreased by increasing levels of replacements, that is to say, it degraded with larger amounts of fake perturbations. To determine whether the factors sample size, number of dichotomous items, and loading value (dichotomized into two levels: <0.5 and ≥ 0.5) affected changes in the average performance of the sample α coefficient, a 2 (I) $\times$ 3 (J) $\times$ 2 (L) analysis of variance (ANOVA) was conducted by pooling the sample α means across all the levels of factor K. The percentage of variance explained by each of the three factors is presented in Table 1. The main effects of factors I, J, and L were all statistically significant (all ps <0.01) as well as the interaction J by L ($p<0.01$). All the other sources were not statistically significant. The percentage of variance explained by factor J and factor L were 9 % and 47 %, respectively. By contrast, factor I as well as the interaction J by L did not account for any substantial variance in the data ($\leq$ 1 %). Finally, all pairwise comparisons between the levels of factor J resulted statistically significant at a Tukey HSD test (all ps <0.01).

We recall that the reliability of a test score is higher if the error variance is smaller relative to the test score variance. Moreover, we would also expect that the sample reliability should approach its maximum with high correlated items and uncorrupted data, but also degrade substantially under massive fake data perturbation. The results of our first SGR simulation study led us to believe that the sample alpha coefficient was clearly sensitive to fake perturbations as, in general, the sample alpha mean decreased by increasing levels of replacements in the data sample. Moreover, the sample alpha coefficient was also sensitive to number of items in the dichotomous scale. This latter result is not difficult to understand and is consistent with those discussed in other researchers about the relationships between alpha performance and number of items in a composite score (e.g., Lord and Novick 1968; Green et al. 1977; Cortina 1993). For example, Cortina (1993) showed that number of items had a profound impact on alpha, especially at low levels of average item intercorrelation and, in general, this relationship was positive and curvilinear (Komorita and Graham 1965).

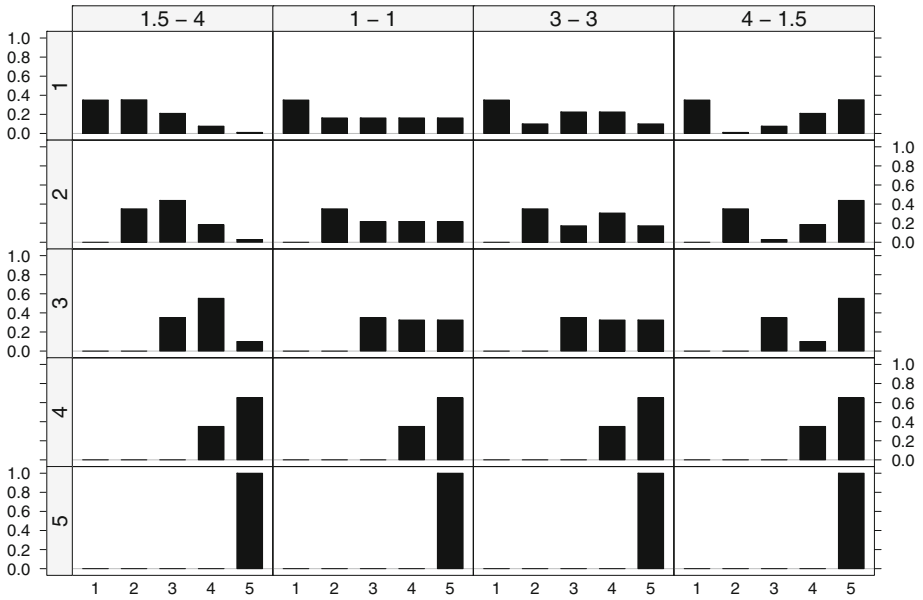

Fig. 4 Four examples of conditional replacement distributions for a 5-point discrete r.v. *Each column* in the graphical representation corresponds to a different conditional replacement distribution with $\pi = 0.4$ and one of the four different assignments for the shape parameters ($\gamma = 1.5, \delta = 4$; $\gamma = \delta = 1$; $\gamma = \delta = 3$; and $\gamma = 4, \delta = 1.5$). *Each row* in the graphical representation corresponds to a different original 5-point discrete value h

The results of our first simulation study extended these findings also in the context of fake corrupted data.

4 SGR simulation study 2 (ordered rating items)

In this second SGR simulation study we evaluated the sensitivity of the sample alpha coefficient to different modulations of graded fake perturbations in 5-point Likert scales.

4.1 Faking models

For the modeling of the faking process we introduced four new modulations of graded fake perturbations: uninformative/neutral faking, average faking, slight faking, and extreme faking. In particular, we used the replacement distribution defined in Eq. (7) to sample fake good observations in the new SGR simulation study.

The *uninformative model* ($\gamma = \delta = 1$) is characterized by the uniform support fake–good distribution [see Eq. (3)] and is based on the PRPP assumption. This principle reflects 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 4 (second column) shows four examples of uniform support fake–good distributions.

The *average model* ($\gamma = \delta = 3$) represents a complication of the uniform support fake–good distribution and is characterized by a symmetric kernel located at around the average distance between the value $h+1$ and the upper bound Q. In this model the chance to replace an

original value h with another greater value k decreases as a function of the distance between k and $(Q + h + 1)/2$. This property reflects that in the process of faking an individual will on average replace the original response with a new fake value which lies at the mean distance between this original response and the most extreme value in the scale. Figure 4 (third column) shows four examples of uniform support fake–good distributions.

The *slight model* ($\gamma = 1.5; \delta = 4$) describes an asymmetric faking good configuration in which the observed self report measure corresponds to a moderate positive shift in the value of the original response. Figure 4 (first column) shows four examples of conditional replacement distributions for slight faking. In this model the chance to replace an original value h with another greater value k decreases as a function of the distance between k and h.

The *extreme model* ($\gamma = 4; \delta = 1.5$) describes an asymmetric faking good configuration in which the observed self report measure corresponds to an exaggerated positive shift in the value of the original response. Figure 4 (fourth column) shows four examples of conditional replacement distributions for extreme faking. Unlike the slight model, in the extreme model the chance to replace an original value h with another greater value k increases as a function of the distance between k and h.

Notice that the uninformative and average models are examples of faking models with symmetric kernels; whereas, the slight and extreme models represent instances of faking models with asymmetric kernels.

4.2 Simulation design and data conditions

The second simulation design included all the factors of the first simulation study plus an additional factor, called type of faking model (M), at four levels: uninformative, average, slight, and extreme. Therefore, five factors were systematically varied in a complete five-factor design.

Let m be a generic level of factor M. The following procedural steps were repeated 4,000 times for each of the 9,576 combinations of levels $(n_i, n_j, \lambda, k, m)$ of the new simulation design:

1. Generate a raw-data set **D** with size n_i according to the single factorial model with n_j variables and equivalent factor loadings λ. The data generation was performed using the same procedure described in the first simulation study.
2. Discretize **D** on a 5-point scale using the method described by Jöreskog and Sörbom (1996b).
3. Sample a fake data matrix **F** using the conditional replacement probability distribution with replacement parameter $\theta_F = (\pi = \frac{k}{100}, \gamma_m, \delta_m)$ and the discretized data **D** (see Eq. 7). The coefficients γ_m and δ_m refer to the shaping parameters of model m.
4. Compute the sample alpha coefficient on the covariance matrix of **F** as well as on the covariance matrix of the original discretized data **D** and save their values for later analyses.

The whole procedure generated a total of $38,304,000 = 4,000 \times 2 \times 3 \times 19 \times 21 \times 4$ new fake matrices as well as an equivalent number of coefficient alpha estimates.

The discretization procedure adopted in Step 2 was a straightforward generalization of the one described in the first simulation study for the dichotomous items. In particular, since a five-category ordinal variable has four distinct thresholds, $-\infty < \upsilon_1 < \upsilon_2 < \upsilon_3 < \upsilon_4 < +\infty$, the normal quantiles, -1.53, -0.49, 0.49, and 1.53, were used as corresponding threshold values. Finally, the original continuous data set **D** was discretized into symmetrically distributed ordinal variables (Step 2) on the basis of these four threshold values.

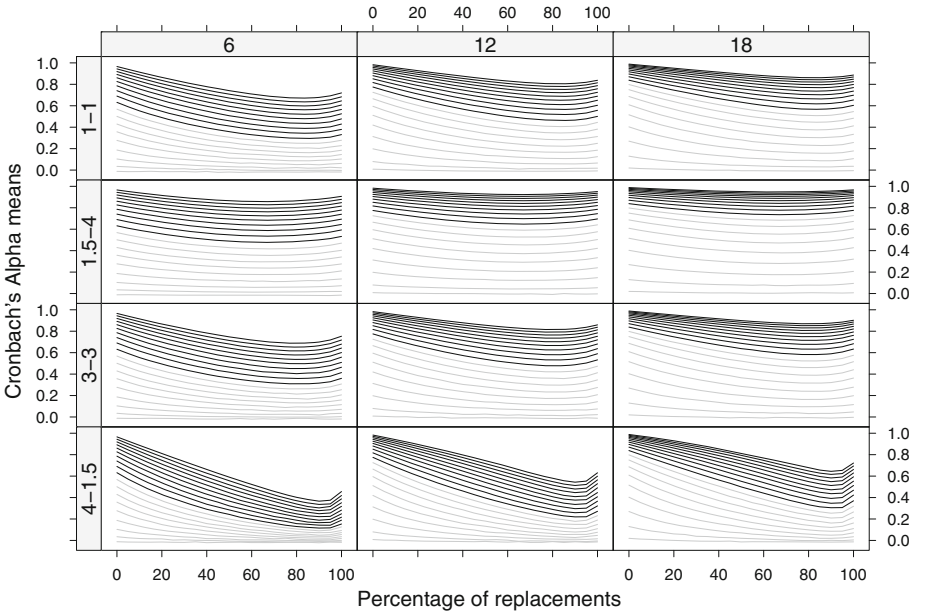

Fig. 5 Means of coefficient alpha as a function of percentage of replacements, factor loading value, number of 5-point rating items, and type of faking model for the larger sample size condition (I = 100). We omitted the graphical representation for the I = 30 sample size condition as it resulted undistinguishable from the other condition I = 100 (see also the main text). Performance curves corresponding to loadings ≥ 0.50 (resp. < 0.50) are represented with *black* (resp. *ligth grey*) *lines*

4.3 Results and discussion

Figure 5 (*faking effect chart*) shows the mean of the sample α coefficient as a function of factors J, K, and M for the I = 100 sample size condition only. Like for the dichotomous case, also in the second simulation study the faking effect chart was characterized by a dominance relation. In general, the sample alpha mean decreased by increasing levels of replacements as it degraded with larger amounts of fake perturbations. To evaluate if the factors sample size, number of items, loading value (dichotomized into two levels: <0.5 and ≥ 0.5) and model of faking influenced changes in the average performance of sample α, a new 2 (I) $\times$ 3 (J) $\times$ 2 (L) $\times$ 4 (M) analysis of variance (ANOVA) was conducted by pooling the sample α means across all the levels of factor K. Table 2 reports the percentage of variance explained by each of the four factors I, J, L, and M, respectively. All the main effects resulted statistically significant (all ps <0.01) as well as the interactions I by L ($p<0.05$), J by L, M by L, and J by M by L (all ps <0.01). All the other interactions were not statistically significant. The percentage of variance explained by J, L, and M were 16 %, 67 % and 14 %, respectively. By contrast, all the other significant sources did not account for any substantial amount of percentage of variance in the data. Like for the dichotomous item case, also in this second simulation study the models with a larger number of items (J) yielded on average better performances (all ps <0.01 at a Tukey HSD test). As expected, the extreme model showed the worse performance (all ps <0.01 at a Tukey HSD test), whereas the slight model was substantially not sensitive to fake perturbation and in general yielded the best performance (all ps <0.01 at a Tukey HSD test). Finally, the uninformative model and the average model did not differ significantly ($p = 0.516$) from one another.

Table 2 Partial $\eta^2 \times 100$ and ANOVA results for the second simulation study

	Partial $\eta^2 \times 100$	Sum sq	Df	F value	p
I	0	1.02	1	34.77	0.0000
J	16	54.48	2	925.91	0.0000
M	14	45.44	3	514.89	0.0000
L	67	558.11	1	18, 970.58	0.0000
I by J	0	0.03	2	0.47	0.6261
I by M	0	0.05	3	0.57	0.6373
J by M	0	0.08	6	0.46	0.8413
I by L	0	0.13	1	4.56	0.0327
J by L	0	0.83	2	14.16	0.0000
M by L	1	2.95	3	33.44	0.0000
I by J by M	0	0.00	6	0.00	1.0000
I by J by L	0	0.00	2	0.04	0.9604
I by M by L	0	0.00	3	0.01	0.9988
J by M by L	0	0.90	6	5.10	0.0000
I by J by M by L	0	0.00	6	0.01	1.0000
Residuals		280.31	9, 528		

The four-way ANOVA was computed on the collection of 9,576 sample coefficient means obtained by pooling all the means across the levels of factor K

In sum, we can read the following ranking from the pairwise comparison analysis in factor M:

$$\text{extreme model} < (\text{uninformative model} \sim \text{average model}) < \text{slight model},$$

where $X < Y$ (resp. $X \sim Y$) means that the performance of alpha in model X is strictly worse than (resp. is not different than) that in model Y.

Our results demonstrate empirically that the extreme model was clearly more sensitive to fake perturbation than the other three models. In what follows, we propose an answer to explain what in the extreme model disposes the coefficient alpha to be affected more by increasing levels of replacements in the observed data. Since the sample alpha coefficient is based on a transformation of the covariance matrix $\mathbf{S}$, we studied how $\mathbf{S}$ was affected by increasing levels of fake replacements in the data in the four faking models. The main result was that, on the one side, the average covariances decreased by increasing amount of fake data perturbations, as fake observations usually tend to weaken the original relationships in the data. One obvious consequence is that the corresponding correlation matrices also tend towards the identity matrix $\mathbf{I}$ and this tendency was more evident (faster) in the extreme model. On the other side, the average variances followed a hump shaped curve as a function of increasing levels of fake replacements. In particular, the largest variances were observed for levels of fake replacements close to 50 %, whereas the lowest variances occurred at the lowest or highest levels of fake replacements. Also for the variances, this tendency was on average more evident in the extreme model. Moreover, in the extreme model the expected values in the fake data matrix were closer to the scale's upper bound (Q) irrespective of the values recorded in the original data matrix. Therefore, when the level of replacement is high the majority of values in the fake data matrix will be close to Q and the corresponding variances will tend toward a minimal value. The latter explains the very low variances observed in

the extreme model with very large percentage of replacements. Nonetheless, in the extreme model the proportion between the sum of all variances and the sum of all the covariances still tended towards 1 relatively faster as compared to the other faking models, thus reflecting a larger sensitivity to fake perturbation.

5 Empirical application

We illustrate the entire procedure using data collected by Vidotto and Marchesini (2000) on the interrelation between personality and student learning[2]. The current section is divided into two subsections: the first introduces the empirical data set and the result of a reliability analysis on the observed data; the second discusses how we can use SGR to reconstruct unknown true reliability scores on the basis of two alternative faking good scenarios: slight faking and extreme faking.

5.1 Empirical dataset and sample alpha result

Participants were 351 undergraduate students at the University of Modena (Italy). Ages ranged from 18 to 31, with a mean of 21.01 and a standard deviation of 2.28. Data consisted of the participants' responses to four of the 155 items of the Modena Resources Personality Inventory (MRPI) (Vidotto and Marchesini 2000) scored on a 5-point agree-disagree scale. The four items were the following:

1. Several times I gave up because what I wanted to reach was too difficult.
2. I am never really relaxed.
3. When I think of my future I have negative feelings.
4. I have difficulties in sleeping because I cannot stop thinking of my problems.

In the 5-point agree-disagree scale, value 1 denotes that a participant totally agrees with the statement, whereas value 5 means total disagreement with the statement. The four items were used as operational indicators of the theoretical construct emotional instability. This psychological construct was validated in a series of factorial studies that showed an overall plausibility of the MRPI questionnaire (Vidotto and Marchesini 2000). The reliability of the MRPI scale based on the four items was modest ($\widehat{\alpha} = 0.64$) denoting that the error variance computed on the observed (351×4) data matrix was not sufficiently smaller relative to the test score variance. However, the observed result may have been affected by fake good observations. More precisely, the participants may have deliberately attempted to manipulate their responses using larger values of the scale to create better impressions (Furnham 1986). This hypothesis was partially supported by the moderate ceiling effects observed in the data (see Table 3). Because no additional items on social desirability was available in the MPRI questionnaire, we decided to perform an SGR analysis on the basis of two hypothetical scenarios: slight faking and extreme faking.

5.2 SGR analysis

An SGR analysis was used to reconstruct the unknown true reliability score (reliability reconstruction problem) on the basis of two alternative faking good scenarios (sligth faking and extreme faking). In the first step of the SGR analysis we defined a simple generative model representing the hypothetical true model for the data. In particular, the generative

[2] We are grateful to Giulio Vidotto and Cristina Marchesini for providing us with such a data set.

Table 3 Descriptive statistics for the four MPRI items

Item	Value				
	1 %	2 %	3 %	4 %	5 %
1	7.12	11.97	17.09	26.50	37.32
2	13.11	15.67	24.50	28.49	18.23
3	15.38	14.35	17.95	19.09	33.05
4	3.13	4.56	15.10	19.37	57.83
Marginal %	9.69	11.64	18.67	23.37	36.62

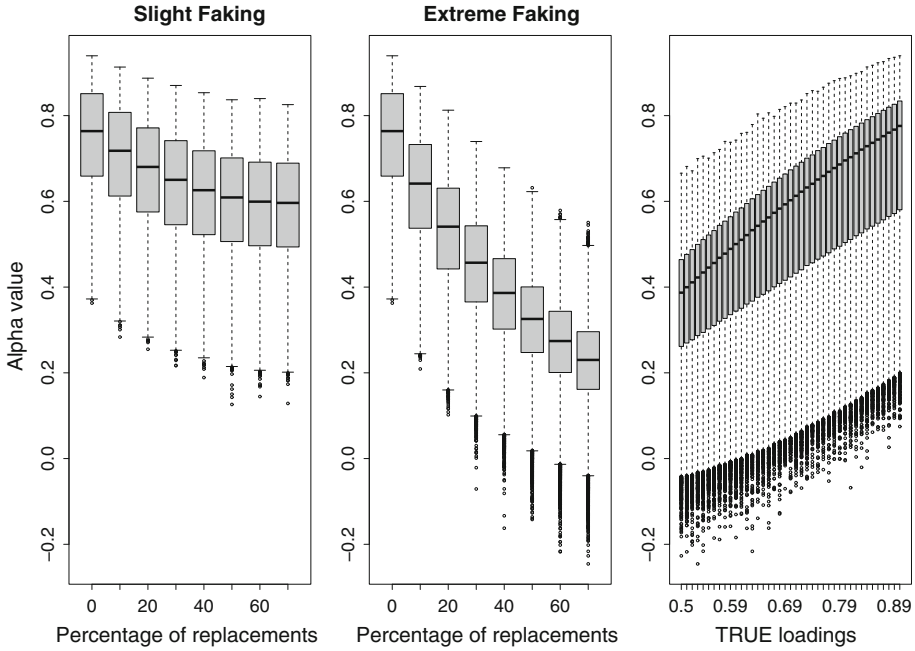

Fig. 6 Alpha coefficient distribution as a function of probability of replacement in the two faking models (*left panel* and *middle panel*). alpha coefficient distribution as a function of loading value λ in the single factorial model (*right panel*)

model consisted of a single factorial model with equal factor loadings for the four items considered in this study. However, since we ignored the original true value for the equal loadings in the factorial model, we simulated original datasets by drawing samples from populations corresponding to single factorial models with equal factor loadings λ ranging from .50 to .90 by step .01 (for a total of 41 distinct levels). The generative models were used to simulated new data without any ceiling effect for the four variables. Moreover, the generation process also allowed the simulation of original sample alpha values spanning the natural range of reliability (see Fig. 6, rightmost panel).

In the second step of the SGR analysis we implemented the slight and extreme faking processes by using the replacement distribution with parameters values $\gamma = 1.5$; $\delta = 4$ and

$\gamma = 4; \delta = 1.5$, respectively. In particular, the percentage of replacements (K) for the four items were assigned at 8 distinct levels ranging from 0 to 70 %. Finally, separately for each faking model and for each combination of levels (λ, k), we generated 5,000 distinct fake data matrices and saved the resulting sample alpha values. The results of the SGR analysis are shown in Fig. 6. As expected, extreme faking disposes the coefficient alpha to be affected more by increasing levels of replacements in the data.

The results of the SGR analysis can also be used to derive approximate CIs for the unknown true reliability score. A CI is usually interpreted as the range of values that encompass the population or true value, estimated by a certain statistic, with a given probability (e.g. Cohen 1990; Rice 1995). The interpretation that CIs provide an envelope within which the parameter value of interest is likely to lie makes sense even when trying to estimate true reliability scores from observed reliability computed on possible fake-corrupted data. We used a combined approach based on SGR simulations and classical CI constructs to derive approximate CIs for the unknown true reliability score. In particular, our procedure boiled down to the following two simple steps. First, we constructed a small fake interval $A = [0.64-0.025, 0.64+0.025]$ centered around the observed sample alpha, $\widehat{\alpha} = 0.64$. Next, separately for each value k of percentage of replacements (K) in the target faking model, we looked for all the simulated original (uncorrupted) data sets $\mathbf{D}$ such that the corresponding fake data sets $\mathbf{F}$ resulted in fake alpha values α^f lying in A. Finally, we used these original (uncorrupted) data sets $\mathbf{D}$ to derive the array of original alpha values α^o with which to compute an approximate 95 % CI for the true alpha in the population. Figure 7 shows the approximate 95 % CI as a function of percentage of replacements (K) in the two faking models. In particular, the approximate CI $\widehat{\text{CI}}$ was based on the empirical quantiles $\widehat{q}_{.025}$ and $\widehat{q}_{.975}$ computed using the array of original alpha values α^o. Note that, $\widehat{\text{CI}}$ is an estimate of the CI such that

$$P(\alpha^o \in \text{CI} | \alpha^f \in A) = 0.95,$$

with $P(\alpha^o \in \text{CI} | \alpha^f \in A)$ being the conditional probability over the two events $\alpha^o \in \text{CI}$ and $\alpha^f \in A$, respectively. Figure 7 also shows two clearly separated patterns for the approximate 95 % CIs in the two faking manipulations. Interestingly, under the slight faking hypothesis, we can reasonable expect a true reliability score for the MRPI scale which is not larger than 0.85 even if a very relevant portion of the original data has been actually corrupted by fake observations. By contrast, for the extreme faking hypothesis, small amounts of fake responses in the data can dramatically affect the true reliability of the MRPI scale. Therefore, in this latter context, we can reasonable expect a very large true reliability score for the MRPI scale.

6 Concluding remarks

The reader may have already noticed some similarities between SGR and standard Monte Carlo experiments. For example, the idea of generating new data sets. However, the two approaches are substantially different. Usually a Monte Carlo experiment uses a hypothesized model to generate new data under various conditions (e.g. Robert and Casella 2004). Therefore the simulated data are used to evaluate some characteristics of the model. This, of course, implies that the distribution of the random component in the assumed model must be known, and it must be possible to generate pseudorandom samples from that distribution under the desired conditions planned by the researcher. Instead of using only the hypothesized model structure to generate simulated data sets, SGR uses the original data sample in order to generate a new family of data sets. In particular, these new data sets are obtained by adding structured perturbations in the original data set. The availability of external knowledge about

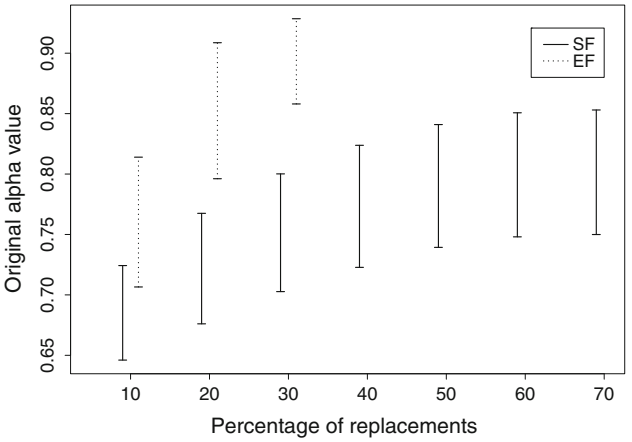

Fig. 7 Approximate 95 % CI as a function of percentage of replacement in the two faking models (sligth faking *SF*, extreme faking *EF*). To achieve reasonable estimation stability in the tail regions of the quantile estimates, we computed only those confidence intervals based on at least 2,000 original alpha values α^o

process faking may suggest the modeling of highly structured malingering scenarios. In the latter case, each new sample represents an alternative malingering scenario which is directly derived from both the original sample and the assumed malingering model. Next, the result of a target criterion, such as alpha, can be compared with the ones obtained from the perturbed samples.

In this paper we used SGR to study how fake observations can affect the behavior of the reliability of a composite score under the essentially tau-equivalent condition. However, such a simplified assumption is potentially a concern when more complex empirical data are considered. But although this limitation, we still preferred to start with this simplified condition to better evaluate our SGR method under the purest and simplest reliability scenario. Nonetheless, it is important to stress that the application of the SGR method is in general not limited to these simplified conditions. In particular, various possible extensions of the SGR procedure to study reliability coefficients could be considered, both from the point of view of the relaxation of the tau-equivalent assumption and the structure of the faking models. For example, if our observations are partitioned in two distinct groups, we may assume different probabilities of faking in the two groups. This latter scenario can be modelled by two distinct replacement models, one for each of the two groups. Similarly, we might relax the tau-equivalent condition by assuming that the loadings in the factorial model are actually sampled from a proper probability distribution that models natural variabilities in the relationship between the items and the factor. Furthermore, we could also consider more complex factorial models based on multidimensional factors with or without equal factor loadings. In these more realistic scenarios, it would be interesting to compare the sensitivity of different reliability indices with respect to fake data perturbations.

References

Allen, M.J., Yen, W.M.: Introduction to Measurement Theory. Brooks/Cole, Monterey (1979)
Bentler, P.A., Woodward, J.A.: Inequalities among lower bounds to reliability: with applications to test construction and factor analysis. Psychometrika **45**, 249–267 (1980)
Cohen, J.: Things I have learned (so far). Am. Psychol. **45**, 1304–1312 (1990)

Cortina, J.M.: What is coefficient alpha? An examination of theory and applications. J. Appl. Psychol. **78**, 98–104 (1993)
Cronbach, L.J.: Coefficient alpha and the internal structure of tests. Psychometrika **16**, 297–334 (1951)
Duhachek, A., Coughlan, A.T., Iacobucci, D.: Results on the standard error of the coefficient alpha index of reliability. Mark. Sci. **24**, 294–301 (2005)
Fan, X., Felsovalyi, A., Sivo, S.A., Keenan, S.: SAS for Monte Carlo Studies: A Guide for Quantitative Researchers. SAS Institute, Inc., Cary (2002)
Ferrando, P.J.: Factor analytic procedures for assessing social desiderability 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 eqution model at the individual and group level for assessing faking-related change. Struct. Equ. Model. **18**, 91–109 (2011)
Fleiss, J.L.: Statistical Methods for Rates and Proportions, 2nd edn. Wiley, New York (1981)
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)
Green, S.B., Lissitz, R.W., Mulaik, S.A.: Limitations of coefficient alpha as an index of test unidimensionality. Educ. Psychol. Meas. **37**, 827–838 (1977)
Guttman, L.: A basis for analyzing test–retest reliability. Psychometrika **10**, 255–282 (1945)
Hall, R.C., Hall, R.C.: Detection of malingered PTSD: an overview of clinical, psychometric, and physiological assessment: where do we stand? J. Forensic Sci. **52**, 717–725 (2007)
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)
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)
Kaiser, H.F., Dickman, K.: Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. Psychometrika **27**, 179–182 (1962)
Komorita, S.S., Graham, W.K.: Number of scale points and the reliability of scales. Educ. Psychol. Meas. **25**, 987–995 (1965)
Leite, W.L., Cooper, L.A.: Detecting social desiderability bias using factor mixture models. Multivar. Behav. Res. **45**, 271–293 (2010)
Li Gwet, K.: K.: Variance estimation of nominal-scale inter-rater reliability with random selection of raters. Psychometrika **73**, 407–430 (2008)
Lissitz, R.W., Green, S.B.: Effect of the number of scale points on reliability: a Monte Carlo approach. J. Appl. Psychol. **60**, 10–13 (1975)
Liu, Y., Wu, A.D., Zumbo, B.D.: The impact of outliers on Cronbachs coefficient alpha estimate of reliability: ordinal/rating scale item responses. Educ. Psychol. Meas. **70**, 5–21 (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)
Lord, F.M., Novick, M.R.: Statistical Theories of Mental Test Scores. Addison-Wesley, Reading (1968)
Marshall, E.: How prevalent is fraud? That's a million-dollar question. Science **290**, 1662–1663 (2000)
Maydeu-Olivares, A., Coffman, D.L., Hartmann, W.M.: Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. Psychol. Methods **12**, 157–176 (2007)
McDonald, R.P.: Factor Analysis and Related Methods. Erlbaum, Hillsdale (1985)
McFarland, L.A., Ryan, A.M.: Variance in faking across noncognitive measures. J. Appl. Psychol. **85**, 812–821 (2000)
Mittenberg, W., Patton, C., Canyock, E.M., Condit, D.C.: Baserates of malingering and symptom exaggeration. J. Clin. Exp. Neuropsychol. **24**, 1094–1102 (2002)
Morgan, M.G., Henrion, M., Small, M.: Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press, New York (1990)
Muthén, B.: A general structural equation model with dichotomous, ordered categorical and continuous latent variables indicators. Psychometrika **49**, 115–132 (1984)
Novick, M.R.: The axioms and principal results of classical test theory. J. Math. Psychol. **3**, 1–18 (1966)
Novick, M.R., Lewis, C.: Coefficient alpha and the reliability of composite measurements. Psychometrika **32**, 1–13 (1967)
Nunnally, J., Bernstein, I.: Psychometric Theory. McGraw-Hill, New York (1994)

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 Social-Psychological Attitudes, pp. 17–59. Academic Press, New York (1991)
R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org (2012)
Revelle, W.: Hierarchical cluster-analysis and the internal structure of tests. Multivar. Behav. Res. **14**, 57–74 (1979)
Rice, J.A.: Mathematical Statistics and Data Analysis, 2nd edn. Duxbury Press, Belmont (1995)
Robert, C.P., Casella, G.: Monte Carlo Statistical Methods, 2nd edn. Springer- Verlag, New York (2004)
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**, 634–644 (1998)
Sijtsma, K.: On the use, the misuse, and the very limited uselfulness of Cronbach's alpha. Psychometrika **74**, 107–120 (2009)
Van der Geest, S., Sarkodie, S.: The fake patient: a research experiment in a Ghanaian hospital. Social Science & Medicine. Medicine **47**, 1373–1381 (1998)
Vidotto, G., Marchesini, C.: La realizzazione professionale. Franco Angeli, Analisi delle risorse personali e dei processi decisionali per l'orientamento scolastico-professionale, Milano (2000)
Weng, L.: Impact of the number of response categories and anchor labels on coefficient alpha and test–retest reliability. Educ. Psychol. Meas. **64**, 956–972 (2004)
Whitby, O.: Estimation of Parameters in the Generalized Beta Distribution (Technical Report No. 29). Department of Statistics, Standford University, Stanford (1971).
Woods, C.M.: Careless responding to reverse-worded items: implications for confirmatory factor analysis. J. Psychopathol. Behav. Assess. **28**, 189–194 (2006)
Yuan, K.-H., Bentler, P.M.: On normal theory based inference for multilevel models with distributional violations. Psychometrika **67**, 539–561 (2002)
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)
Zickar, M.J., Robie, C.: Modeling faking good on personality items: an item-level analysis. J. Appl. Psychol. **84**, 551–563 (1999)
Zinbarg, R.E., Revelle, W., Yovel, I., Li, W.: Cronbach's α, Revelle's β, and McDonald's ω_H: their relations with each other and two alternative conceptualizations of reliability. Psychometrika **70**, 123–133 (2005)