On the sensitivity of Cronbach's alpha to fake data

Massimiliano Pastore[1] and Luigi Lombardi[2]

[1] University of Padua, Department of Developmental and Social Psychology, via Venezia 10, Padova, Italy
(E-mail: mpastore@unipd.it)

[2] University of Trento (Rovereto Pole), Department of Cognitive Science and Education, corso Bettini 31, Rovereto (TN), Italy
(E-mail: luigi.lombardi@unitn.it)

Abstract. In this paper, we address the problem of evaluating Cronbach's alpha (Cronbach[1]) when fake data are considered. Starting from the introduction of an integrated simulation approach, called MC+SGR, we evaluate the sensitivity of alpha to two relevant scenarios of perturbation: symmetric faking and asymmetric faking. We also report a simple example of correction table for the α statistic.

Keywords: Cronbach's alpha, Fake data, Sample Generation by Replacements, Monte Carlo simulations.

1 Introduction

How can we evaluate the impact of fake information in real life contexts? In some circumstances, individuals may tend to distort their behaviors or actions in order to reach specific goals. For example, in personnel selection job applicants may misrepresent themselves on a personality survey hoping to increase the likelihood of being offered a job. 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. A major problem in self-report measures is that in many occasions there is no basis to assume that subjects are responding honestly. This problem is common in areas like psychology (Hopwood *et al.*[4]), organizational and social science (Van der Geest and Sarkodie[10]), forensic medicine (Gray *et. al*[3]), and scientific frauds (Marshall[7]).

In such a contexts, self-report measures confront the researcher with a crucial question: *If data included fake data (fd) points, 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 data analysis is the situation in which a researcher needs to estimate the internal consistency or reliability of survey scores. Within this context the FEQ can be rewritten as: *If data contained k% fd points, what would the true reliability of my survey score be?* In particular, we would expect that a reliability index should approach its maximum under uncorrupted data, but also degrade substantially under massive data perturbation.

One of the most popular statistics to compute items reliability is Cronbach's alpha (Cronbach[1]) which is widely used in psychology, business, nursing, and many other disciplines. Several studies have been conducted to evaluate some important features of Cronbach's alpha. For example, MC simulations have been used to study the robustness of alpha's confidence intervals under violation of symmetry (e.g., Maydeu-Olivares *et al*[8]), as well as the impact of sample size (e.g., Duhachek *et al.*[2]) and type of estimation method on the alpha statistic (van Zyl *et al.*[11]). Notwithstanding, in the data analysis literature the issue of evaluating the sensitivity of Cronbach's alpha to fake data has been substantially neglected.

In this paper we try to fill this gap by proposing an integrated approach, called MC+SGR, which combines standard Monte Carlo techniques and the data perturbation procedure SGR (Sample Generation by Replacements; Lombardi *et al.*[6]; Pastore and Lombardi[9]) to evaluate the FEQ on the alpha statistic.

The paper is organized as follows. Section 2 outlines the SGR approach. Section 3 describes the MC+SGR simulation study for studying Cronbach's alpha under two relevant scenarios of faking. In Section 4 we discuss results of the simulation study. Finally, Section 5 reports some concluding remarks.

2 SGR

In many self-reported surveys the resulted dataset often includes incomplete records (missing data) and/or fake records (fake data). In such a context it is worth considering data analytic tools to evaluate the FEQ. Let $\mathbf{X}$ and ϕ be the original data set prior to any fake perturbation and a statistic on $\mathbf{X}$, respectively. Since we usually observe data only after that the faking process has taken place, many times the original data $\mathbf{X}$ cannot be directly observable. The main goal of a replacement analysis is the evaluation of ϕ under the so called *fake data space* (FDS) of $\mathbf{X}$. FDS consists of a collection of new data sets $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_B$, where each $\mathbf{F}_b$ ($b = 1, \ldots, B$) represents a different fake perturbation of $\mathbf{X}$. The perturbation is carried out by means of a probabilistic model $\mathbf{R}$ that mimics the faking process of interest. Next, the distribution of results $\phi(\mathbf{F}_1), \ldots, \phi(\mathbf{F}_B)$ is finally evaluated and eventually compared to the original result $\phi(\mathbf{X})$.

A case of particular interest is when $\mathbf{X}$ is actually observable like, for example, in Monte Carlo simulation studies. More precisely, we think of the simulated dataset $\mathbf{X}$ as being represented by an $I \times J$ matrix with I observations each containing J Likert-type items. A Likert-type item being an ordinal variable that takes values on a small set $\mathcal{Q} = \{1, ..., Q\}$ (e.g., $Q = 2$ for dichotomous items and $Q = 5$ for five-point Likert items). The main idea of the SGR approach is to construct a new $I \times J$ data matrix $\mathbf{F}$ (called the *fake data matrix*) from $\mathbf{X}$ by manipulating each element x_{ij} in $\mathbf{X}$ ($\forall i = 1, \ldots, I; \forall j = 1, \ldots, J$) according to a $Q \times Q$ replacement matrix $\mathbf{R}$ that models the faking process. In particular, element r_{pq} of $\mathbf{R}$ $(\forall(p,q) \in \mathcal{Q} \times \mathcal{Q})$ denotes the conditional probability $p(f_{ij} = q | x_{ij} = p)$ of replacing an original observed value p in entry (i,j) of $\mathbf{X}$ with the new value q. By repeating the same procedure B times we can generate the FDS: $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_B$, and thus provide the distribution of ϕ under the FDS.

SGR grounds on three basic assumptions: 1) the *principle of indifference* 2) the *local replacement property* and 3) the *total error decomposition condition.* The first assumption grounds on the idea that in the absence of further knowledge about the process of faking all entries in $\mathbf{X}$ are assumed to be equally likely in the process of replacement. The second assumption entails that the conditional probability of replacing an initial value in $\mathbf{X}$ with another value in $\mathcal{Q}$ only depends on that initial value. More precisely, all the entries in $\mathbf{X}$ showing a same observed value, say p, must have an identical conditional probability distribution of replacement:

$$\mathbf{R}_p = (r_{p1}, \ldots, r_{pQ}).$$

Finally, the third assumption regards the way total error can be decomposed in SGR. Let $\mathbf{T}$ be the true (but unknown) $I \times J$ data matrix, then the following identities hold:

$$\begin{aligned} \mathbf{F} &= \mathbf{X} + \mathbf{E}^* && (1) \\ &= (\mathbf{T} + \mathbf{E}) + \mathbf{E}^* && (2) \\ &= \mathbf{T} + \mathbf{E}^{\text{tot}} && (3) \end{aligned}$$

where $\mathbf{E}^* = \mathbf{X} - \mathbf{F}$ is the *fake data error* matrix, $\mathbf{E} = \mathbf{T} - \mathbf{X}$ is the *standard data error* matrix, and $\mathbf{E}^{\text{tot}} = \mathbf{E} + \mathbf{E}^*$ is the *total error* matrix. The error components $\mathbf{E}$ and $\mathbf{E}^*$ are assumed to be stochastically independent. Two relevant instances of the replacement matrix $\mathbf{R}$ will be presented in the next section, whereas more sophisticated assumptions about the replacement matrix $\mathbf{R}$ will be discussed in the last section of this paper.

3 Simulation study

We first introduce the main tokens of our simulation study. Next, we will report all details of the MC+SGR design.

3.1 Cronbach's alpha

Cronbach's alpha coefficient (Cronbach[1]) is used as a measure of internal consistency of a survey or test with a limited number of Likert items. More precisely, consider a survey composed of J items, $X_1, \ldots, X_J$, intended to measure a single latent attribute or dimension. The main goal is to determine the reliability of the test score $Y = X_1 + \ldots + X_J$, that is, the percentage of variance of the survey score that is due to the attribute of which the items are indicators. In a sample of I respondents, the coefficient alpha statistic is

$$\alpha(\mathbf{X}) = \frac{J}{J-1}\left(1 - \frac{\sum_{j=1}^{J} s_{jj}}{\sum_{j=1}^{J}\sum_{j'=1}^{J} s_{jj'}}\right) \tag{4}$$

where $s_{jj'}$ is the (j, j')-element of the covariance matrix $\mathbf{S}$ of an $I \times J$ response matrix $\mathbf{X}$. Clearly, the value of α does not exceed 1 which, in turn, denotes a perfect internal consistency. In our study Cronbach's alpha will play the role of dependent variable in the MC+SGR simulation design.

3.2 Models of faking

Several malingering contexts can be proposed according to the sensitivity of self-report measures. In this paper we will limit our attention to two simple, but relevant faking scenarios: symmetric faking and asymmetric faking. In the first case we assume the total absence of knowledge about the process of faking. Given an observed value p in $\mathbf{X}$, all values in $\mathcal{Q} \setminus \{p\}$ are assumed to be likely in the process of replacement of p. In other words, this scenario assumes a pure random malingering model. In contrast, the availability of external knowledge about the process of faking may suggest the modeling of more complex scenarios. For example, in personnel selection some subjects are likely to fake a personality questionnaire to match the ideal candidate's profile (positive impression management or fake-good process). In this second case it could be reasonable to consider a conditional replacement model in which the conditioning is a function of response polarity.

We used two different replacement matrices $\mathbf{R}_{su}$ and $\mathbf{R}_{asu}$ to provide a simple characterization of the faking scenarios described above. $\mathbf{R}_{su}$ is called the *symmetric uniform replacement* matrix and represents a context in which responses are subject to unpolarized uniform random faking. $\mathbf{R}_{su}$ is defined as

$$r_{pq} = \begin{cases} \frac{1-\gamma}{Q-1}, & \forall p \neq q \\ \gamma, & p = q \end{cases} \tag{5}$$

In Eq. (5), γ denotes the probability of non-replacement. $\mathbf{R}_{asu}$ is called the *asymmetric uniform replacement* matrix and represents a fake-good scenario in which $f_{ij} > x_{ij} (\forall i = 1, ..., I; \forall j = 1, ..., J)$. In particular,

$$r_{pq} = \begin{cases} \frac{1-\gamma}{Q-p}, & \forall p < q \\ \gamma, & p = q \\ 0, & \forall p > q \end{cases} \tag{6}$$

also in Eq. (6), γ indicates the probability of non-replacement.

3.3 Extended MC simulations: MC+SGR

In order to evaluate the impact of fd points on Cronbach's alpha we used a standard MC approach to generate a series of artificial data sets $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_B$ on the basis of statistical properties of a factorial model. For sake of simplicity, in our study we opted for a single factorial model with four observed items and identical loadings. Next, for each MC simulated data $\mathbf{X}_b$ $(b = 1, \ldots, B)$ we generated a new perturbed matrix $\mathbf{F}_b$ on the basis of a particular model of faking $\mathbf{R}$. Therefore, we may think of each new perturbed data $\mathbf{F}_b$ as an alternative *informative scenario* which is directly derived from the original simulated MC sample $\mathbf{X}_b$. Next, the behavior of a the reliability coefficient α was evaluated with respect to the collection of perturbed samples $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_B$ that corresponds to our particular FDS. In this case, of course, the distributional properties of α are not those that simply hold under a particular model hypothesis (like for standard Monte Carlo simulation studies); rather they are the properties under a model whose parameters corresponds to values fitted from both the MC generating process and the structured collection of perturbed SGR samples that are generated from the given MC samples. We call this integrated procedure the MC+SGR approach.

3.4 Simulation design

Now we are in the position to provide all details of our simulation design. In order to generate the initial data matrices, we used a single factorial model with four observed items and identical loadings $\lambda_v = \lambda$ $(v = 1, \ldots, 4)$. Moreover, by varying the values of the loading λ we were able to generate data with α values ranging from .34 (very mild α) to .91 (very high α). The following procedural steps were repeated for each value of λ:

1. 5000 raw-data sets $\mathbf{X}_b$ with $J = 4$ items and $I = 100$ observations were generated according to a standard MC procedure. Next, each $\mathbf{X}_b$ $(b = 1, ..., 5000)$ was converted to a Q-point scale $(Q = 2, 5)$ using the method described by Jöreskog and Sörbom[5]. This substep allowed to simulate Likert-type data at two different levels: dichotomous items and five-point Likert items that correspond to the most common data in self-report surveys.
2. For each discrete matrix $\mathbf{X}_b$ we constructed a collection of fake matrices ${}_z\mathbf{F}_{b,k}$ by using the replacement matrix $\mathbf{R}_z$ $(z = su, asu)$ with non-replacement proportion $\gamma = 1 - (k/100)$ and $k = 5, 10, \ldots, 100$.
3. For each perturbed data matrix ${}_z\mathbf{F}_{b,k}$ the reliability coefficient α was finally evaluated.

In sum, four parameters were systematically varied in a complete four-factorial design:

1. the factorial loading λ, at 23 levels: $.400, .425, \ldots, .950$;

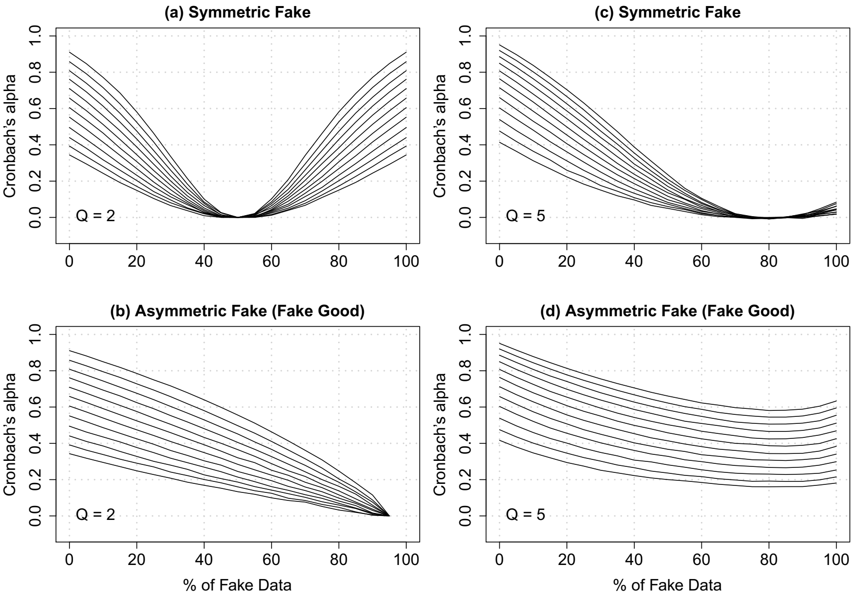

Fig. 1. Fake effect charts for the reliability coefficient α. Fake effect for dichotomous items in the symmetric (a) and asymmetric (b) conditions. Fake effect for five-point Likert items in the symmetric (c) and asymmetric (d) conditions.

2. the item scale Q, at two levels: 2 and 5;
3. the model of faking z, at two levels: symmetric uniform fake (su) and asymmetric uniform (asu) fake (fake good);
4. the percentage of replacements k, at 20 levels: $5, 10, \ldots, 100$.

The whole procedure generated a total of $9200000 = 5000 \times 23 \times 2 \times 2 \times 20$ new perturbed data matrices.

4 Results

Figure 1 shows the medians of the reliability coefficient α as a function of % of replacement, model of faking, and item scale for twelve initial population values of α ranging from .34 to .91. Note that in our simulation design the initial alpha values always correspond to the hypothetical non-replacement condition ($k = 0$). So, for example in Figure 1a, the first curve indicates the median values of α as a function of percentage of replacement in dichotomous data with an initial α value of approximately .91. The second curve that for $\alpha = .88$, and so on.

Figure 1a shows the fake effect chart for dichotomous items in the symmetric uniform condition. In this panel we may observe a perfect symmetric convex pattern with minimum at $k = 50$ which, in turn, corresponds to the

α \ fake %	5	10	15	20	30	50	80
$\leq .05$	.00 − .32	.02 − .37	.07 − .40	.12 − .45	.17 − .51	.24 − .59	.27 − .65
(.05,.10]	.07 − .30	.14 − .39	.17 − .43	.18 − .50	.24 − .53	.28 − .63	.31 − .70
(.10,.15]	.11 − .34	.16 − .41	.20 − .47	.23 − .51	.26 − .55	.31 − .65	.33 − .73
(.15,.20]	.17 − .37	.20 − .44	.24 − .49	.26 − .52	.28 − .59	.32 − .68	.35 − .75
(.20,.25]	.22 − .40	.24 − .46	.26 − .51	.28 − .55	.31 − .61	.35 − .70	.37 − .79
(.25,.30]	.26 − .44	.29 − .49	.30 − .54	.31 − .58	.34 − .65	.38 − .74	.41 − .82
(.30,.35]	.30 − .47	.32 − .53	.34 − .57	.35 − .61	.38 − .68	.42 − .78	.45 − .86
(.35,.40]	.35 − .50	.36 − .56	.38 − .61	.39 − .65	.42 − .71	.46 − .82	.50 − .89
(.40,.45]	.40 − .54	.41 − .60	.43 − .65	.44 − .69	.47 − .75	.53 − .86	.56 − .92
(.45,.50]	.45 − .59	.46 − .64	.48 − .69	.50 − .73	.53 − .80	.60 − .89	.66 − .94
(.50,.55]	.50 − .63	.52 − .69	.54 − .73	.56 − .77	.60 − .84	.67 − .92	.73 − .95
(.55,.60]	.56 − .68	.58 − .73	.60 − .78	.63 − .81	.67 − .87	.75 − .94	.80 − .95
(.60,.65]	.61 − .72	.64 − .78	.66 − .82	.69 − .86	.74 − .91	.81 − .95	.85 − .96
(.65,.70]	.67 − .77	.70 − .82	.72 − .86	.75 − .89	.80 − .94	.86 − .96	.88 − .96
(.70,.75]	.72 − .81	.75 − .86	.78 − .90	.81 − .93	.85 − .95	.89 − .96	.90 − .96
(.75,.80]	.78 − .86	.80 − .90	.83 − .93	.86 − .95	.89 − .96	.92 − .96	.90 − .96
(.80,.85]	.83 − .90	.86 − .94	.88 − .95	.90 − .96	.92 − .96	.96 − .96	−
(.85,.90]	.88 − .94	.90 − .96	.92 − .96	.93 − .96	.94 − .97	−	−
(.90,.95]	.93 − .96	.94 − .97	.95 − .97	.96 − .97	−	−	−
(.95,1]	.96 − .97	−	−	−	−	−	−

Table 1. Alpha correction table for samples with size $I = 100$ and $J = 4$ (five-point Likert items). Given an empirically observed alpha [rows] and an hypothetical percentage of fake good [column], the table then yields the MC+SGR 90%-percentile intervals of the true population alpha.

maximum loss for alpha. Note that, if $k = 50$, then $\gamma = 1 - (50/100) = .50$. In other words, $r_{pq} = .50$ for all possible combinations of dichotomous values p and q. This latter case characterizes the maximum entropy condition for the generated FDS and, therefore, minimum value for alpha. By contrast, if $k = 100$, then $\gamma = 0$; that is to say, we replace the whole data matrix with its dual representation ($1 \mapsto 2$ and $2 \mapsto 1$). Since, α is based on the covariance matrix (see Eq. (4)), its value will not change when computed on the dual representation of the data. Finally, a large dominance relation can be read from the alpha curves shown in Figure 1a. In particular, if α_1 and α_2 denote two distinct initial values for alpha such that $\alpha_1 > \alpha_2$, then the α_1-curve dominates the α_2-curve.

Figure 1b shows the fake effect chart for dichotomous items in the asymmetric uniform condition (fake good). This panel shows a decreasing monotonic pattern converging at $k = 95$, which corresponds to the maximum loss for alpha. Note that $k = 95$ is also the maximal replacement percentage that is allowed in this condition. Figure 1c and Figure 1d show the fake effect charts for the 5-point Likert items in the symmetric uniform and asymmetric uniform conditions, respectively. Results can be interpreted accordingly.

Our MC-SGR study can also be used to construct correction tables for empirically observed α (see Table 1). So, for example, let us assume that we observed $\alpha = .67$ for a data set with 100 observations and four 5-point Likert items. Moreover, we also assume that an asymmetric process of faking (i.e., fake good) has affected approximately 20% of our data. Then, by using

Table 1 we can infer that the true population alpha lies within .75 and .89 (MC+SGR 90%-percentile interval).

5 Concluding remarks

In this paper we presented an integrated MC+SGR approach to study how fake observations can affect the behavior of the well known Cronbach's alpha statistic. Various possible extensions of the MC+SGR approach could be considered, both from the point of view of the relaxation of the three main SGR assumptions and the structure of the replacement matrix **R**. 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 modeled by two distinct replacement matrices, one for each of the two groups.

References

1.L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334, 1951.

2.A. Duhachek, A. T. Coughlan, and D. Iacobucci. Results on the standard error of the coefficient alpha index of reliability. *Marketing Science*, 24, 294-301, 2005.

3.N. S. Gray, M. J. MacCulloch, J. Smith, M. Morris, and R. J. Snowden. Violence viewed by psychopathic murderers. *Nature*, 423, 497, 2003.

4.C. J. Hopwood, C. A. Talbert, L. C. Morey and R. Rogers. 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, 2008.

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

6.L. Lombardi, M. Pastore and M. Nucci. 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, Kluwer Academic Publishers, 2004.

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

8.A. Maydeu-Olivares, D. L. Coffman, W. M. Hartmann. Asymptotically distribution-free (ADF) interval estimation of coefficient Alpha. *Psychological Methods*, 12, 157–176, 2007.

9.M. Pastore and L. Lombardi. Evaluating the sensitivity of goodness-of-fit indices to data perturbation: An integrated MC-SGR approach; in: J. Janssen and P. Lenca (Eds.), *Proceedings of the XIth International Symposium on Applied Stochastic Models and Data Analysis*, pp. 1452–1459, ENST, 2005.

10.S. Van Der Geest and S. Sarkodie. The fake patient: A Research experiment in a ghanaian hospital. *Social Science & Medicine*, 47, 1373–1381, 1998.

11.J. M. van Zyl, H. Neudecker, and D. G. Nel. On the distribution of the maximum likelihood estimator of Cronbachs alpha. *Psychometrika*, 65, 271-280, 2000.