sgr: A Package for Simulating Conditional Fake Ordinal Data

by Luigi Lombardi and Massimiliano Pastore

Abstract Many self-report measures of attitudes, beliefs, personality, and pathology include items that can be easily manipulated by respondents. For example, an individual may deliberately attempt to manipulate or distort responses 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. This article introduces the package **sgr** that can be used to perform fake data analysis according to the sample generation by replacement approach. The package includes functions for making simple inferences about discrete/ordinal fake data. The package allows to quantify uncertainty in inferences based on possible fake data as well as to study the implications of fake data for empirical results.

Introduction

How can we evaluate the impact of fake information in real life contexts? In nature, some individuals tend to distort their behaviors or actions in order to reach specific goals. In some species, for example, wimpy animals may not signal their real social value by faking a higher status to deceive other competitors. Similarly, in personnel selection some job applicants may misrepresent themselves on a personality test hoping to increase the likelihood of being offered a job. Being able to discriminate between honest and fraudulent signals and evaluating the impact of counterfeit elements crucially depend on the way we can reason about the whole process of faking. A coherent knowledge of the type or structure of faking processes may lead to stronger inferences that lie on or close to what we may call the genuine, but probably hidden representation of a manifest behavior. In general fake data may alter a large variety of self-report measures. This problem is particularly relevant for discrete/ordinal data collected in sensitive environments such as, for example, risky sexual behaviors, drug addictions, tax evasion, political preferences, financial compensation, and personnel selection. More in general, researchers interested in the study of human behavior in areas like psychology (Hopwood et al., 2006), organizational and social science (Van der Geest and Sarkodie, 1998), psychiatry (Beaber et al., 1985), forensic medicine (Gray et al., 2003), scientific frauds (Marshall, 2000), and economics (Crawford, 2003) may face the fake data problem when analyzing and interpreting empirical data.

In this article, we discuss the **sgr** package that we have developed for running fake data analysis according to the *sample generation by replacement* (SGR) approach (Lombardi and Pastore, 2012). SGR is a data simulation procedure to generate artificial samples of fake discrete/ordinal data. The main characteristic of the SGR approach is that it allows detailed explorations of what outcomes are produced by particular sets of faking assumptions. By changing the input in the faking model parameters and showing the effect on the outcome of a model, SGR provides a *what-if-analysis* of the faking scenarios. Therefore, SGR can be used to quantify uncertainty in inferences based on possible fake data as well as to evaluate the implications of fake data for statistical results. To illustrate, let us consider the following example where a researcher is interested in studying the relationship between therapy-uncompliance indicators (e.g., forgetting the treatment) and unsafe behaviors indicators (e.g., drinking alcohol) in a group of liver transplant patients. Generally, patients diagnosed with alcohol dependence who follow a pharmaceutical regimen after the liver transplant would deliberately answer fraudulently a question about drinking alcohol due to abstinence from ethanol and social desirability factors (e.g. Foster et al., 1997). In this context, an SGR analysis can help in testing for potential influence of faking the drinking alcohol self-report measure on the strength of the relationship between therapy-uncompliance and unsafe behaviors indicators. More specifically, how sensitive are the empirical associations to possible fake observations in the drinking alcohol self-report measure? Are the conclusions still valid under one or more scenarios of faking (e.g., slight, moderate, and extreme faking) for the drinking alcohol variable?

In general, SGR takes an interpretation perspective by incorporating in a global model all the available information about the process of faking and the underlying true model representation. This makes SGR related in spirit to other statistical approaches such as, for example, uncertainty and sensitivity analysis (Helton et al., 2006) and prospective power analysis (Cohen, 1988) which are all characterized by an attempt to directly quantify uncertainty of general statistics computed on the data by means of specific hypotheses.

The rest of the paper is organized as follows. The next section reviews the SGR framework and its basic implementations using the **sgr** package. The following section provides three examples

illustrating the application of **sgr** to faking scenarios. The final section discusses limitations and future implementations in **sgr** components beyond the general scheme presented here.

The SGR framework

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 (*data generation process*) and the model representing the faking process to perturb the data (*data replacement process*). By repeatedly sampling data from the SGR procedure we can generate the so called fake data sample (FDS) and eventually study the distribution of some relevant statistics computed on this simulated data space. In SGR the data generation process is modeled by means of standard Monte Carlo procedures for ordinal data whereas the data replacement process is implemented using ad hoc probabilistic faking models. In sum, the overall generative process is split into two conceptually independent and possibly simpler components (divide and conquer strategy).

With regard to the fake-data problem in general, we think of the data in the generation process as being represented by an $n \times m$ matrix $\mathbf{D}$, that is to say, n i.i.d. observations (hypothetical participants) each containing m elements (hypothetical participant's responses). We assume that entry d_{ij} of $\mathbf{D}$ ($i = 1, \ldots, n$; $j = 1, \ldots, m$) takes values on a small ordinal range $V_q = \{1, \ldots, q\}$ (for the sake of simplicity, in this presentation we assume identical ordinal scales). In particular, let $\mathbf{d}_i$ be the $(1 \times m)$ array of $\mathbf{D}$ denoting the pattern of responses of participant i. The response pattern $\mathbf{d}_i$ is a multidimensional ordinal random variable with probability distribution $p(\mathbf{d}_i|\theta_D)$, where θ_D indicates the vector of parameters of the probabilistic model for the data generation process. The main idea of the replacement approach is to construct a new $n \times m$ ordinal data matrix $\mathbf{F}$, called the *fake data matrix* of $\mathbf{D}$, by manipulating each element d_{ij} in $\mathbf{D}$ according to a replacement probability distribution. Let $\mathbf{f}_i$ be the $(1 \times m)$ array of $\mathbf{F}$ denoting the replaced pattern of fake responses of participant i. The fake response pattern $\mathbf{f}_i$ is a multidimensional ordinal random variable with conditional replacement probability distribution

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

where θ_F indicates the vector of parameters of the probabilistic faking model.

It is important to note that in the standard SGR framework the replacement distribution $p(\mathbf{f}_i|\mathbf{d}_i, \theta_F)$ is restricted to satisfy the *conditional independence* (CI) assumption (see Lombardi and Pastore, 2012; Pastore and Lombardi, 2014). More precisely, in the replacement distribution each fake response f_{ij} only depends on the corresponding data observation d_{ij} and the model parameter θ_F. Therefore, because the patterns of fake responses are also i.i.d. observations, the simulated data array $(\mathbf{D}, \mathbf{F})$ is drawn from the joint probability distribution

$$p(\mathbf{D}, \mathbf{F}|\theta_D, \theta_F) = \prod_{i=1}^{n} p(\mathbf{d}_i|\theta_D) p(\mathbf{f}_i|\mathbf{d}_i, \theta_F) \tag{2}$$

$$= \prod_{i=1}^{n} p(\mathbf{d}_i|\theta_D) \prod_{j=1}^{m} p(f_{ij}|d_{ij}, \theta_F) \tag{3}$$

In the last section of this article we will discuss some potential limitations of the conditional independence assumption in real application domains of the SGR approach.

Data generation process

In general, several options are available to represent the data generation process (Muthén, 1984; Jöreskog and Sörbom, 1996; Moustaki and Knott, 2000; Samejima, 1969). In the current version of the **sgr** package we implemented a procedure based on the multivariate latent variable framework which is called *underlying variable approach* (UVA, Muthén, 1984; Jöreskog and Sörbom, 1996). This approach assumes that the observed ordinal variables are treated as metric through assumed underlying normal variables. In particular, we assume that there exists a continuous data matrix $\mathbf{D}^*$ underlying the ordinal data matrix $\mathbf{D}$. Let $\mathbf{d}_i^*$ be the $(1 \times m)$ array of $\mathbf{D}^*$ denoting the pattern of underlying continuous values of the ith observation. It is convenient to let $\mathbf{d}_i^*$ have the multivariate standard normal distribution with density function $\phi(\mathbf{0}, \mathbf{R})$ where $\mathbf{R}$ denotes the $(m \times m)$ model correlation matrix. The connection between the ordinal variable d_{ij} and the underlying variable d_{ij}^* in $\mathbf{D}^*$ is given by

$$d_{ij} = h \qquad \text{iff} \qquad \tau_{h-1}^j < d_{ij}^* \leq \tau_h^j$$

with $h = 1, \ldots, q; i = 1, \ldots, n; j = 1, \ldots, m$ and where

$$-\infty = \tau_0^j < \tau_1^j < \tau_2^j < \ldots < \tau_{q-1}^j, \tau_q^j = +\infty,$$

are threshold parameters. Therefore, the joint probability of $\mathbf{d}_i = (h_1, \ldots, h_m)$ is given by

$$p(\mathbf{h}|\theta_M) = \int_{\tau_{h_1-1}^1}^{\tau_{h_1}^1} \cdots \int_{\tau_{h_m-1}^m}^{\tau_{h_m}^m} \phi(\mathbf{z}|\mathbf{0}, \mathbf{R}) d\mathbf{z} \tag{4}$$

with $\theta_M = (\tau, \mathbf{R})$ and $\mathbf{z} = (z_1, \ldots, z_m)$ being the parameter vector of the original data generation model and the values for the continuous variables $\mathbf{d}_i^*$, respectively.

In SGR the data generation process is obtained by first generating the continuous data $\mathbf{D}^*$ according to a model correlation matrix $\mathbf{R}$ and then by transforming it to its discrete counterpart $\mathbf{D}$ using the model thresholds τ. In the following example, we used the **sgr** function `rdatagen` to sample $n = 100$ random observations from a data generation model with two symmetrically distributed ordinal variables with five levels each and correlation value .4.

```
> library(sgr)
> require(MASS)
> require(polycor)
> set.seed(367)
> R <- matrix(c(1,.4,.4,1),2,2)
> th <- list(c(-Inf,qnorm(c(0.04,0.27,0.73,0.96)),
+              Inf),c(-Inf,qnorm(c(0.06,0.31,0.69,0.94)),Inf))
> Dx <- rdatagen(n=100,R=R,Q=c(5,5),th=th)
> Dx$data
```

In this example, the threshold values are derived from the quantiles of the standard normal distribution in such a way that the first simulated variable shows a slightly larger variance than the second simulated variable. Generally, the threshold values can be derived in two different ways. In the first case, we can use empirically based knowledge (e.g., an already existing data set) to estimate the threshold values on the basis of the observed distribution function of the levels of the discrete variable (e.g., Jöreskog and Sörbom, 1996). In the second case, some simple statistical knowledge can be used to simulate threshold values according to desired properties. For example, the normal quantiles used as corresponding threshold values can be computed using the inverse of the binomial cumulative distribution function (e.g., Jöreskog and Sörbom, 1996). In the `rdatagen` function call the parameter `Q` specifies the number of levels for each ordinal variable. To compare the model correlation matrix $\mathbf{R}$ with the sample polychoric correlation, we can use the `polychor` function in the **polycor** package (Fox, 2010)

```
> d1 <- factor(Dx$data[,1],ordered = T)
> d2 <- factor(Dx$data[,2],ordered = T)
> polychor(d1, d2, ML=TRUE, std.err=TRUE)

Polychoric Correlation, ML est. = 0.3627 (0.09832)
```

Data replacement process

To generate the fake ordinal data we used a parametrized replacement distribution based on a discrete beta kernel (Pastore and Lombardi, 2014). Some examples of replacement distributions are shown in Figure 1. Let $p_{k|h} \equiv p(k|h, \theta_F)$ be the conditional probability of replacing an original ordinal value h with the new ordinal value k. 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-).

The conditional replacement distribution can be described according to the following equation

$$p_{k|h} = \begin{cases} DG(k; a^+, b^+, \theta_F^+)\pi^+, & 1 \leq h < k \leq q \\ DG(k; a^-, b^-, \theta_F^-)\pi^-, & 1 \leq k < h \leq q \\ 1 - (\pi^+ + \pi^-), & 1 < k = h < q \\ 1 - \pi^+, & k = h = 1 \\ 1 - \pi^-, & k = h = q \end{cases} \tag{5}$$

with DG being the generalized beta distribution for discrete variables (Pastore and Lombardi, 2014). Note that in Eq. (5), the function DG is used with two different set of parameters. More precisely, in the first line the function DG models the behavior of the faking distribution for fake positive values

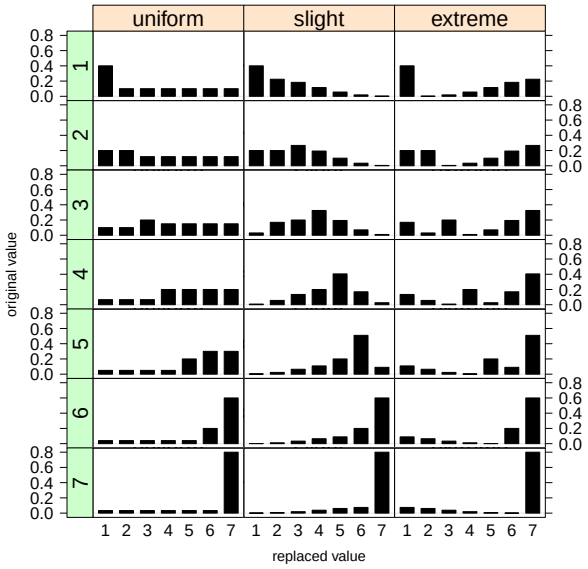

Figure 1: Three examples of conditional replacement distributions for a 7-point discrete r.v. Each column in the graphical representation corresponds to a different conditional replacement distribution (see Table 1). For each example the overall probabilities are $\pi^+ = .6$ and $\pi^- = .2$. Each row in the graphical representation corresponds to a different original 7-point discrete value h.

($k > h$) by means of the governing shape parameter $\theta_F^+ = (\gamma^+, \delta^+)$ with bounds $(a^+ = h+1, b^+ = q)$. By contrast, the second line represents the behavior of the faking distribution for fake negative values ($k < h$) modelled by the governing shape parameter $\theta_F^- = (\gamma^-, \delta^-)$ with bounds $(a^- = 1, b^- = h-1)$. Some examples of faking models with their parameters assignments are reported in Table 1 (see also Pastore and Lombardi, 2014). In general, the governing shape parameters θ_F^+ and θ_F^- must be strictly positive. In particular, if $\gamma^+ = \delta^+ = 1$, the right part of the replacement distribution reduces to a *uniform* support fake positive distribution (Fig. 1 first column). By contrast, if $1 \leq \gamma^+ < \delta^+$ (resp. $1 \leq \delta^+ < \gamma^+$), the model mimics asymmetric faking configurations corresponding to moderate positive shifts (resp. exaggerated positive shifts) in the value of the original response (Fig. 1, second and third columns). More specifically, in the *slight* positive faking configuration the chance to replace an original value h with another greater value k decreases as a function of the distance between k and h. By contrast, in the *extreme* positive faking configuration the chance to replace an original value h with another greater value k increases as a function of the distance between k and h. Unlike the asymmetric configurations (slight faking and extreme faking), the uniform support distribution ($\gamma^+ = \delta^+ = 1$) mimics a kind of random world model that can be used whenever we believe to deal with purely random fake data. This principle requires the simplest quantitative representation for the replacement process and reflects the lack of information about the distributional properties of the faking behavior. Similar configurations can be described also for the left part of the replacement distribution which represents the negative faking process [$\theta_F^- = (\gamma^-, \delta^-)$]. However, for this latter component the ordinal relation characterizing the shape parameters must be reversed (see Table 1). Finally, in the conditional replacement distribution the parameters π^+ and π^- denote the overall probability of faking positive and the overall probability of faking negative, respectively. These probabilities act as weights to rescale the discrete beta distribution DG such that $(\pi = \pi^+ + \pi^-) \leq 1$. In general, π^+ and π^- represent *a priori* or empirically based knowledge about the distribution of faking for the two components (e.g., the chance of observing a positive or negative fake observation in the data). The third, fourth, and fifth lines of Eq. (5) show the probability of non-replacement ($k = h$). Note that, if we set $\pi^+ = 0$ (resp. $\pi^- = 0$), then the replacement model boils down to a pure faking negative (resp. positive) model which corresponds to a context in which responses are exclusively subject to negative (resp. positive) faking (see fig. 2).

In the following example, we applied a pure (slight) positive faking model (see Table 1) to generate a fake data matrix **F** from the original data matrix **D**.

```
> RM <- replacement.matrix(Q=5,p=c(.5,0),fake.model="slight")
> Fx <- rdatarepl(Dx$data,RM)
46% of data replaced.
> Fx$Fx
```

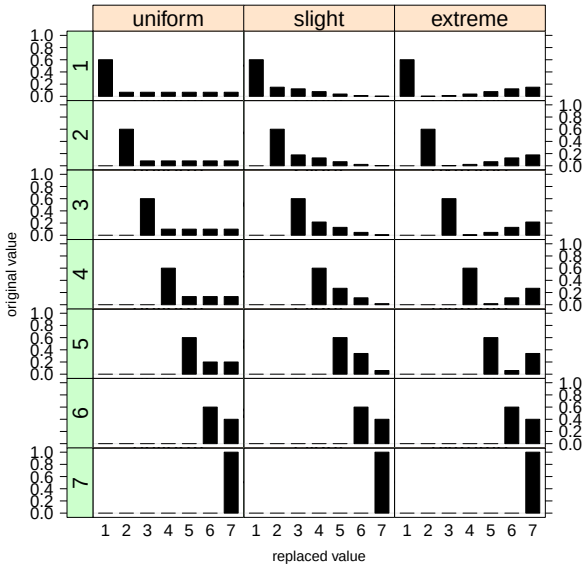

Figure 2: Three examples of conditional replacement distributions for a 7-point discrete r.v. Each column in the graphical representation corresponds to a different conditional replacement distribution. For each example the overall probabilities are $\pi^+ = .6$ and $\pi^- = .0$ (faking positive condition). Each row in the graphical representation corresponds to a different original 7-point discrete value h.

We used the `replacement.matrix` function to construct the conditional replacement probability distribution and save the result in the variable `RM` which is used as the argument of the the data replacement generation function `rdatarepl`. Note that the argument `fake.model` in the `replacement.matrix` function allows to set the options reported in Table 1. However, all the model parameters can be set manually by the user to any array of consistent values if so desired. For example, an equivalent syntax would have been

```
> RM <- replacement.matrix(Q=5,p=c(.5,0),gam=c(1.5,0),del=c(4,0))
```

We can evaluate the impact of positive faking on the new fake data matrix by comparing the frequencies of the ordinal categories in **D** and **F**. For example, for the first ordinal variable we have

```
> table(Dx$data[,1])

 1  2  3  4  5
 5 29 40 24  2
> table(Fx$Fx[,1])

 1  2  3  4  5
 2 17 36 31 14
```

which shows how the positive faking has shifted the values of the first ordinal variable towards larger ones. In a similar way, we could also evaluate the impact of faking on the sample polychoric correlation matrix of **F**.

Model	γ^+	γ^-	δ^+	δ^-
uniform	1	1	1	1
slight	1.5	4	4	1.5
extreme	4	1.5	1.5	4

Table 1: Examples of default parameters assignments for some relevant faking models (Pastore and Lombardi, 2014).

Illustrative examples

By way of illustration we consider three simple SGR examples. The first is for the evaluation of a correlational analysis computed on five-point rating data. This example is hypothetical and serves to introduce the main features and functions implemented in the **sgr** package. The second application considers real data about illicit drug use among young people aged 14 to 27. This second example shows how to model directional faking hypotheses (e.g., faking good or faking bad). It is also important because illustrates how the replacement functions can be applied to dichotomous data. Finally, the third application extends the second example by analyzing a new set of data about cannabis consumption in young people using log-linear models for ordinal data.

Example 1

We begin with a simple SGR analysis about a hypothetical observed difference ($\hat{\Delta} = \hat{\rho}_1 - \hat{\rho}_2 = .3$) between two ordinal correlations computed on two five-point rating variables X and Y for the groups of subjects, G_1 ($n_1 = 50$) and G_2 ($n_2 = 50$). For example, in a risky sexual behaviors scenario the rating variables X and Y can represent, in two groups of young adults (females and males), the self-report attitude to contraceptive use during a sexual intercourse and the declared number of sexual partners in the last three months, respectively. Normally, an effect size of .3 denotes a relevant difference between two correlations. However, how sensitive may this result be to possible fake data? Is this effect still observed under one or more scenarios of faking? In this example, we are interested in testing whether the observed correlation difference can still be consistent with a true generative model reflecting an identical moderate correlation $\rho_1 = \rho_2 = .25$ for the two groups. Moreover, we also assume a perturbation process represented by two distinct uniform faking models: $\pi_1^+ = .2$ and $\pi_1^- = .1$ for G_1, and $\pi_2^+ = .3$ and $\pi_2^- = .2$ for G_2. We can easily reformulate this example using a Fisher significance testing (Lehmann, 1993). More precisely, we can construct the corresponding hypothesis

$$\begin{aligned} H \quad : \quad & \rho_1 = \rho_2 = .25\ (\Delta = 0), \\ & \pi_1^+ = \pi_2^- = .2, \pi_1^- = .1, \pi_2^+ = .3, \\ & \gamma_s^+ = \gamma_s^- = \delta_s^+ = \delta_s^- = 1, \quad s = 1, 2 \end{aligned}$$

and examine whether or not the observed correlation difference $\hat{\Delta}$ is consistent with H. In particular, we are interested in the p-value

$$Pr[\Delta > \hat{\Delta}|H].$$

The code below illustrates the SGR analysis

```
> require(polycor)
> set.seed(367)
> obs.stat <- .3; mc.stat <- NULL
> Rmc <- matrix(c(1,.25,.25,1),2)
> PM <- matrix(c(rep(1,100),rep(2,100)),ncol=2,byrow=TRUE)
> Pparm <- list(p=matrix(c(.2,.3,.1,.2),2),gam=matrix(1,2,2),del=matrix(1,2,2))
> for (b in 1:1000) {
+   mcD <- rdatagen(n=100,R=Rmc,Q=5)$data
+   Fx <- partition.replacement(mcD,PM,Pparm=Pparm)
+   for (j in 1:ncol(Fx)) {
+     Fx[,j] <- ordered(Fx[,j])
+   }
+   mcpc1 <- hetcor(Fx[1:50,])$correlations[1,2]
+   mcpc2 <- hetcor(Fx[51:100,])$correlations[1,2]
+   Delta <- mcpc1-mcpc2
+   mc.stat <- c(mc.stat,Delta)
+ }

> hist(mc.stat)
> sum(mc.stat>=obs.stat)/1000
[1] 0.226
```

An empirical p-value can be computed by a Monte Carlo experiment. In our example, the test procedure $\Delta^* = \hat{\rho}_1^* - \hat{\rho}_2^*$ is replicated 1000 times under the condition of the hypothesis. Next, the approximate p-value is computed as the proportion of the simulated Δ^* values which are larger than the observed correlation difference .3. More precisely, for each replicate $b = 1, \ldots, 1000$, we first generate a 100×2 ordinal data matrix `mcD` according to the generative model with correlation

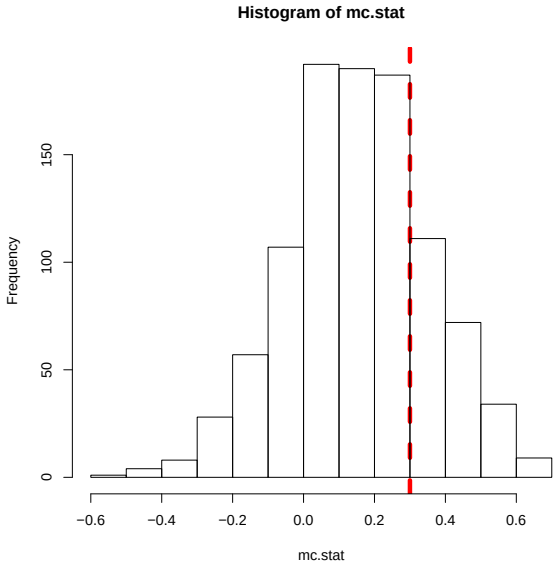

Figure 3: Distribution of the test procedure Δ^* under H.

matrix `Rmc`. This matrix contains two symmetrically distributed ordinal variables (default value[1] in the `rdatagen` function.) Next, the ordinal matrix is transformed according to the faking models. In particular, the function `partition.replacement` allows to set different replacement distributions for the two groups of subjects and returns the perturbed data matrix. This function has three main arguments: `Dx=mcD`, the data frame or matrix to be replaced; `PM`, the partition matrix to cluster the observations into the distinct groups; `Pparm`, the list of replacements parameters for each of the different faking models. Note that the partition matrix must have the same dimension as the matrix to be replaced and a numeric code for each distinct cluster (group) in the partition. If a cell of the partition matrix contains 0, then the corresponding cell value in the original data matrix is not modified (*no replacement* condition is applied). In our example, `Pparm` is a list containing three elements. Each element is a 2 (number of groups) $\times$ 2 (faking positive vs faking negative) matrix. So for example, `p[1,1]` and `p[1,2]` denote the overall faking positive probabilities for G_1 and G_2, respectively. Similarly, `gam[1,1]` (resp. `gam[2,1]`) indicates the first shape parameter for the faking positive (resp. faking negative) model in group G_1. The same figure follows for the second shape parameter `del`. Figure 3 shows the distribution of the test procedure under H (approximate p-value $= .226$). According to the distribution of the test procedure the observed correlation difference $\hat{\Delta}$ seems consistent with the hypothesis of faking.

Example 2

Table 2 refers to a real prospective study about illicit drug use among young people aged 14-27 (Pastore et al., 2007). In particular, we evaluated the relationship between age (dichotomized into two categories: adults, > 17, and minors) and ecstasy drug consumption. We expected that each individual would deliberately answer the question either honestly or fraudulently depending on her/his beliefs and intentions which, in turn, could be influenced by the context. How can the researcher evaluate the impact of possible fake answers when trying to provide an overall picture of the investigated phenomena? Although the example is specific, a similar problem may occur in a variety of situations about stigmatizing characteristics (e.g., habitual gambling, experience of induced abortion, tax evasion, rash driving, risky sexual behavior).

The result of a log linear model for independence for the two-way table showed a significant likelihood-ratio chi-squared statistic ($G^2_{(1)} = 5.29, p < .05$). Hence the independence assumption was rejected. By a quick inspection of the counts shown in table 2 we can easily recognize that only 29% of adults answered affirmatively to the question. By contrast, more than 50% of minors replied affirmatively. Therefore, we suspected that the adults might have shown a larger level of

[1]The default setting requires that the quantiles are computed using the inverse of the binomial cumulative distribution (see for example, Jöreskog and Sörbom, 1996).

	drug	
	yes (1)	no (2)
adults (1)	10	25
minors (2)	32	29

Table 2: Observed frequencies for testing independence of age and drug use for the question: Have you ever made use of ecstasy?

social desirability (Paulhus, 1984) as compared to the minors. This might have artificially boosted the observed difference between the two groups. To test this hypothesis we performed a new SGR analysis on the two-way table by assuming a) a generative model implementing the independence assumption with marginal probability $\Pr(\text{yes}) = .75$ b) a fake good model for the variable drug consumption. 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). 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. In this application the scenario corresponds to a context in which all fakers respond negatively to the question. Finally, we also assumed two distinct levels of faking for the two groups: $\pi_1^+ = .8$ for the adults and $\pi_2^+ = .4$ for the minors. We reformulated the problem within a pure significance test setting:

$$\begin{aligned} H \quad : \quad & G^2 = 0 \text{ (independence assumption)}, \\ & \Pr(\text{yes}) = .75, \\ & \pi_1^+ = .8,\ \pi_2^+ = .4,\ \pi_1^- = \pi_2^- = .0, \\ & \gamma_s^+ = \delta_s^+ = 1,\ \gamma_s^- = \delta_s^- = 0, \quad s = 1,2 \end{aligned}$$

The following code illustrates the SGR analysis

```
> require(MASS)
> set.seed(367)
> data(smokers)
> ecstasy.table <- table(smokers$drug,smokers$age,dnn=c("drug","age"))
> obs.lrt <- loglm(~drug+age,data=ecstasy.table)$lrt
>
> PM <- matrix(0,nrow(smokers),2)
> PM[smokers$age==1,2] <- 1
> PM[smokers$age==2,2] <- 2
> Pparm <- list(p=matrix(c(.8,.4,0,0),2),gam=matrix(c(1,1,0,0),2),
+               del=matrix(c(1,1,0,0),2))
> mc.lrt <- NULL
> for (b in 1:1) {
+   smokers$simdrug <- rdatagen(nrow(smokers),R=matrix(1),Q=2,
+                               probs=list(c(.75,.25)))$data
+   Fx <- partition.replacement(smokers[,c("age","simdrug")],PM,Pparm=Pparm)
+   mc.lrt <- c(mc.lrt,loglm(~simdrug+age,data=table(Fx$simdrug,Fx$age,
+                               dnn=c("simdrug","age")))$lrt)
+ }

> hist(mc.lrt)
> sum(mc.lrt>=obs.lrt)/1000
[1] 0.812
```

Note that for dichotomous variables ($q = 2$) all faking positive models reduce to the following uniform conditional replacement distribution (Pastore and Lombardi, 2014).

$$p_{k|h} = \begin{cases} 1, & h = k = 2 \\ \pi^+, & h = 1, k = 2 \\ 1 - \pi^+, & h = k = 1 \\ 0, & h = 2, k = 1 \end{cases} \tag{6}$$

Figure 4 shows the distribution of the test procedure under the hypothesis (approximate p-value $= .812$). According to the approximate G^2 distribution the observed likelihood ratio seems consistent with the hypothesis of faking.

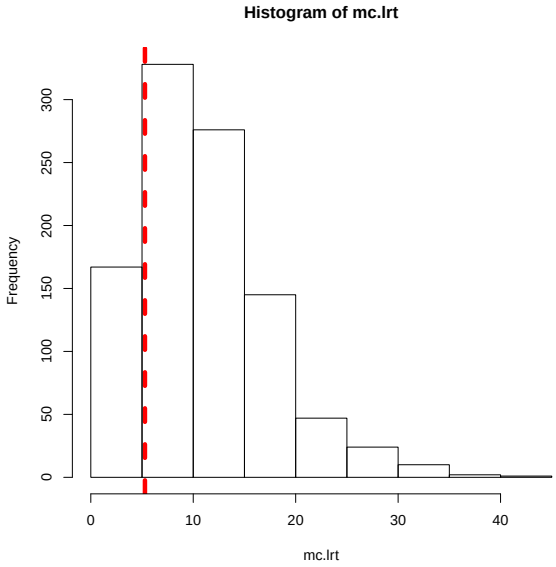

Figure 4: Reproduced distribution for the test statistic G^2 under H.

Example 3

In this application we extend the results reported in the second example by analyzing a new set of ordinal data about illicit drug use among young people (see table 3). This new two-way table relates an independent categorical variable, *age*, minors (< 18 years old) vs adults, to a dependent ordinal variable, *cannabis consumption*. In particular, the dependent variable uses a four-point ordinal scale ranging from *never* (1) to *often* (4) (with intermediate levels being *once* (2) and *some times* (3), respectively). When response categories are ordered, logit models can directly incorporate the ordering (Agresti, 1990). In general, this results in model representations having simpler interpretations than ordinary multicategory logit models at least when the proportional odds model holds.

	cannabis			
	(1)	(2)	(3)	(4)
adults (1)	20	5	7	0
minors (2)	27	5	18	10

Table 3: Observed frequencies for testing independence of age and drug use.

The following code illustrates the results of applying an ordered logistic model to the data represented in table 3. For the analysis we used the function `polr` in the **MASS** package (Venables and Ripley, 2002) that allows to fit a logistic or probit regression model to an ordered factor response.

```
> Y <- data.frame(list(age=gl(2,4),response=gl(4,1,8,ordered=TRUE),
+           counts=c(20,5,7,0,27,5,18,10)))
> fit0 <- polr(response~1,data=Y,weight=counts)
> fit1 <- polr(response~age,data=Y,weight=counts)
> lrt.obs <- -2*(logLik(fit0)-logLik(fit1))
```

The likelihood ratio statistic $\Lambda = -2(L_0 - L_1)$ for the observed sample showed a significant result ($\Lambda_{(3)} = 5.22, p < .05$). Hence, the independence assumption was rejected in the logit model. About the model of faking also in this application we expected that individuals' responses were strictly subject to faking good manipulations. However, unlike the previous example, this time we speculated that only the group of adults showed a social desirability bias whereas the minors' responses were assumed not to be fake dependent ($\pi_2 = \pi_2^+ + \pi_2^- = 0$). In particular, we supposed that the adults were showing a moderate level of faking good (10%) and that their responses were characterized by a slight faking behavior (see table 1). Note that because of the meaning of the categories of the ordinal scale for cannabis consumption, in this application the faking good manipulations are modelled by means of the fake negative parameters (π^-). Finally, for the data generation process we constructed a

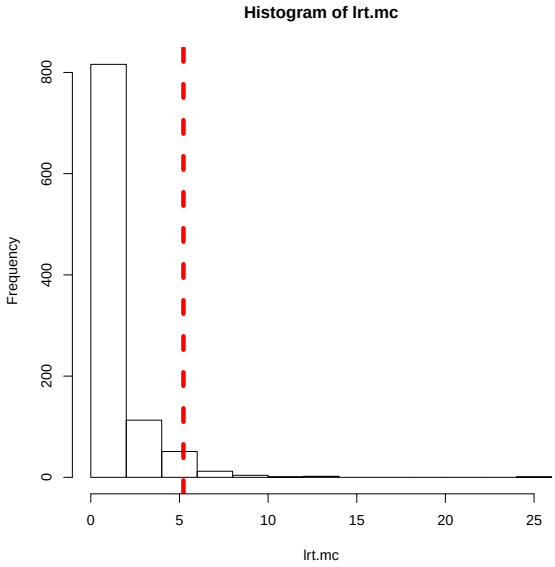

Figure 5: Reproduced distribution for the likelihood ratio statistic under H.

generative model under the assumption of no relation between age and cannabis consumption ($\Lambda = 0$) and with true response proportions being equal to the empirical response proportions for the group of minors. We can collect all the information in the following hypothesis:

$$\begin{aligned} H \quad : \quad & \Lambda = 0 \text{ (independence assumption)}, \\ & \Pr(0) = .45, \Pr(1) = .08, \\ & \Pr(2) = .30, \Pr(3) = .17, \\ & \pi_1^- = .1,\ \pi_2^- = \pi_1^+ = \pi_2^+ = .0, \\ & \gamma_1^- = 1.5,\ \delta_1^- = 4, \\ & \text{all the other shape parameters are set to } 0 \end{aligned}$$

The following code illustrates the SGR analysis

```
> set.seed(367)
> Z <- na.omit(smokers[,c("age","druguse")])
> PM <- matrix(0,nrow(Z),ncol(Z))
> PM[Z$age==1,2] <- 1
> lrt.mc <- NULL
> for (b in 1:1) {
+   Z$simdrug <- rdatagen(nrow(Z),R=matrix(1),Q=4,
+                         probs=list(c(27,5,18,10)/60))$data
+   Dx <- Z[,-2]
+   Fx <- partition.replacement(Dx,PM,p=matrix(c(0,.1),1),fake.model="slight")
+   Tmc <- table(Fx$age,Fx$simdrug)
+   Ymc <- data.frame(list(age=gl(2,4),response=gl(4,1,8,ordered=TRUE),
+           counts=c(Tmc[1,],Tmc[2,])))
+   fit0 <- polr(response~1,data=Ymc,weight=counts)
+   fit1 <- polr(response~age,data=Ymc,weight=counts)
+   lrt.mc <- c(lrt.mc,-2*(logLik(fit0)-logLik(fit1)))
+ }

> sum(lrt.mc>=lrt.obs)/1000
[1] 0.039
```

Figure 5 shows the distribution of the test procedure under the hypothesis. This time the observed likelihood ratio statistic seems not consistent with H (approximate p-value .039). In substantive terms, the observed association between age and cannabis consumption cannot be explained by an independent generative model and slight faking good manipulations for the adult group.

Example 3 (continued)

In this section we provide a full exploratory SGR analysis for the data presented in table 3. In particular, we show how it is possible to vary the parameters (γ_1^-, δ_1^-) of the fake negative distribution and evaluate how these changes effect the results of the approximate p-value. Figure 6 shows the contour plot of the approximate p-value as a function of different levels for the shape parameters γ_1^- and δ_1^- in the group of adults. More specifically, the value of parameter γ_1^- was varied at 21 distinct levels from 0.5 to 5.5 (by a 0.25 step). The same set of values was also applied for the second shape parameter δ_1^-. In this application we also changed the overall probability of faking good π^- by replacing the original value 0.1 (used in the previous example) with the new value 0.2. By contrast, all the other parameters' values were left unchanged in the SGR simulation by keeping the same values reported in the previous analysis. The results show how the value of the observed statistic, $\Lambda = 5.22$, is consistent with an independent true model ($\Lambda = 0$) that has been corrupted by a moderate amount of faking good perturbation (20%), and which is also characterized by an extreme faking pattern in the replacement distribution. This is evident from a quick inspection of figure 6 where the parameters assignments that resulted consistent with the earlier faking hypothesis are restricted to the left portion of the main diagonal ($\gamma_1^- < \delta_1^-$) in the graphical representation. By contrast, the parameters assignments corresponding to the right portion of the main diagonal ($\gamma_1^- > \delta_1^-$) are not consistent with the hypothesis. Note that these latter values represent slight faking configurations in the replacement distribution. In sum, the results are in line with a moderate faking good process which is characterized by a general property of extremeness in the way the original true values are replaced with the fake ones in the replacement distribution. That is to say, in general the chance to replace an original true value with another lower value seem to increase as a function of the distance between two values.

In what follows we present a short code example that the reader may easily manipulate to set the desired values for the parameters in the simulation study (shape parameters, overall probabilities of faking, number of runs in the SGR simulations). Note that in this exploratory setting the overall time required to complete the SGR simulation may widely vary according to the complexity (e.g., number of different values for the parameters) of the simulation design.

```
> data(smokers)
> Z <- na.omit(smokers[,c("age","druguse")])
>
> fit0 <- polr(ordered(druguse)~1,data=Z)
> fit1 <- polr(ordered(druguse)~age,data=Z)
> lrt.obs <- -2*(logLik(fit0)-logLik(fit1)) # observed LRT
>
> ### SGR algorithm
```

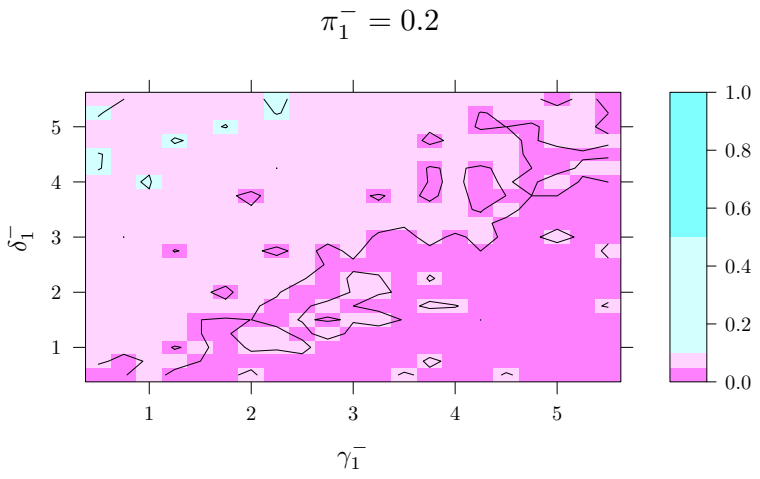

Figure 6: Contour plot for the approximate p-value as a function of shape parameter γ_1^- and shape parameter δ_1^-. In this example the overall probability of faking good for the adult group was set equal to $\pi_1^- = 0.2$. Note that the points represented to the left of the main diagonal correspond to extreme faking conditions, whereas the points represented to the right of the main diagonal correspond to slight faking conditions. The total number of runs in the SGR simulation analysis was equal to 500.

```
> PI <- .2; B <- 10  # for real simulations set B at least 500
> lrt.mc <- ga.mc <- de.mc <- p.mc <- NULL
> PM <- matrix(0,nrow(Z),ncol(Z)) # partition matrix
> PM[Z$age==1,2] <- 1
>
> for (GA in seq(.5,5.5,.5)) {
+   for (DE in seq(.5,5.5,.5)) {
+
+     Pparm <- list(p=matrix(c(0,PI),1),gam=matrix(c(0,GA),1),del=matrix(c(0,DE),1))
+
+     for (b in 1:B) {
+       Z$simdrug <- rdatagen(nrow(Z),R=matrix(1),Q=4,
+                             probs=list(c(27,5,18,10)/60))$data
+       Dx <- Z[,-2]
+       Fx <- partition.replacement(Dx,PM,Pparm=Pparm)
+
+       Tmc <- table(Fx$age,Fx$simdrug)
+       Ymc <- data.frame(list(age=gl(2,ncol(Tmc)),response=gl(ncol(Tmc),1,
+               ordered=TRUE,labels=colnames(Tmc)),counts=c(Tmc[1,],Tmc[2,])))
+
+       fit0 <- polr(response~1,data=Ymc,weight=counts)
+       fit1 <- polr(response~age,data=Ymc,weight=counts)
+       statval <- -2*(logLik(fit0)-logLik(fit1))
+       lrt.mc <- c(lrt.mc,statval)
+
+       ga.mc <- c(ga.mc,GA); de.mc <- c(de.mc,DE)
+       p.mc <- c(p.mc,ifelse(statval>lrt.obs,1,0))
+     }
+   }
+ }

> LRT <- data.frame(list(gam=ga.mc,del=de.mc,lrt=lrt.mc))
> aggregate(p.mc,list(gam=LRT$gam,del=LRT$del),mean)
```

Summary, limitations, and future works

This paper illustrated the usage of a new R package, **sgr**, for simulating and analyzing ordinal fake data. As far as we know, **sgr** is the first statistical package that is devoted to the analysis of fake data. Overall, the essential characteristic of this approach is its explicit use of mathematical models and appropriate probability distributions for quantifying uncertainty in inferences based on possible fake data. Moreover, it involves the derivation of new statistical results as well as the evaluation of the implications of such new results: Are the substantive conclusions reasonable? How sensitive are the results to the modeling assumptions about the process of faking? In sum, SGR takes an interpretation perspective by incorporating in a global model all the available information about the process of faking. In this contribution we illustrated the use of **sgr** on three simple scenarios of faking. More complex examples of SGR applications can be found in Lombardi and Pastore (2012) and Pastore and Lombardi (2014).

As with many Monte Carlo-type approaches, also SGR involves simplifying assumptions that may result in lower external validity. For example, one relevant limitation regards the assumption that restricts the conditional replacement distribution to satisfy the CI property. Unfortunately, this restriction clearly limits the range of empirical faking processes that can be mimicked by the current SGR simulation procedure. In particular, because the replacement distribution under the CI assumption acts as a perturbation process for the original data, the resulting new fake data sets will in general show covariance patterns that are (on average) weaker than the ones observed for the original uncorrupted data. In general, this may not be a serious problem as different studies have shown that self-report measures under faking motivating conditions tend to have smaller variances and lower reliability (covariance) estimates than those observed for measures collected under uncorrupted conditions (Ellingson et al., 2001; Eysenck et al., 1974; Hesketh et al., 2004; Topping and O'Gorman, 1997). However, opposite results have also been observed where simple fake good instructions tend to increase the intercorrelations between the manipulated or faked items (Ellingson et al., 1999; Galic et al., 2012; Pauls and Crost, 2005; Zickar and Robie, 1999; Ziegler and Buehner, 2009). Therefore, although encouraging, the promise of this approach should be examined across more varied conditions. We acknowledge that more work still needs to be done. We are in the process of extending **sgr** to

include new replacement distributions other than the ones presented in this article which will allow to modulate different levels of correlational patterns in the simulated fake data.

Bibliography

A. Agresti. *Categorical Data Analysis*. Wiley, New York, NY, 1990. [p9]

R. Beaber, A. Marston, J. Michelli, and M. Mills. A brief test for measuring malingering in schizophrenic individuals. *American Journal of Psychiatry*, (142):1478–1481, 1985. [p1]

J. Cohen. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988. [p1]

V. Crawford. Lying for strategic advantage: Rational and boundedly rational misrepresantion of intentions. *The American Economic Review*, (93):133–149, 2003. [p1]

J. Ellingson, P. Sackett, and L. Hough. Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal Of Applied Psychology*, 84(2):155–166, APR 1999. ISSN 0021-9010. doi: {10.1037/0021-9010.84.2.155}. [p12]

J. E. Ellingson, D. B. Smith, and P. R. Sackett. Investigating the influence of social desirability on personality factor structure. *Journal Of Applied Psychology*, 86:122–133, 2001. [p12]

S. B. Eysenck, H. J. Eysenck, and L. Shaw. The modification of personality and lie scale scores by special 'honesty' instructions. *The British journal of social and clinical psychology*, 13:41–50, 1974. [p12]

P. F. Foster, E. Fabrega, S. Karademir, N. H. Sankary, D. Mital, and J. W. Williams. Prediction of abstinence from ethanol in alcoholic recipients following liver transplantation. *Hepatology*, 25:1469–1477, 1997. [p1]

J. Fox. *polycor: Polychoric and Polyserial Correlations*, 2010. URL http://CRAN.R-project.org/package=polycor. R package version 0.7-8. [p3]

Z. Galic, Z. Jerneic, and M. P. Kovacic. Do Applicants Fake Their Personality Questionnaire Responses and How Successful are Their Attempts? A Case of Military Pilot Cadet Selection. *International Journal Of Selection And Assessment*, 20(2):229–241, JUN 2012. ISSN 0965-075X. doi: {10.1111/j.1468-2389.2012.00595.x}. [p12]

N. S. Gray, M. J. MacCulloch, J. Smith, M. Morris, and R. J. Snowden. Forensic psychology: Violence viewed by psychopathic murderers. *Nature*, (423):497–498, 2003. [p1]

J. Helton, J. Johnson, C. Salaberry, and C. Storlie. Survey of sampling based methods for uncertainty and sensitivity analysis. *Reliability Engineering and System Safety*, (91): 1175–1209, 2006. [p1]

B. Hesketh, B. Griffin, and D. Grayson. Applicants faking good: evidence of item bias in the neo pi-r. *Personality And Individual Differences*, 36:1545–1558, 2004. [p12]

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, 2006. [p1]

K. Jöreskog and D. Sörbom. *PRELIS 2: User's Reference Guide*. Scientific Software International, Inc., Lincolnwood, IL, 1996. [p2, 3, 7]

E. L. Lehmann. The fisher, neyman-pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, (424):1242–1249, 1993. [p6]

L. Lombardi and M. Pastore. Sensitivity of fit indices to fake perturbation of ordinal data: A sample by replacement approach. *Multivariate Behavioral Research*, (47):519–546, 2012. [p1, 2, 12]

E. Marshall. How prevalent is fraud? that's a million-dollar question. *Science*, (290): 1662–1663, 2000. [p1]

I. Moustaki and M. Knott. Generalized latent trait models. *Psychometrika*, 65(3):391–411, SEP 2000. ISSN 0033-3123. doi: {10.1007/BF02296153}. [p2]

B. Muthén. A general structural equation model with dichotomous, ordered categorical and continuous latent variables indicators. *Psychometrika*, (49):115–132, 1984. [p2]

M. Pastore and L. Lombardi. The impact of faking on cronbach's alpha for dichotomous and ordered rating scores. *Quality & Quantity*, 48:1191–1211, 2014. doi: 10.1007/s11135-013-9829-1. [p2, 3, 4, 5, 8, 12]

M. Pastore, L. Lombardi, and F. Mereu. Effects of malingering in self-report measures: A scenario analysis approach. In C. H. Skiadas, editor, *Recent Advances in Stochastic Modeling and Data Analysis*, pages 483–491. World Scientific Publishing, 2007. [p7]

D. L. Paulhus. Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, (46):598–609, 1984. [p8]

C. Pauls and N. Crost. Effects of different instructional sets on the construct validity of the NEO-PI-R. *Personality And Individual Differences*, 39(2):297–308, JUL 2005. ISSN 0191-8869. doi: {10.1016/j.paid.2005.01.003}. [p12]

F. Samejima. *Estimation of Latent Ability Using a Response Pattern of Graded Scores*. Psychometric Monograph No. 17. Richmond, VA: Psychometric Society, 1969. [p2]

G. D. Topping and J. O'Gorman. Effects of faking set on validity of the neo-ffi. *Personality and Individual Differences*, 23(1):117–124, 1997. ISSN 0191-8869. doi: http://dx.doi.org/10.1016/S0191-8869(97)00006-8. URL http://www.sciencedirect.com/science/article/pii/S0191886997000068. [p12]

S. Van der Geest and S. Sarkodie. The fake patient: A research experiment in a ghanaian hospital. *Social Science & Medicine*, (47):107–120, 1998. [p1]

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0. [p9]

M. J. Zickar and C. Robie. Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology*, 84:551–563, 1999. [p12]

M. Ziegler and M. Buehner. Modeling Socially Desirable Responding and Its Effects. *Educational And Psychological Measurement*, 69(4):548–565, AUG 2009. ISSN 0013-1644. doi: {10.1177/0013164408324469}. [p12]

Luigi Lombardi
Department of Psychology and Cognitive Science, University of Trento
Corso Bettini, 31, 38068 Rovereto, TN
Italy
luigi.lombardi@unitn.it

Massimiliano Pastore
Department of Developmental and Social Psychology, University of Padova
Via Venezia, 8, 35131 Padua
Italy
massimiliano.pastore@unipd.it