Some regression procedures for randomized response data

van der Heijden, Peter G.M.
Utrecht University, Department of Methodology and Statistics
P.O.Box 80.140
3508 TC Utrecht, the Netherlands
E-mail: p.g.m.vanderheijden@uu.nl

Frank, Laurence
Utrecht University, Department of Methodology and Statistics
E-mail: l.e.frank@uu.nl

Cruyff, Maarten
Utrecht University, Department of Methodology and Statistics
E-mail: m.cruyff@uu.nl

Böckenholt, Ulf
North Western University, Kellogg School of Management
2001 Sheridan Rd
Evanston, IL 60208, USA
E-mail: u-bockenholt@kellogg.northwestern.edu

1. Introduction

Is it possible to measure sensitive behavior such as noncompliance with rules and regulations that govern public life using surveys? Because it is well–known that questions about compliance behavior with rules and regulations may not yield truthful responses, the randomized response (RR) method has been proposed as a survey tool to get more honest answers to sensitive questions (Warner, 1965). In the original RR approach, respondents were provided with two statements, A and B, with statement A being the complement of statement B. For example, statement A is 'I used hard drugs last year' and statement B is 'I did not use hard drugs last year'. A randomizing device, for instance, in the form of a pair of dice determines whether statement A or B is to be answered. The interviewer records the answer *yes* or *no* without knowing the outcome of the randomizing response device. Thus the interviewee's privacy is protected but it is still possible to estimate the probability that the sensitive question (A and not-B) is answered positively.

Recent meta–analyses have shown that RR methods can outperform more direct ways of asking sensitive questions (Lensvelt, Hox, van der Heijden and Maas, 2005). Importantly, the relative improvements in the validity increased with the sensitivity of the topic under investigation.

This paper will show recent developments of in the analysis of RR data, focusing on regression models relating the sensitive question(s) measured with RR to explanatory variables. Thus we show the developments in this field since pioneering work by Maddela (1983) and Scheers and Dayton (1988), who showed how this relation can be studied by adjusting the logistic regression model.

In this manuscript the examples stem from surveys that we conducted for the Dutch government on noncompliance in social welfare. In the Netherlands Dutch employees must be insured under the Disability Insurance Act, the Unemployment Insurance Act, and the Welfare Insurance Act. Under each of these acts, a (previously) employed person is eligible for financial benefits provided certain conditions are met. For details on the design of the 2002 study we refer to Lensvelt-Mulders, van der Heijden, Laudy and van Gils (2006).

Most of the examples focus on six RR questions, four of which are health– and the remaining

two are work–related. The health questions are:

1. Has a doctor or specialist ever told you that the symptoms your disability classification is based upon have decreased without your informing the Department of Social Services of this change?
2. At a Social Services check–up, have you ever acted as if you were sicker or less able to work than you actually are?
3. Have you yourself ever noticed an improvement in the symptoms causing your disability, for example in your present job, in volunteer work or the chores you do at home, without informing the Department of Social Services of this change?
4. For periods of any length at all, do you ever feel stronger and healthier and able to work more hours without informing the Department of Social Services of this change?

The work–related questions are:

1. Have you recently done any small jobs for or via friends or acquaintances, for instance in the past year, or done any work for payments of any size without reporting it to the Department of Social Services? (This only pertains to monetary payments.)
2. Have you ever in the past 12 months had a job or worked for an employment agency in addition to your disability/unemplyment/welfare benefit without informing the Department of Social Services?
3. Have you worked off the books in the past 12 months in addition to your disability/ unemployment/welfare benefit?

The remainder of the manuscript is structured as follows. In Section 2 we discuss the analysis of RR data without explanatory variables. In Section 3 we discuss work on logistic regression adjusted for RR data. In Section 4 we describe extensions to the situation of multivariate RR data. Section 5 discusses how regression approaches are to be adjusted in the light of new models that adjust for the fact that part of the sample is not following the RR design laid out by the researcher.

2. Univariate and multivariate RR data, no explanatory variables

This section discusses the analysis of univariate and multivariate RR data, where *no* explanatory variables are involved. We start off using the forced response design (Boruch, 1971) as an example of an RR design.

Assume that the sensitive question asks for a *yes* or *no* answer. The forced response design is as follows. After the sensitive question is asked, the respondent throws two dice and keeps the outcome hidden from the interviewer. If the outcome is 2, 3 or 4, the respondent answers *yes*. If the outcome is 5, 6, 7, 8, 9 or 10, the respondent answers according to the truth. If the outcome is 11 or 12, the respondent answers *no*.

Let Y be the latent binary RR variable that denotes the true status on the sensitive item, and let Y^* be the observed RR variable that denotes the observed answer on the randomized sensitive question, with *yes* $= 1$ and *no* $= 2$. Then

$$\begin{aligned} P(Y^* = 1) &= P(Y^* = 1|Y = 2)P(Y = 2) + P(Y^* = 1|Y = 1)P(Y = 1) \\ &= 1/6 + 3/4P(Y = 1). \end{aligned} \tag{1}$$

If we write $P(Y^* = 1) = c + dP(Y = 1)$, other designs can be described in the same way (compare Böckenholt and van der Heijden, 2007).

Following Chaudhuri and Mukerjee (1988) we simplify notation by using matrix notation. We do this for the general situation that Y^* as well as Y have K categories, indexed by j, k; this will turn out to be useful in the situation of multivariate RR data, i.e. the situation that more than one sensitive question is asked using an RR design. We collect the probabilities $P(Y^* = j) = \pi_j$ referring to the observed variable Y^* in a vector $\boldsymbol{\pi^*} = (\pi_1^*, ..., \pi_K^*)^t$, and we collect the probabilities $P(Y = k) = \pi_k$ referring to the latent variable Y in a vector $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)^t$. We collect the randomizing probabilities given by the RR design in a matrix $\boldsymbol{P}$ with elements $p_{jk} = P(Y^* = j|Y = k)$. Then the general form of RR designs can be written as (Chaudhuri and Mukerjee, 1988; van den Hout and van der Heijden, 2002):

$$\boldsymbol{\pi^*} = \boldsymbol{P\pi}, \tag{2}$$

As a first example, consider the forced response design. Here

$$\boldsymbol{P}_{FR} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 11/12 & 2/12 \\ 1/12 & 10/12 \end{pmatrix}. \tag{3}$$

As a second example, let there be two RR variables, leading to a first couple (Y_1^*, Y_1) with matrix $\boldsymbol{P_1}$, and to a second couple (Y_2^*, Y_2) with matrix $\boldsymbol{P_2}$. Then the vector $\boldsymbol{\pi^*} = (\pi_{11}^*, \pi_{12}^*, \pi_{21}^*, \pi_{22}^*)^t$, the vector $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})^t$ and $\boldsymbol{P} = \boldsymbol{P_1} \otimes \boldsymbol{P_2}$, where $\otimes$ is the Kronecker product.

The estimation of equation (2) is straightforward when the parameters are in the interior of the parameter space. If we use sample proportions of *yes* and *no* answers as estimates of $\hat{\boldsymbol{\pi}}^*$, then

$$\hat{\boldsymbol{\pi}} = \boldsymbol{P^{-1}}\hat{\boldsymbol{\pi}}^* \tag{4}$$

yields the unbiased moment estimates for $\boldsymbol{\pi}$. If the parameters are in the interior of the parameter space, then the unbiased moment estimates are equal to the maximum likelihood estimates (MLEs) (compare van den Hout and van der Heijden, 2002).

Parameters are not necessarily in the interior of the parameter space. An example where they are not in the interior of the parameter space can be derived from equation (1), when $P(Y^* = 1) < 1/6$. In other words, the observed proportions are below chance level. If this happens, the moment estimate for $P(Y = 1)$ will be negative and the moment estimate for $P(Y = 2)$ will be larger than one. As MLEs cannot be negative, they will be found at the boundary of the parameter space (i.e. it is 0).

MLEs can always be found by maximizing the kernel of the multinomial loglikelihood. Let $\boldsymbol{n} = (n_1, ..., n_K)^t$ be the vector of observed frequencies related to the probabilities for the observed response Y^* and let u be a unit vector of length K, then the kernel of the loglikehood is

$$\ell = \boldsymbol{u}^t (\boldsymbol{n} \log \boldsymbol{\pi^*}) = \boldsymbol{u}^t (\boldsymbol{n} \log \boldsymbol{P\pi}). \tag{5}$$

Maximizing ℓ over the parameters $\boldsymbol{\pi}$ can be done using an Expectation-Maximization algorithm, or by maximizing the likelihood directly (compare van den Hout and van der Heijden, 2002). We note that, when only one sensitive question is involved and the two moment estimates are outside the parameter space, the two MLEs simply end up on the boundary as 0 and 1. When more than one sensitive question is involved and one or more moment estimates are outside the parameter estimates, then, in general, the MLEs cannot be derived analytically and L has to be maximized using iterative methods.

Example In the 2002 survey on regulatory non–compliance w.r.t. disability benefit the six RR questions introduced in Section 1 were asked. The sample size was 1,760. Point estimates with 95 percent bootstrap confidence intervals are, for health item 1 it was .03 (.004 – .052), for health item 2 it was .04 (.012 – .061), for health item 3 it was .07 (.046 – .098) and for health item 4 it was .13 (.102 – .156). Note that the estimates go up for the later, less severe items. For work item 1 the point estimate was .16 (.128 – .184) and for work item 2 it was .08 (.050 – .103).

3. Logistic regression of univariate RR data

In logistic regression the dependent variable is predicted from one or more covariates. Let the z covariates of individual i be collected in covariate vector $\boldsymbol{x}_i$ of length $z \times 1$, and let the regression parameters be collected in a parameter vector $\boldsymbol{\beta}$ of length $z \times 1$. Then one way to formulate logistic regression is as

$$\text{(6a)} \quad \pi_{1i} = \frac{\exp(\beta_0 + \boldsymbol{x}_i^t\boldsymbol{\beta})}{1 + \exp(\beta_0 + \boldsymbol{x}_i^t\boldsymbol{\beta})}$$

$$\text{(6b)} \quad \pi_{2i} = \frac{1}{1 + \exp(\beta_0 + \boldsymbol{x}_i^t\boldsymbol{\beta})}$$

(compare Agresti, 2002, for a general introduction to logistic regression).

Now assume that we deal with RR data, and let elements π_{1i} and π_{2i} refer to the probability of the true status 1 and 2 for individual i, and let π^*_{1i} and π^*_{2i} refer to the probabilities of the responses 1 and 2 for individual i. If we collect the elements π_{1i} and π_{2i} into a vector $\boldsymbol{\pi}_i$ and elements π^*_{1i} and π^*_{2i} into a vector $\boldsymbol{\pi}^*_i$, then

$$\text{(7)} \quad \boldsymbol{\pi}^*_i = \boldsymbol{P}\boldsymbol{\pi}_i$$

with $\boldsymbol{\pi}_i$ defined as in (6a) and (6b). Thus the probability of a sensitive true status of individual i is a function of covariates $\boldsymbol{x}_i$.

The earliest reference to logistic regression for RR data that we came across was Maddela (1983, pages 54-56), and an elaborate treatment can be found in Scheers and Dayton (1988) and van der Heijden and van Gils (1996). Lensvelt-Mulders et al. (2006) extended the logistic regression procedure so that it can incorporate person weights that make it possible to weight a sample toward population characteristics, if known.

The logistic regression model for RR data is estimated by setting up the likelihood and maximizing over the parameters. Let the observed RR data for individual i be (n^*_{1i}, n^*_{2i}), with (n^*_{1i}, n^*_{2i}) = (1,0) if individual i has answer 1, and (n^*_{1i}, n^*_{2i}) = (0,1) if individual i has answer 2. Then the loglikelihood

$$\text{(8)} \quad \ell(\beta_0, \boldsymbol{\beta}) = \sum_i \left(n^*_{1i} \log \pi^*_{1i} + n^*_{2i} \log \pi^*_{2i}\right).$$

For further model development it is useful to write this in matrix terms, similar as in (5). Let $\boldsymbol{u}$ be a unit vector of length 2×1, and collect (n^*_{1i}, n^*_{2i}) in vector $\boldsymbol{n}^*_i$ and of length 2×1 respectively. Then equation (8) can be written as

$$\text{(9)} \quad \ell(\beta_0, \boldsymbol{\beta}) = \sum_i \boldsymbol{u}^t \left(\boldsymbol{n}^*_i \log \boldsymbol{\pi}^*_i\right) = \sum_i \boldsymbol{u}^t \left(\boldsymbol{n}^*_i \log \boldsymbol{P}\boldsymbol{\pi}_i\right),$$

where the elements of $\boldsymbol{\pi}_i$ are defined in (6). Thus maximizing $\ell(\beta_0, \boldsymbol{\beta})$ over the parameters $(\beta_0, \boldsymbol{\beta})$ yields the maximum likelihood estimates. The incorporation of person weights w_i is simply accomplished by reformulating the loglikelihood as

$$\text{(10)} \quad \ell(\beta_0, \boldsymbol{\beta}) = \sum_i \boldsymbol{u}^t \left(w_i \boldsymbol{n}^*_i \log \boldsymbol{P}\boldsymbol{\pi}_i\right).$$

Maddala (1983) provides first and second order derivatives of the loglikelihood and suggests to use the Newton-Raphson method to maximize the loglikelihood (see also Scheers and Dayton, 1988,

and van der Heijden and van Gils, 1996). Van den Hout, van der Heijden and Gilchrist (2007) show that the model is a member of the family of generalized linear models and propose to fit the model with the iterative reweighted least-squares algorithm, which is a very stable fitting procedure.

Example. As an illustration we report an example taken from Lensvelt-Mulders et al. (2006). The dependent variable is the work item "In the last 12 months have you taken on a small job alone or together with your friends that you got paid for without informing the social welfare agency?". As an explanatory variable we choose "I think it is more beneficial to me not to follow the rules connected to my disability insurance benefit", abbreviated as "benefit", that is measured on a five–point scale and has mean 3.67 and standard deviation .777. The logistic regression model logit(non-compliance) = constant + b * benefit has estimates .765 for the constant and .751 for b. In order to study the impact of these estimates, we compare the estimated probability of non-compliance for the mean value of benefit (i.e. 3.67), and the mean plus or minus one standard deviation (i.e. 3.67 + .78 = 4.45 and 3.67 - .78 = 2.89). For the mean value of benefit the estimated probability of non–compliance is 12 percent, for 4.45 the estimated probability is 20 percent and for 2.89 the estimated probability is 7 percent. This shows that benefit has a strong relation with the decision not to comply with the above disability insurance benefit regulation: the more people perceive their benefit if they do not comply, the more they do not comply with this work regulation.

Further developments. Space limitations withhold us to discuss the following developments. First, Frank et al. (2009) discusses the situation that repeated cross sections are carried out with the aim to assess whether compliance with regulations has changed over time. They discuss this in the specific situation that the randomized response design has changed over time. A second development of interest is the development of a linear regression model where the RR variable is the *in*dependent variable. See Van den Hout and Kooiman (2006) for details.

4. Extensions of regression approaches to multivariate RR data

Here multivariate RR data are analyzed directly using an appropriate regression model. As a first example we discuss the analysis of a summary of the multivariate data, namely of a sum score (see Cruyff, van den Hout and van der Heijden, 2008) . For example, if there is a set of M sensitive questions indexed by m ($m = 1, ..., M$) it may be interesting to know how many sensitive questions are answered affirmatively. This problem can be considered as the problem of estimating a sum score variable Z from M RR variables Y_m, where the sum ranges from 0,...,M. A second question is then in what way this sum score can be related to explanatory variables. In this section we describe the approach of Cruyff et al. (For a different approach, see Fox, 2008, who developed what a beta-binomial ANOVA model for randomized response sum score data.)

Let the sum score variable denoting the number of true *yes* responses be defined by

$$(11) \quad Z = \sum_{m=1}^{M} Y_m.$$

Analogously, let the sum score variable $Z^* = \sum_{m=1}^{M} Y_m^*$ denote the number of observed *yes* responses. The probability of observing sum score s on variable Z^*, for $s \in \{0, \ldots, M\}$, is given by the RR sum score model

$$(12) \quad \pi_s^* = \sum_{t=0}^{M} q_{st}\pi_t,$$

where $\pi_s^* = I\!P(Z^* = s)$, $\pi_t = I\!P(Z = t)$ and $q_{st} = I\!P(Z^* = s|Z = t)$. A general definition of the elements q_{st} can be found in Cruyff et al. (2008).

The model is estimated as follows. Note that, similar to equation (2) we can write equation (11) in matrix terms by collecting the elements q_{st} in a matrix $\boldsymbol{Q}$ and it follows that

$$\boldsymbol{\pi^*} = \boldsymbol{Q\pi}. \quad (13)$$

Similar to equation (4), we can find a moment estimate of $\boldsymbol{\pi}$ by

$$\hat{\boldsymbol{\pi}} = \boldsymbol{Q^{-1}}\hat{\boldsymbol{\pi}}^*, \quad (14)$$

where $\boldsymbol{\pi} = (\pi_0, \ldots, \pi_M)', \boldsymbol{\pi}^* = (\pi_0^*, \ldots, \pi_M^*)'$ and π_s^* estimated by n_s^*/n, with n_s^* denoting the frequency of the observed sum score s on variable Z^*. The matrix $\boldsymbol{Q}$ is an $(M+1)\times(M+1)$ transition matrix with entries $(s+1, t+1)$ given by the conditional misclassification probabilities q_{st}, for $s, t \in \{0, \ldots, M\}$. The method of moment solution always fits the data, but can result in probability estimates outside the boundaries of parameter space defined by $(0, 1)$. The maximum likelihood estimates of the RR sum score model are obtained by maximizing the kernel of the observed-data log likelihood

$$\ln \ell(\boldsymbol{\pi}|n_0^*, \ldots, n_M^*), = \sum_{s=0}^{M} n_s^* \ln \left(\sum_{t=0}^{M} q_{s|t}\pi_t \right), \quad (15)$$

for $\pi_t \in (0, 1)$. Kuha and Skinner (1997) have provided EM algorithms. Van den Hout and van der Heijden (2002) show that if the method of moments estimates are in the interior of the parameter space, the maximum likelihood solution is identical to the method of moments solution. Otherwise, one or more maximum likelihood estimates will be on the boundary.

We now present the model for the regression of an RR sum score variable on a set of covariates. Assume that the sum scores are on an ordinal scale and let $I\!P(Z = t|\mathbf{x})$ denote the probability that the sum score variable Z takes on the value t given the covariate vector $\mathbf{x}$. Define $\gamma_t = I\!P(Z \leq t|\mathbf{x})$. Then the proportional odds model (McCullagh, 1980) states that

$$\gamma_t = \frac{\exp(\alpha_t - \mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\alpha_t - \mathbf{x}'\boldsymbol{\beta})}, \quad (16)$$

where the threshold parameters α_t can be thought of as the values on a latent trait variable that mark the transition from $Z = t - 1$ to $Z = t$. The threshold parameters satisfy the condition

$$-\infty < \alpha_0 \leq \alpha_1 \leq \ldots \leq \alpha_M \equiv \infty. \quad (17)$$

Note that for $M = 1$, the order of the threshold parameters is $-\infty < \alpha_0 \leq \alpha_1 \equiv \infty$, and expression (16) reduces to the binary logistic regression model (with a negative sign for $\boldsymbol{\beta}$).

In the RR design, Z is not directly observed. Therefore, the cumulative probabilities $I\!P(Z \leq t|\mathbf{x})$ are modeled through the observed variable Z^*, with the relation between Z^* and Z given by the RR sum score model. The RR proportional odds model is given by

$$\gamma_s^* = \sum_{j=0}^{s} \sum_{t=0}^{M} q_{j|t}(\gamma_t - \gamma_{t-1}), \quad (18)$$

where $\gamma_s^* = I\!P(Z^* \leq s|\mathbf{x})$. For more details, see Cruyff et al. (2008). *Example* Cruyff et al. (2008) report the following example concerning the following three sensitive questions discussed in the introduction: "At a Social Services check-up, have you ever acted as if you were sicker or less able to work than you actually are?", "For periods of any length at all, do you ever feel stronger and healthier and able to work more hours without informing the Department of Social Services of this change?", and "Have you recently done any small jobs for or via friends or acquaintances, for instance in the past year, or done any work for payments of any size without reporting it to the Department of Social Services? (This only pertains to monetary payments.)". They analyzed the sum score variable

Table 1: Parameter estimates of the RR proportional odds model.

Parameters	*Estimates (se)*	*t-value*
α_1	0.99 (0.31)	3.10
α_2	2.46 (0.38)	6.46
Intercept	-0.85 (0.46)	-1.84
Gender	-0.81 (0.26)	-3.14
Education	0.32 (0.16)	2.05
Age	-0.57 (0.28)	-2.23
Time unemployed	0.13 (0.16)	0.80
Last job contract	-0.57 (0.29)	-1.99
Degree of disability	-0.26 (0.25)	-1.05

$Z = \sum_{m=1}^{3} Y_m$, denoting the number of *yes* responses to these three questions of the Social Security Survey, with the RR sum score model and the RR proportional odds model. The frequencies of the sum scores $0, 1, 2, 3$ observed in the sample are given by the vector $\boldsymbol{n}^* = (811, 649, 245, 55)$.

The respective method of moments (MM) sum score probability estimates of the RR sum score model are $\widehat{\boldsymbol{\pi}} = (0.850, 0.075, 0.058, 0.017)$. Since the MM estimates are all in the interior of the parameter space, the ML solution is identical. The log likelihood of ML solution is -1949.54. The same probability estimates and log likelihood can also be obtained with the RR proportional odds null model, i.e. the model without any covariates except the intercept. The parameter estimates of the null model are $\widehat{\beta}_0 = -1.74$, $\widehat{\alpha}_1 = 0.77$ and $\widehat{\alpha}_2 = 2.32$, and the sum score probabilities are found by plugging these estimates into $\widehat{\gamma}_t$ defined in (16), and using expression $\widehat{\pi}_t = \widehat{\gamma}_t - \widehat{\gamma}_{t-1}$.

Table 1 presents the parameter estimates of the RR proportional odds model with all six covariates. The log likelihood of this model is -1937.84, yielding a likelihood ratio test statistic of 23.4 with 6 degrees of freedom in relation to the null model. The parameter estimates of the covariates gender, age, last job contract and education are significant. *Further developments* The multivariate logistic regression model proposed by Glonek and McCullagh (1995) is worked out by van den Hout, van der Heijden and Gilchrist (2007). If there are, for example, two RR variables, the model by Glonek and McCullagh may consist of two univariate logistic regressions as well as a regression model to predict the odds ratio between the two responses. A second further development the so-called Rasch model, which is a model that assumes, among others, a latent variable for the persons, and given the latent variable the answers to the items are independent. For RR data this model was independently adapted by Bockenholt and van der Heijden (2004, 2007) and Fox (2005). Here we just mention that the Rasch model can be extended by relating the latent variable to explanatory variables. Bockenholt and van der Heijden (2007) provide an example. Again, space constraints withhold us from discussing these developments in detail.

5. Discussion: implications of self protective responses for the analysis of univariate and multivariate RR data

Despite the fact that the respondents' privacy is protected by the RR design, it is not always perceived as such by the respondents. Because RR forces respondents to give a potentially self-incriminating answer for something they did not do, it is susceptible to self-protective responses (SP), i.e. respondents answer *no* although they should have responded *yes* according to the randomizing device (see for example, Edgell, Himmelfarb, and Duchan, 1982). The online questionnaires that we used were designed in such a way that the outcome of the dice is not recorded and this was mentioned in the instructions given to the respondents. As a result, the respondents were free to give a different answer than the forced *yes* or *no* induced by the dice. Although RR performs relatively well, by eliciting more admittances of fraud than direct-questioning or computer-assisted self-interviews (Lensvelt-Mulders et al., 2005), non-compliance probabilities might still be underestimated if SP is not taken into account.

Recently, several studies have focussed on the detection or estimation of SP in the setting of RR. Clark & Desharnais (1998) showed that by splitting the sample into 2 groups and assigning each group a different randomization probability, it is possible to detect the presence of SP responses and to measure its extent. Böckenholt and van der Heijden (2007) use a multivariate approach to estimate SP by proposing an item randomized-response model discussed shortly in Section 4. As an extension of this model, the response behavior that does not follow the RR design is approached by introducing mixture components in the IRR models with a first component consisting of respondents who answer truthfully and follow an item response model, and a second component consisting of respondents who systematically say *no* to every item in a subset of items. A similar approach is adopted by Cruyff et al. (2007) who work out the same idea in the context of log linear models. A different approach for sum scores (i.e. different from the approach taken in the proportional odds model discussed in section 5.2) is employed in Cruyff, Böckenholt, van den Hout and van der Heijden (2008) . They introduce a regression model that allows for SP in randomized response sum score data. The model assumes a Poisson distribution for the true sum score variable assessing the individual number of sensitive characteristics. The model further assumes that the observed sum score variable denoting the number of incriminating responses is partly generated by the randomized response design, and partly by SP. Since SP by definition results in an observed sum score of zero, the distribution of the observed sum score variable is zero-inflated with respect to the Poisson randomized response distribution of the true sum score variable. The model allows for predictors that explain individual differences in the Poisson parameters as well predictors that explain individual differences in the probability of SP.

It should be noted that it is not possible to estimate SP from univariate RR data, simply because multivariate RR data are needed to estimate the probability of SP. If we want to the existence of SP into account in the analysis of univariate RR data, we have to "borrow" information from the SP-parameters estimated in multivariate models. To solve this problem, Frank et al. (2008) propose to employ a two-step approach. At the first step we estimate the amount of SP on each wave using multivariate data consisting of three additional RR questions about health conditions, which are part of the full data set. In a second step we use the estimates of SP as external information in our trend analyses. Applying this approach, SP is estimated in the first step using the Profile Likelihood method proposed by Cruyff et al. (2007). Given the estimates of SP for each wave, a correction for SP is carried out by ignoring *no* responses from the sample. For example in 2002, 11% of the sample size is reduced by ignoring observed *no*-responses. In the second step, change in time is modeled using the frequencies adjusted for SP.

In a regression context where individuals have explanatory variables, similar results may be obtained by weighting down those individuals that systematically say *no*. For example, when SP is

estimated to be .10, then the sample size has to be reduced with 10 percent by giving respondents that systematically say *no* lower person weight. For example, if the sample size is 1,000, then a reduction has to be made of 100 respondents. If the number of respondents that systematically say *no* is 300, then each of these respondents should get a person weight of .666 so that their effective number is reduced to .666 * 300 = 200, which is a reduction with 100.

REFERENCES

Agresti, A. (2002). Categorical data analysis. New Jersey : John Wiley and Sons.

Böckenholt, U., and van der Heijden, P. G. M. (2004). Measuring noncompliance in insurance benefit regulations with randomized response methods for multiple items. In A. Biggeri, E. Dreassi, C. Lagazio, and M. Marchi (Eds.), 19th international workshop on statistical modelling (p. 106-110). Florence, Italy.

Böckenholt, U., and van der Heijden, P. G. M. (2007). Item randomized-response models for measuring noncompliance: Risk-return perceptions, social influences, and self-protective responses. Psychometrika, 72, 245-262.

Boruch, R. F. (1971). Assuring confidentiality of responses in social research: a note on strategies. The American Sociologist, 6, 308-311.

Chaudhuri, A. and Mukerjee, R. (1988). Randomized response: Theory and techniques. New York : Marcel Dekker.

Clark, S. J. and Desharnais, R. A. (1998). Honest answers to embarrissing questions: detecting cheating in the randomized response model. Psychological Methods, 3, 160-168.

Cruyff, M. J. L. F., van den Hout, A., van der Heijden, P. G. M., and Böckenholt, U. (2007). Log-linear randomized-response models taking self-protective response behavior into account. Sociological Methods and Research, 36, 266-282.

Cruyff, M. J. L. F., van den Hout, A., van der Heijden, P. G. M., and Böckenholt, U. (2008a). Accounting for self-protective responses in randomized response data from a social security survey using the zero-inflated poisson model. Annals of applied statistics, 2, 316-331.

Cruyff, M. J. L. F., van den Hout, A. and van der Heijden, P. G. M. (2008b). The analysis of randomized-response sum score variables. Journal of the Royal Statistical Society, Series B, 71, 21-30.

Edgell, S. E., Himmelfarb, S., and Duncan, K. L. (1982). Validity of forced response in a randomized response model. Sociologial Methods and Research, 11, 89-110.

Fox, J.-P. (2005). Randomized item response theory models. Journal of Educational and Behavioral Statistics, 30, 1-24.

Fox, J.-P. (2008). Beta-binomial ANOVA for multivariate randomized response data. British Journal of Mathematical and Statistical Psychology, 61, 453-470.

Frank, L. E., van den Hout, A., and van der Heijden, P. G. M. (2009). A logit model for repeated cross-sectional reandomized response data. Methodology, 5, 145-152.

Glonek, G. and McCullagh, P. (1995). Multivariate logistic models. Journal of the Royal Statistical Society, Series B, 57, 533-546.

Kuha, J., and Skinner, C. (1997). Categorical data analysis and misclassification. In L. Lyberg (Ed.), Survey measurement and process quality. New York : Wiley.

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., and Maas, C. J. M. (2005). Meta-analysis of randomized response research: Thirty-five years of validation. Sociological Methods and Research, 33, 319-348.

Lensvelt-Mulders, G. J. L. M., van der Heijden, P. G. M., and Laudy, O. (2006). A validation of a computer-assisted randomized response survey to estimate the prevalence of fraud in social security. Journal of the Royal Statistical Society A, 169, 305-318.

Maddala, G. S. (1983). Limited Dependent and Qualitative Variables in Econometrics. Cambridge : Cambridge University Press.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). Journal of the Royal Statistical Society, Series B, 42, 109-142.

Scheers, N. J., and Dayton, C. M. (1988). Covariate randomized response models. Journal of the American Statistical Association, 83, 969-974.

Van den Hout, A., and Kooiman, P. (2007). Estimating the linear regression model with categorical

covariates subject to randomized response. Computational Statistics and Data Analysis, 50, 3311-3323.

Van den Hout, A., and van der Heijden, P. G. M. (2002). Randomized response, statistical disclosure control and misclassification: A review. International Statistical Review, 70, 269-288.

Van den Hout, A., van der Heijden, P. G. M., and Gilchrist, R. (2007). The logistic regression model with response variables subject to randomized response. Computational Statistics and Data Analysis, 51, 6060-6069.

Van der Heijden, P. G. M., and van Gils, G. (1996). Some logistic regression models for randomized response data. In R. H. A. Forcina G.M. Marchetti and G. Galmatti (Eds.), Statistical modelling. proceedings of the 11th international workshop on statistical modelling. (p. 341-348). Orvieto, Italy.

Van der Heijden, P. G. M., van Gils, G., Bouts, J., and Hox, J. J. (2000). A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning. eliciting sensitive information in the context of welfare and unemployment benefit. Sociological Methods and Research, 28, 505-537.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating answer bias. Journal of the American Statistical Association, 60, 63-69.