Repeated cross-sectional randomized response data: taking design change and self-protective responses into account

L.E. Frank[1], A. van den Hout[2], and P.G.M. van der Heijden[1]
[1]Department of Methodology and Statistics,
Utrecht University, The Netherlands
[2]MRC Biostatistics Unit, Institute of Public Health,
Cambridge, UK

September 3, 2009

Abstract

Randomized response (RR) is an interview technique that can be used to protect the privacy of the repondents when faced with sensitive questions. This paper describes how to measure change in time when a binary RR question has been asked on several time points. In cross-sectional research settings, often new insights gradually become available. In our setting, a switch to another RR procedure occurred and it necessitated the development of a trend model that estimates the effect of the covariate time when the dependent variable is measured by different RR designs. Additionally, we show that it is possible to deal with the presence of self-protective responses, thereby accommodating our trend model with the latest developments in the analysis of RR data. The model proposed is not limited to change in time, but generalizes to every cross-classification of a RR variable with a categorical variable.

keywords: linear trend, longitudinal data, misclassification, randomized response, repeated cross-sections, self-protective responses

1 Introduction

Randomized response (RR) is an interview technique that can be used when sensitive questions have to be asked and respondents are reluctant to answer directly (Warner, 1965; Chaudhuri and Mukerjee, 1988). Examples of sensitive questions are questions about fraud, alcohol consumption or sexual behavior. A recent meta-analysis study showed that randomized response designs lead to more valid answers compared to other conventional question-and-answer methods (Lensvelt-Mulders et al., 2005b). An RR design can be defined in various ways, but all designs have in common that a specified probability mechanism protects the privacy of the individual respondent. The resulting RR variables represent misclassified responses on categorical variables where conditional misclassification probabilities are fixed by design (van den Hout & van der Heijden, 2004). The true status of the individual respondent is not revealed because his or her observed answer depends not only on the true status but also on the misclassification design.

Besides the randomized response setting, misclassification probabilities occur in several other fields of research. Most related to randomized response is the post randomization method (PRAM, Kooiman, Willenborg & Gouweleeuw, 1997), where values of categorical variables are misclassified after the data have been collected to protect the privacy of the respondents. PRAM can be considered as applying RR after the data collection. Misclassification also plays a role in medicine and epidemiology with the probabilities to be correctly classified as a case (sensitivity) or as a non-case (specificity), see Chen, 1989; Copeland, Checkoway, McMichael, & Holbrook, 1977; Greenland, 1980, 1988; Magder & Hughes, 1997. Misclassified data can be analyzed with loglinear models or the general framework of latent variable models and latent class models (see for example, Haberman, 1979; Hagenaars, 1990, 1993; Rabe-Hesketh & Skrondal, 2007; Skrondal & Rabe-Hesketh, 2004; van den Hout & van der Heijden, 2002; Walter, Irwig & Glasziou, 1999)

This paper proposes a model to measure change in time when RR is used to ask a sensitive question at several time points. The model will be illustrated with data from a Dutch repeated cross-sectional study on non-compliance to rules in the area of social benefits. Data have been collected every two years since 2000, and given that measures to prevent regulatory non-compliance have been intensified during this period, the question rises whether the prevalence of regulatory non-compliance has changed over the years and how the change can be modeled.

Considering time as a covariate, we propose a new approach to measure the effect of this covariate when the dependent variable is measured by randomized response. Several aspects of this cross-sectional study make it impossible to use standard analysis methods and necessitate a new approach in the field of the analysis of randomized response data, to be able to deal with this type of research questions. First, the fact that RR variables represent misclassified responses on categorical variables precludes the use of, for example, the *linear logit model* (Agresti, 2002, p. 180), to test for a linear trend. The trend tests proposed in this paper take into account the misclassification induced by the RR design,

using the framework of Van den Hout and Van der Heijden (2004) and the results obtained by Maddala (1983) and Scheers and Dayton (1988). Second, as a consequence of increasing knowledge about the efficiency of randomized response designs, a change in the design occurred during the cross-sectional study. We show in this paper how to accommodate the trend model for design changes. Third, accounting for self-protective responses (SP) being a new development in the field of the analysis of RR data (Böckenholt & van der Heijden, 2007; Cruyff et al., 2007), we also show a way to incorporate the presence of SP into the trend model.

The outline of the paper is as follows. The next section explains the randomized response design as a misclassification design and shows how to deal with changes in the randomized response design over time. Section 3 introduces the trend model for RR variables with an additional procedure to account for self-protective responses. Section 4 shows an application of the model and section 5 ends the paper with a discussion.

2 The randomized response design

RR designs use a randomizing device that perturbs the answers of the respondents. The basic idea behind RR is that the perturbation induced by the misclassification design protects the privacy of the respondent. At the same time, the researcher knows the nature of the perturbation and this allows for a correct analysis of the observed data where the misclassification is taken into account. There are several randomized response designs (*cf.* Fox & Tracy, 1986, Chapter 2). Two of these designs are used in the application of this paper and will be discussed below. Each design uses a different randomizing device, namely playing cards and dice.

For Kuk's randomized response design (Kuk, 1990), the randomizing device consists of two stacks of cards. The idea of using a randomizing device is to generate binary outcomes, i.e. the *yes* and *no* answers, according to two Bernoulli distributions with known parameters. A way to elicit the required binary outcomes it to use two stacks of cards with varying proportions of red cards. Assume answering *yes* to the sensitive question is associated with the color red, the Kuk design can be implemented by creating a stack that contains more red cards than black cards, $\frac{8}{10}$ and $\frac{2}{10}$, respectively. The other stack, representing the *no* answer contains more black cards than red cards, with a proportion of $\frac{2}{10}$ of red cards. After shuffling each stack, the respondent is asked to draw a card at random from each stack. Instead of answering *yes* or *no*, the respondent expresses the answer *yes* by naming the color of the card that came from the right stack (the one with the higher proportion of red cards), or if he wishes to express the answer *no*, by naming the color of the card from the left stack (the one with the higher proportion of black cards). More details about the implementation of the Kuk design are available in Van der Heijden et al. (2000).

The forced response design (Boruch, 1971) uses dice as the randomizing device. The binary responses are kow generated according to the known distribution of the sum of the outcomes of two dice. After the sensitive question is asked, the respondent throws two dice and keeps the outcome hidden from the interviewer. If the outcome of the two dice is 2, 3 or 4, the respondent answers *yes*. If the outcome is 5, 6, 7, 8, 9 or 10, he answers according to the truth. If the outcome is 11 or 12, he answers *no* (for the implementation of the forced response method, see Lensvelt-Mulders et al., 2006).

In the RR design by Kuk, violations are associated with the color *red* (expressing the answer *yes*). As a result, the probability to be correctly classified is $8/10$ both for respondents who violated regulations and for those who did not. The RR matrix that contains the conditional misclassification probabilities

$$p_{ij} = P(\text{category } i \text{ is observed} | \text{ true category is } j) \tag{1}$$

is therefore given by

$$\boldsymbol{P}_{Kuk} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 8/10 & 2/10 \\ 2/10 & 8/10 \end{pmatrix}. \tag{2}$$

Similarly, the forced response design yields the following transition matrix

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

As an illustration, given the forced response design with the transition matrix of Equation 3, the probability of a forced *yes* is equal to $p_{12} = \frac{1}{6}$, the probability of a forced *no* is $p_{21} = \frac{1}{12}$ and the probability of a truthful answer is $1 - (p_{12} + p_{21}) = \frac{3}{4}$. The probability of an observed *yes* response is $\pi_1^* = p_{12} + (1 - (p_{12} + p_{21}))\pi_1$. Taking the observed proportion of *yes* answers as an estimate $\hat{\pi}_1^*$ of π_1^*, an estimate of π_1 can be obtained with:

$$\hat{\pi}_1 = \frac{\hat{\pi}_1^* - p_{12}}{1 - (p_{12} + p_{21})} \tag{4}$$

and the estimated variance of $\hat{\pi}_1$ is given by (*cf.* Fox & Tracy, 1986, p.21):

$$\hat{\sigma}_{\hat{\pi}_1} = \frac{\hat{\pi}_1^*(1 - \hat{\pi}_1^*)}{N(p_{12} + p_{21})^2}. \tag{5}$$

The general form of RR designs is (Chaudhuri & Mukerjee, 1988; van den Hout & van der Heijden, 2002):

$$\boldsymbol{\pi^*} = \boldsymbol{P\pi}, \tag{6}$$

where in case of dichotomous items, $\boldsymbol{\pi^*} = (\pi_1^*, \pi_2^*)'$ is a vector with the probabilities of the observed answers, $\boldsymbol{\pi} = (\pi_1, \pi_2)'$ is the vector of the probabilities of the latent status and $\boldsymbol{P}$ is the 2×2 matrix defined in Equation 2 or Equation 3. If $\boldsymbol{P}$ is non-singular and the observed proportion of *yes* and *no* answers are unbiased point estimates $\hat{\boldsymbol{\pi}}^*$ of $\boldsymbol{\pi^*}$, $\boldsymbol{\pi}$ can be estimated by the unbiased moment estimator (Chaudhuri & Mukerjee, 1988; Kuha & Skinner, 1997)

$$\hat{\boldsymbol{\pi}} = \boldsymbol{P}^{-1}\hat{\boldsymbol{\pi}}^*. \tag{7}$$

However, in practice it is possible to obtain estimates that are outside the parameter space (0,1) when, for example, the observed proportion of *yes* answers is very small. Van den Hout and van der Heijden (2002) have demonstrated that the maximum likelihood estimator (MLE) is in general a good alternative to the moment estimator (ME) and, in case of boundary solutions, the authors propose to use the maximum of the likelihood for point estimation and the bootstrap percentile method for confidence intervals.

2.1 Repeated cross-sectional randomized response data and different RR designs

Estimation efficiency and perceived privacy protection In research settings with repeated measurements, design changes can occur when more knowledge about the properties of the designs gradually becomes available. For example, the repeated cross-sectional study on regulatory non-compliance that serves as an illustration in the application section of this paper, uses two different RR designs: Kuk's method was used in 2000 and for the remaining years the forced response design was used. The switch to the forced response design in 2002 follows from more insight into the advantages of this design.

The probabilities of the randomization device result from a compromise between estimation efficiency and perceived privacy protection by the respondents. Fox and Tracy (1986, pp. 25-26) provide an extensive discussion of this issue: to offer optimal protection to the respondents, the probability for giving a truthful response should be as small as possible. However, the smaller the truthful response probability, the larger the sampling variance of the estimator, leading to less efficiency because fewer respondents provide relevant information about the sensitive behavior.

Increasing research findings suggest that the forced response design is more efficient (Lensvelt-Mulders et al., 2005a), is comparatively easy for respondents to follow and, the probabilities of a forced *yes* or *no* tend to be overestimated by the respondents (Moriarty & Wiseman, 1976). Kuk's design has the advantage that respondents answer the questions with colors instead of the more self-incriminating *yes* answer, but it will not be able to give the same level of privacy protection, even if it is made as efficient as the forced response design.

With regard to the perception of the privacy protection offered by the forced response design, the choice of the value $\frac{3}{4}$ for the probability of a truthful answer, as described in Section 2, does not seem to be the smallest possible probability, but it follows from the results obtained by Moriarty and Wiseman (1976) and Soeken and Macready (1982) who demonstrate that the probability of a truthful answer can

be chosen between .7 and .8 without interfering with the perceived grade of anonymity. Given their results, by choosing .75, there is a probability of .25 to be divided between the forced *yes* and the forced *no* probability. The *yes* answer represents the acknowledgement of non-compliance and because of the respondents' reluctancy to admit non-compliance, the forced *yes* probability is twice as large as the forced *no* answer to make the respondent more comfortable with answering *yes*. At the same time the forced *yes* probability is approximately in the same range as the expected prevalence of the sensitive topic in the population, as recommended by Clark and Desharnais (1998).

Accommodating changes in the RR designs Given the switch of RR design, the misclassification probabilities can be arranged in such a way that it becomes possible to estimate prevalences for RR variables that have been collected in a repeated cross-section. In our application (see Section 4), RR variables were measured on three time points and the matrix of misclassification probabilities in (6) can be generalized as follows. First, the probabilities of the observed answers have to be restructured. The randomized response variable with two categories $i = 1, 2$ was observed on three time points ($t = 1, 2, 3$), leading to the probabilities for the observed answers π^*_{it}. The 2×3 table of observed answer probabilities can be represented as a vector $\boldsymbol{\pi}^* = (\pi^*_{11}, \pi^*_{21}, \pi^*_{12}, \pi^*_{22}, \pi^*_{13}, \pi^*_{23})$, and similarly, we obtain the vector $\boldsymbol{\pi} = (\pi_{11}, \pi_{21}, \pi_{12}, \pi_{22}, \pi_{13}, \pi_{23})$. The transition matrix $\boldsymbol{P}$ in Equation 6 can be extended to a block diagonal matrix $\boldsymbol{P}$ composed of blocks $\boldsymbol{P}_t$ for time point t. The result is the following 6×6 matrix:

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{P}_1 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}_3 \end{pmatrix} \tag{8}$$

To accomodate the different RR designs used in our application, which consists of a combination of Kuk's method and the forced response method, the block diagonal matrix can be changed in the following way: $\boldsymbol{P}_1$ is defined by the missclassification probabilities of the Kuk design in (2) and $\boldsymbol{P}_2$ and $\boldsymbol{P}_3$ are defined by the missclassification probabilities of the forced response design in (3).

3 Logit model for trend

We now return to our research question whether the prevalence of non-compliance has changed over the years and how the change can be modeled. Consider the sensitive question as the dependent variable and the time points as the independent variable with scores $t = 1, 2, 3$. If one expects a monotone trend, this hypothesis can be tested with the *linear logit model* (*cf.* Agresti, 2002, p. 180):

$$\pi_{1t} = \frac{\exp(\beta_0 + \beta_1 t)}{1 + \exp(\beta_0 + \beta_1 t)} \tag{9a}$$

$$\pi_{2t} = \frac{1}{1 + \exp(\beta_0 + \beta_1 t)} \tag{9b}$$

where the category 1 refers to a *yes* answer. The *independence* model is the special case where $\beta_1 = 0$. An expansion of the model to include a quadratic trend is obtained by adding the term $\beta_2 t^2$ to $\beta_0 + \beta_1 t$. The log likelihood is given by

$$\ell(\boldsymbol{\beta} \mid n) = \sum_t n_{1t} \log \pi_{1t} + \sum_t n_{2t} \log \pi_{2t}, \tag{10}$$

which can be expressed more concisely as

$$\ell(\boldsymbol{\beta} \mid \boldsymbol{n}) = \boldsymbol{u}(\boldsymbol{n} \ \log \boldsymbol{\pi}), \tag{11}$$

where $\boldsymbol{u}$ is the unit vector.

When the dependent variable is an RR variable, the log likelihood must take into account the misclassification induced by the RR design. Using the misclassified observed frequencies n^*_{it} and the mis-

classified observed probabilities π^*_{it}, the adaptation of the log likelihood in Equation 10 becomes:

$$\ell(\boldsymbol{\beta} \mid \boldsymbol{n}^*) \quad = \quad \sum_t n^*_{1t} \log \pi^*_{1t} + \sum_t n^*_{2t} \log \pi^*_{2t}.$$

In analogy to (11) the log likelihood can also be expressed in matrix algebra as:

$$\ell(\boldsymbol{\beta} \mid \boldsymbol{n}^*) = \boldsymbol{u}(\boldsymbol{n}^* \log \boldsymbol{\pi}^*) = \boldsymbol{u}(\boldsymbol{n}^* \log \boldsymbol{P}\boldsymbol{\pi}), \tag{12}$$

where elements π_{1t} and π_{2t} of vector $\boldsymbol{\pi}$ are defined in (9a) and (9b) and $\boldsymbol{P}$ is block diagonal matrix as in (8). Maximizing the log likelihood in (12) over parameters $\boldsymbol{\beta}$ leads to estimated probabilities for the *yes* and *no* answers on each time point, $\hat{\pi}_{1t}$ and $\hat{\pi}_{2t}$.

A goodness-of-fit measure for the trend models can be obtained with the likelihood ratio statistic using the log likelihood defined in (12). It allows for testing the hypotheses of no change (the independence model) and linear or quadratic trend (the last named being only possible, of course, if there are enough time points to leave degrees of freedom). It is well known that the use of the order in the time points leads to more efficient estimates of π_{it} as well as to more powerful tests (*cf.* Agresti, 2002, Section 6.4, p. 236). We note that the framework proposed here is not restricted to model changes in time of an RR variable, and applies to each cross-classification of an RR variable with any categorical variable. The R-code to fit the models described in this section is available from the authors.

3.1 Accounting for self-protective responses

Despite the fact that the repondents' 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, & Duchan, 1982). In our application, the online questionnaires were designed in such a way that the outcome of the dice is not recorded and this fact 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 admissions of fraud than direct-questioning or computer-assisted self-interviews (Lensvelt-Mulders et al., 2005b), 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 and 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 (IRR), where a common sensitivity scale is assumed for a set of RR variables. 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 and co-authors (2007) who work out the same idea in the context of log linear models.

As we feel that this new development to correct RR estimates for SP responses is important, we propose the following procedure to incorporate the existence of SP into the trend model. 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 a percentage of the observed *no* responses from the sample that is equal to the estimated amount of SP. In the second step, change in time is modeled using the frequencies adjusted for SP.

The recent developments in the estimation of SP offer possibilities for longitudinal research settings. It allows for adopting the misclassification probabilities in the RR design according to the proportion of SP estimated on the previous time point, and, consequently, leading to a better balance between estimation efficiency and privacy protection (as discussed in Section 2.1). It should be noted that in the

application we describe in this paper, methods to estimate SP were not available on the first and the second time point, and, consequently, SP for the earlier time points could only be estimated retrospectively and no changes could be made to the misclassification probabilities.

A drawback of the two-step approach we propose is that the uncertainty about the estimates of SP in the first step is not automatically taken into account by the theoretical standard errors in the trend model in the second step. Therefore, empirical standard errors are obtained for the regression coefficient of the trend model using the non-parametric bootstrap (Efron & Tibshirani, 1993).

The details of the bootstrap procedure are as follows:

- For each of the time point, sample B times n respondents with replacement, where n is equal to the sample size at each time point.
- For each of the time points estimate SP for each bootstrap sample. Adjust the bootstrap sample frequencies for SP on each time point (first step of the two-step approach).
- Fit the independence model and the linear trend model to each of the B bootstrap samples. This results in B estimates of the intercept in the independence model (or intercept only model) and B estimates of the intercept and the slope in the linear trend model. The standard deviation of the distribution of these B estimates for intercepts and slope yields the bootstrap estimates of standard errors and the 95% bootstrap percentile intervals (second step of the two-step approach).

4 Application: prevalence of regulatory non-compliance in social benefit area

4.1 The data

Dutch employees are obliged to be insured under the Invalidity Insurance Act and, provided certain conditions are met, a formerly employed person is entitled to receive financial benefits, which can amount to as much as 70% of the person's last income. The welfare system being rather costly, the Department of Social Affairs in the Netherlands monitors the prevalence of non-compliance to the rules on a regular basis. After a pilot in 1998, three waves followed in the years 2000, 2002, and 2004. A detailed description of the 2002 cross-sectional study is given in Lensvelt-Mulders et al. (2006). The Department of Social Affairs has intensified the measures to prevent regulatory non-compliance during these years and is interested in knowing whether the prevalence of regulatory non-compliance has changed over the years and how the change can be modeled.

The application focusses on the following sensitive question concerning the health status of the respondent: *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?* If non-compliance is detected, it

Table 1: Observed weighted frequencies of *yes* (n_1^*) and *no* (n_2^*) answers, and estimated probabilities ($\hat{\pi}_1$) with 95% confidence intervals of non-compliance corrected for the RR design for the cross-sectional data on regulatory non-compliance, measured on three time points. Person weights were used to weight the sample towards population characteristics (*cf.* Lensvelt-Mulders et al., 2006)

	2000 (n=1308)	2002 (n=1760)	2004 (n=830)
n_1^*	388	466	197
n_2^*	920	1294	633
$\hat{\pi}_1$	0.16	0.10	0.07
95% CI	[0.13, 0.19]	[0.07, 0.13]	[0.03, 0.11]

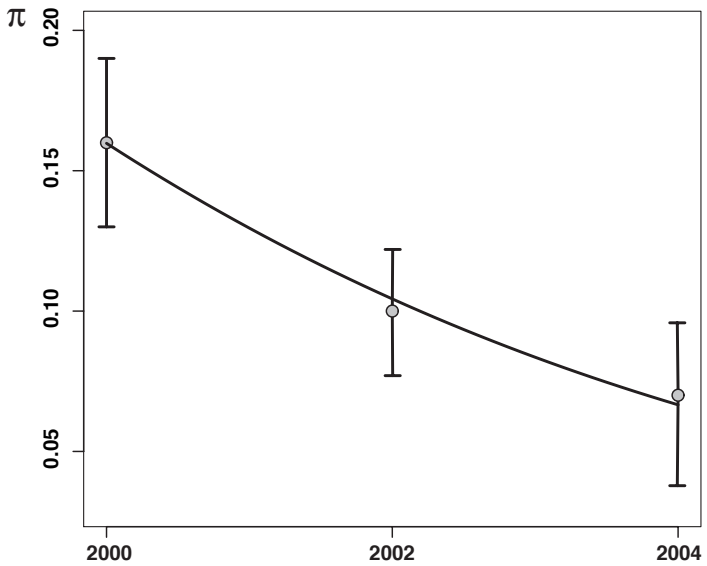

Figure 1: Univariately estimated regulatory non-compliance probabilities (represented by the dots) with 95% confidence intervals, and fitted trend line on the RR question *For periods of any length at all, do you ever feel stronger and healthier and able to work more hours without informing the Department of Social Services of this change?*

can lead to sanctions and sometimes even to loss of invalidity insurance benefits. Given the sensitivity of the topic, asking respondents directly whether they violate these obligations, will not yield valid results (*cf*. Van der Heijden et al., 2000). Therefore, a randomized response design has been used at each wave to ensure the confidentiality of the answers. For the question just described, Table 1 displays the observed, misclassified, frequencies of *yes* (n_1^*) and *no* (n_2^*) answers for the three time points, as well as the estimated probabilities of regulatory non-compliance corrected for the RR design (as explained in section 2). A change of the randomized response design occurred at time point 2002, where the Kuk's design was replaced with the forced response design. Accordingly, the block diagonal matrix of misclassification probabilities in (8) has been used to accommodate this change of RR design.

4.2 Results

Two models were fitted to the RR data: the independence model and the linear trend model. The goodness-of-fit of the models was evaluated with the likelihood ratio statistic. Figure 1 shows that the estimated regulatory non-compliance probabilities decrease monotonically over the time points, and, accordingly, the estimated logistic regression parameter for the linear trend has a negative value (see Table 2). The value of the likelihood ratio statistic L^2 in Table 2 indicates that the independence

Table 2: Results trend analyses (the model of choice is in bold typeface)

Model	L^2	df	p	$\hat{\beta}_0$ ($\hat{\sigma}_{\hat{\beta}_0}$)	$\hat{\beta}_1$ ($\hat{\sigma}_{\hat{\beta}_1}$)
[1] Independence	10.415	2	.001	-2.09 (0.10)	
[2] **Linear**	**0.004**	**1**	**.95**	**-1.17 (0.28)**	**-0.49 (0.16)**
	ΔL^2	Δdf	p		
[1] - [2]	10.411	1	.001		

Table 3: Results trend analyses on univariately estimated regulatory non-compliance probabilities corrected for SP with likelihood ratio statistics (L^2), bootstrap estimates of standard errors ($\hat{s}_B$) and bootstrap percentile confidence intervals

Model	$\hat{\beta}$	$\hat{s}_B$	95% CI
[1] Independence ($L^2 = 17.62, df = 2, p = .00$)			
$\hat{\beta}_0$	-1.61	0.10	[-1.75, -1.41]
[2] Linear ($L^2 = 1.27, df = 1, p = .26$)			
$\hat{\beta}_0$	-0.67	0.30	[-1.07, -0.09]
$\hat{\beta}_1$	-0.50	0.16	[-0.81, -0.26]

model does not fit ($L^2 = 10.415$, $df = 2$, $p = .001$), whereas the linear model produces good fit values: $L^2 = 0.004$, $df = 1$, $p = .95$. Testing the linear model against the independence model (see the ΔL^2-values in the lower part of Table 2) leads to the conclusion that the linear trend model forms a significant improvement. Also, both parameters of the linear trend model depart significantly from zero. Summarizing, this means that among the persons entitled to receive social benefits, the proportion of respondents who do not comply to the rule of informing the Department of Social Services about their health improvement, has significantly decreased in the period from 2000 to 2004.

Accounting for self-protective responses We now present the results of the trend analysis that takes SP into account, using the two-step approach. It should be noted that it is not possible to model change in time and SP behavior simultaneously, simply because multivariate RR data are needed to estimate the probability of SP, whereas we have repeated univariate data here (note that at each time point a distinct sample is used, see Table 1). 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, the proportion of SP was estimated on the data in Table 1 with the Profile Likelihood method proposed by Cruyff et al. (2007). The resulting proportion of SP-answers for the three waves are: .13, .15, and .11, for the years 2000, 2002, and 2004 respectively. Adjusting the observed frequencies for these SP proportions, yields the following estimates of regulatory non-compliance probabilities $\widehat{\pi}$: .24, .17, and .11 for the years 2000, 2002, and 2004, respectively. Comparing these SP-corrected non-compliance probabilities with the uncorrected probabilities in Table 1, clearly shows that the SP-corrected probabilities are higher and that not accounting for SP leads to underestimation of non-compliance probabilities.

In the second step, change in time is modeled by fitting the trend model to the observed frequencies adjusted for SP, and leads to the results displayed in Table 3. The likelihood ratio statistic L^2 is equal to $17.62 (df = 2, p = .00)$ for the independence model, and for the linear trend model $L^2 = 1.27 (df = 1, p = .26)$. The model fit clearly improves after adding a parameter to account for linear trend. The variability of the logistic regression parameters is estimated by bootstrap standard deviations, following the bootstrap set-up explained in Section 3.1. The results are based on the observed frequencies of *yes/no* responses in Table 1 and from three additional RR questions about health conditions for each of the three waves. For each of the years 2000, 2002 and 2004, $1,000$ times n respondents were sampled with replacement, where n is equal to the sample size 1308, 1760 and 830 for the years 2000, 2002 and 2004 respectively.

The 95% bootstrap confidence intervals (percentile method) show that both parameters of the linear model depart significantly from zero, leading to the conclusion that the SP-corrected non-compliance probabilities decrease monotonically over the three time points. This means that the proportion of persons not complying to the rule of giving information about health improvement, has decreased in the period between 2000 and 2004.

5 Discussion

This paper showed how to measure the effect of the covariate time on repeated cross-sectional randomized response data and how to take into account RR design changes and the presence of self-protective responses. Due to the misclassification design, traditional trend models cannot be used. One of the key elements of the method proposed, is the construction of a block diagonal matrix with conditional classification probabilities for each time point, which allows for using different RR designs over time.

The method is not limited to modeling changes in time and can be applied to each cross-classification of an RR variable with a categorical variable. Consider for example a study where one is interested whether residents of large city's are more or less disposed to report non-compliance behavior than residents of smaller communities. If the RR question has the two categories *yes* and *no*, and if population size of the place of residence is measured in five categories, the result is a 2×5 factorial design. One could use the method proposed in this paper to test whether the non-compliance probabilities change monotonically according to the population size of the place of residence.

Despite the fact that the RR method offers protection of the respondents' privacy, it does not entirely exclude the presence of evasive response bias. As a result, non-compliance probabilities might still be underestimated if self-protective responses are not taken into account. We showed in this paper that it is possible to correct RR estimates for SP in the trend model, using a two-step procedure where the amount of SP is estimated in the first step and the trend model is fitted on frequencies corrected for the SP estimates in the second step. The new approach we propose, becomes a powerful tool in longitudinal research settings when it is combined with estimates of the occurrence of SP, as it is now possible to adjust the misclassification probabilities of the RR design according to the estimates of SP on the previous time point, thereby offering a better balance between estimation efficiency and privacy protection.

The two-step procedure has the disadvantage that it is computationally demanding, however, it seems inevitable in the setting of repeated univariate data, where at each time point a distinct sample is used. In this situation, it is not possible to model change in time and SP behavior simultaneously, as shown in the recently developed methods to correct RR estimates for SP responses. It should be noted that the two-step procedure uses the same data twice: to estimate the SP probabilities and to fit the model with associated standard errors for the model parameters. A possible solution would be to use cross-validation: estimating SP and fitting the model on the training set and obtaining the variability estimates for the model parameters in the test set, although this would lead to a very complex simulation set-up. Estimating SP is a recent development and requires further research.

References

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

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

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

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

Chen, T. T. (1989). A review of methods for misclassified categorical data in epidemiology. *Statistics in Medicine*, *8*, 1095-1106.

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

Copeland, K. T., Checkoway, H., McMichael, A. J., & Holbrook, R. H. (1977). Bias due to misclassification in the estimation of relative risk. *American Journal of Epidemiology*, *105*, 488-495.

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

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

Efron, B., & Tibshirani, R. (1986). *An introduction to the Bootstrap*. London: Chapman and Hall.

Fox, J. A., & Tracy, P. E. (1986). *Randomized Response. A Method for Sensitive Surveys* (No. 58). Newbury Park: Sage Publications.

Greenland, S. (1980). The effect of misclassification in the presence of covariates. *American Journal of Epidemiology*, *112*, 564-569.

Greenland, S. (1988). Variance estimation for epidemiologic effect estimates under misclassification. *Statistics in Medicine*, *7*, 745-757.

Haberman, S. J. (1979). *Analysis of Qualitative Data: New Developments (Vol. 2)*. New York: Academic Press.

Hagenaars, J. A. (1990). *Categorical longitudinal data. Loglinear analysis of panel, trend and cohort data*. Newbury Park: Sage.

Hagenaars, J. A. (1993). *Loglinear models with latent variables*. Newbury Park: Sage.

Kooiman, P., Willenborg, L. C. R. J., & Gouweleeuw, J. M. (1997). *PRAM: a method for disclosure limitation of microdata* (Research paper No. 9705). Voorburg/Heerlen: Statistics Netherlands.

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

Kuk, A. Y. C. (1990). Asking sensitive questions indirectly. *Biometrika*, *77*, 436-438.

Lensvelt-Mulders, G. J. L. M., Hox, J. J., & van der Heijden, P. G. M. (2005a). How to improve the efficiency of randomized response designs. *Quality and Quantity*, *39*, 253-265.

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. J. M. (2005b). 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., & 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.

Magder, L. S., & Hughes, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*, *146*, 195-203.

Moriarty, M., & Wiseman, F. (1976). On the choice of a randomization technique with the randomized response model. In *Proceedings of the social statistics section of the american statistical association* (p. 624-26). Washington D.C.: American Statistical Association.

Rabe-Hesketh, S., & Skrondal, A. (2007). Multilevel and latent variable modeling with composite links and exploded likelihoods. *Psychometrika*, *72*(2), 123-140.

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

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling*. Boca Raton: Chapman and Hall/CRC.

Soeken, K. L., & Macready, G. B. (1982). Respondents' perceived protection when using randomized response. *Psychological Bulletin*, *92*, 487-489.

Van den Hout, A., & 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. (2004). The analysis of multivariate misclassified data with special attention to randomized response. *Sociological Methods and Research*, *32*, 384-410.

Van der Heijden, P. G. M., van Gils, G., Bouts, J., & 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.

Walter, S. D., Irwig, L., & Glasziou, P. P. (1999). Meta-analysis of diagnostic tests with imperfect reference standards. *Journal of Clinical Epidemiology*, *10*, 943-951.

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