Do randomized–response designs eliminate response biases? An empirical study of non–compliance behavior. *

Ulf Böckenholt[1], Sema Barlas[1], and Peter G.M. van der Heijden[2].

[1] Department of Management, McGill University,

[2]Department of Methodology and Statistics, Utrecht University

*Correspondence should be addressed to Ulf Böckenholt (ulf.bockenholt@mcgill.ca), Faculty of Management, McGill University, 1001 Sherbrooke Street West, Montreal, QC H3A 1G5, Canada; Sema Barlas (sema.barlas@mcgill.ca), Faculty of Management, McGill University, 1001 Sherbrooke Street West, Montreal, QC H3A 1G5, Canada; or to Peter van der Heijden (p.vanderheijden@fss.uu.nl), Department of Methodology and Statistics, Utrecht University, P.O. Box 80.140, 3508 TC Utrecht, Netherlands. The authors are grateful to Rolf Langeheine and to two reviewers whose suggestions helped to improve the paper substantially. The authors also wish to thank the Dutch Ministry of Social Affairs and Employment for making the reported data available. This research was supported in parts by grants from the Social Sciences and Humanities Research Council of Canada and the Canadian Foundation of Innovation.

Abstract

Do randomized–response designs eliminate response biases? An empirical study of non–compliance behavior.

Out of the set of survey methods for obtaining honest answers to sensitive issues, the method of randomized responses (RR) has proven to be the most effective one. So far, in applications of RR methods it has been assumed that they eliminate response biases. We investigate the validity of this assumption by applying multivariate RR models that allow for different types of response biases. Our data analyses show that RR methods do not eliminate response biases but that the biases can be modeled in informative ways: Accounting for response biases leads to estimates that are at least twice the size of the estimates obtained when response biases are ignored.

1 Introduction

Models for the analysis of self–report data have become increasingly sophisticated over the years in market–research applications. Going beyond the consideration of random sources of measurement error (Bound, Brown, & Mathiowetz, 2001), a number of models have been developed to allow for individual differences in response processes to survey questions (Baumgartner & Steenkamp, 1992; Benitez-Silva, Buchinsky, Chan, Cheidvasser & Rust, 2004; de Jong, Steenkamp, Fox & Baumgartner, 2005; Hsiao & Sun, 1998; Hsiao, Sun, & Morwitz, 2002; Mittal & Kamakura, 2001). The importance of this development for more informative analyses of self–reports cannot be underestimated in view of the considerable body of literature in psychology and survey research on individual response behavior (McFadden et al., 2005; Schwarz, 2003; Tourangeou, Rips, & Rasinski, 2000). A large number of seemingly irrelevant factors have been shown to influence greatly the different stages of a response process ranging from comprehension of a question to the editing of answers. When not taken into account, these factors may bias substantially the measurement of subjective phenomena. The factors' influences do not end with a self–report but can carry over to actual behaviors. A recent study by Chandon, Morwitz and Reinarzt (2005) showed that simply asking respondents about their intentions yielded a substantially higher correlation between latent intentions and purchase behavior among surveyed consumers than among similar non–surveyed consumers.

This article focuses on models for the analysis of self–reports on personal and sensitive issues. For example, in marketing research studies interviews may be conducted on product usage of male cosmetics or of drugs to overcome impotence, on health problems such as urinary incontinence, or on medical prescriptions, but also on such potentially

less sensitive issues such as shopping expenditures and default rates in credit card usage (Blair, Sudman, Bradburn, & Stocking, 1977; Lamb & Stem, 1978). Other topics in marketing concern the "dark side" of consumer behavior such as stealing or, more generally, non–compliance with rules and regulations that govern public life (Benitez–Silva, Buchinsky, Chan, Cheidvasser & Rust, 2004). It is well–known that direct questions about such topics may lead to response refusals or to deliberate misreports. In general, the likelihood of an edited response increases as the topic becomes more sensitive, and it is higher among those respondents with something to hide (Tourangeou et al., 2000).

Response–elicitation methods that protect the privacy of a respondent provide one avenue to reduce the incidence of edited answers. In particular, randomized response (RR) methods (Warner, 1965) can outperform significantly more direct ways of asking sensitive questions, especially when the sensitivity of the topic under investigation is high (Lensvelt, Hox, van der Heijden and Maas, 2005). The work by Lara, Strickler, Olavarrieta, and Ellerston (2004) represents a recent sociological example that demonstrates the usefulness of RR methods. These authors found that in comparison to face–to–face interviews, audio computer–assisted self–interviews and self–administered questionnaires, the RR method was the most promising methodology to measure rates of induced abortion in Mexico.

Despite the apparent usefulness of RR methods in reducing a person's likelihood to edit their responses, applications utilizing this methodology have focused almost exclusively on the estimation of population proportions (e.g., frequency of gambling, drinking, or sexual activities) rather than on understanding the determinants of the behavior under consideration. To study and relate the behavior of interest to a wider context, however, necessitates broadening the simple objective of past studies since two different aspects

of individual differences in the response process need to be taken into account. First, respondents are likely to differ in their propensity to exhibit the phenomenon under study. Second, respondents may differ in the degree to which they provide truthful answers to personal or sensitive questions. This behavior is likely to depend on the extent to which their studied behavior deviates from social norms, and the degree to which a person is motivated to present her- or himself in a positive way. In the extreme, a person may be committed to display an ideal public self and to convey an image of being flawless (Hewitt et al., 2003). Both sources of individual differences, actual behavior and positive self–representation, may be correlated and influence a person's response behavior jointly. For example, respondents who have to hide something (e.g., drug users as opposed to non–drug users) may be more likely to edit their responses. Thus, in view of these observations, it seems most appropriate to conclude that RR methods may reduce but not fully eliminate the influence of response biases.

To facilitate more informative analyses of RR data, this paper proposes statistical models which allow explicitly for the possibility that respondents may edit their responses. Our work builds on and extends a recent study by Böckenholt and van der Heijden (2004a, 2004b)[1] who proposed parametric item–response models for the analysis of RR data. Here, we consider semi–parametric item response models that follow directly from latent–class models for the analysis of RR data (Dayton and Scheers, 1997). These models have the advantage that they do not require presumably arbitrary assumptions about the distribution of the individual–difference parameters in the population of respondents. In addition, the proposed models will be extended to allow for two different types of response biases. The first response bias is assumed to be caused by the

[1] A similar class of models was developed independently by Fox (2005).

condition that, under the considered RR schemes, respondents are forced occasionally to give a self–incriminating answer for something they did not do (Edgell, Himmelfarb, and Duchan, 1982). Since not all respondents may be at ease with incriminating themselves, we allow for the possibility that a subset of the respondents may not follow the RR instructions in this case. Second, some respondents may not trust the privacy protection mechanism and give a negative response regardless of the response determined by the RR scheme or the question asked. A less extreme case is obtained when respondents are willing to be truthful for some of the items (which, presumably, they do not consider to be personal or sensitive) but give a negative response for more serious social norm violations. We consider both types of response biases and include them in the proposed semi–parametric item–response models.

The remainder of the paper is structured as follows. After presenting the RR data that are used to illustrate the proposed approach in more detail, Section 3 describes the semi–parametric item–response models and extensions that allow for different response biases. The results of the data analysis are presented in Section 4. We conclude the paper with several discussion points.

2 The 2000 and 2002 Compliance Surveys

The models to be presented are applied to RR data obtained in two large–scale surveys on non–compliance with conditions that provide social security insurance benefits in the Netherlands. About 12% of the Dutch workforce receives substantial financial support under the Dutch disability act. To remain entitled to the financial benefits, recipients have to comply with regulations about extra income and health-related behavior. These regulations are made operational in simple, non-legal terms with the objective that all

recipients can understand them (Lee, 1993).

Infringements of social security regulations can occur when a person does not comply with regulations about the provision of information to organizations that provide financial benefits to this person. Quantitative measures on the prevalence of this non-compliance was lacking but deemed necessary by the Dutch government for evaluating and monitoring policy interventions. To determine the percentage of recipients who do not comply with the disability rules, the Dutch Ministry of Social Affairs and Employment issued several RR studies.

In 1996, a validation experiment was carried out to investigate the usefulness of RR in surveys on benefit support (van der Heijden, van Gils, Bouts and Hox, 2000; see Lensvelt et al, 2005, for an overview of RR validation studies). Only people who had been identified previously as non–compliant were selected and asked whether they complied with a list of regulations for benefit support. This experiment compared four ways to elicit answers from the participants: direct questions, computer–assisted self–interviewing, and two different RR approaches. Since the actual compliance behavior of the respondents was known, the study allowed to detect false negative responses. Compared to other methods, the RR methods performed best (van der Heijden et al., 2000). Further support for the usefulness of the RR method was obtained in a nationwide large–scale survey in 1998. As a result of the satisfactory performance of the RR method in these studies, a decision was made by the Dutch Ministry to carry out nationwide RR surveys on a two-year basis that investigate infringements of regulations that govern the Disability Benefits Act, the Unemployment Act and the National Assistance Act.

This paper focusses on the 2000 and 2002 surveys about non–compliance with the Disability Benefits Act (Lensvelt–Mulders et al., 2006). In both surveys, regulatory

non–compliance was measured with the following four items:

A. Have you been told by your physician about a reduction in your disability symptoms without reporting this improvement to your social welfare agency?

B. On your last spot–check by the social welfare agency, did you pretend to be in poorer health than you actually were?

C. Have you noticed personally any recovery from your disability complaints without reporting it to the social welfare agency?

D. Have you felt for some time now to be substantially stronger and healthier and able to work more hours, without reporting any improvement to the social welfare agency?

Clearly, these questions are ordered according to their degree of intentional violations of the regulations. A person who does not report the outcome of a medical check–up may also avoid reporting any personally noticed improvements of their health status. In contrast, persons who notice personal improvements may or may not misreport their health status. This ordering suggests that non–compliance with the regulations can be measured along a one–dimensional continuum.

Different RR schemes were used in the two surveys: In 2000, the Kuk (1990) method was applied. Under this scheme, respondents are presented with two stacks of cards. In the left stack, 80% of the cards are red and in the other stack 20% of the cards are red. Then the sensitive question is asked. When the answer is "Yes", the respondents should name the color of the left stack; when it is "No", the respondent should name the color of the right stack. Thus, aside from randomizing the response, this method has the advantage that it does not require the respondent to provide a "Yes" or a "No"

answer but instead to name a color which may be viewed as more neutral. The forced choice (FC) design was adopted for the 2002 survey because it requires smaller sample sizes than the Kuk method to yield comparable levels of precision in estimating the incidence parameters of interest. Under this method respondents were asked to "throw" two computerized dice and to answer "Yes" for the summative outcomes 2, 3 and 4, to answer "No" for the outcomes 11 or 12, and to answer honestly in all other cases.

The responses are reproduced in Table 1. We note that about 40% of the participants respond the color "Black" (533 out of 1308) in 2000 or give a "No"–response (769 out of 1760) in 2002 to all four questions. Although both numbers need to be adjusted for the respective randomization method they are rather substantial.

3 Multivariate Analysis of Randomized–Response Data

Methods for the multivariate analysis of RR data are reviewed by Fox and Tracy (1986), Chaudhuri and Mukerjee (1988) and van den Hout and van der Heijden (2004). This section discusses different approaches for the analysis of multivariate RR data: restricted as well as unrestricted versions of latent class models, semi–parametric, and parametric item–response models. These models have proven useful for the analysis of categorical data in many social and behavioral science applications (Lazarsfeld & Henry, 1968; Lindsay, Clogg, & Grego, 1991). We modify these model to accommodate RR structures and, subsequently, extend them to allow for two different types of response biases which may result when respondents do not follow the randomization scheme. Specifically, we distinguish between respondents who choose not to reveal any information regardless of the question asked and between respondents who do not give a self–incriminating answer when forced to do so by the randomization procedure.

Originally, latent class analysis was proposed for analyzing associations among categorical variables. This method partitions data into homogenous subgroups or classes under the assumption of local independence. Thus, respondents belonging to the same class share the same response probabilities, and associations among the responses within each latent class are assumed to be explained completely by their relationship with the latent grouping variable.

In this basic formulation of the latent class model, no relationship among the different latent classes are imposed. However, if there are some commonalities among the observed variables, it is frequently possible to order the classes along some continuum. By modeling relationships among the latent classes explicitly, a more parsimonious and easier to interpret representation of the data can be obtained. In the limit, when increasing the number of latent classes, heterogeneity can be represented by a latent–trait variable with a continuous distribution in the population.

3.1 Latent Class Randomized–Response Models

The usefulness of randomization schemes depends on their face validity. In the forced choice (FC) design, respondents are given two dice that they can roll themselves. In the Kuk method, respondents are asked to draw cards from two decks that are mixed ostensibly. To account for these randomization procedures, we write the conditional probability of observing person's i response vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ})$ given membership in latent class t as:

$$\Pr(\mathbf{Y}_i = \mathbf{y}_i|t) = \prod_{j=1}^{J} [c + e\lambda_{j|t}]^{y_{ij}} [1 - (c + e\lambda_{j|t})]^{1-y_{ij}}, \tag{1}$$

where $\lambda_{j|t}$ is the conditional probability to answer item j negatively given latent class t, and c and e are constants that are determined by the randomization method. For the

FC scheme, that was used in the 2002 survey[2], we obtain $c = \frac{1}{12}$ and $e = \frac{3}{4}$. Under Kuk's (1990) randomization scheme that was used in the 2000 survey, the probabilities of a red card were $\frac{4}{5}$ and $\frac{1}{5}$ in the two decks, respectively, which leads to $c = \frac{1}{5}$ and $e = \frac{3}{5}$.

Within each latent class, the probabilities of the J responses are independent with unknown and item–specific response probabilities $\lambda_{j|t}$. Denoting the probability of being in latent class t as π_t, the likelihood function can be written as

$$L(\boldsymbol{\lambda}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \sum_{t=1}^{T} \pi_t \Pr(\mathbf{Y}_i = \mathbf{y}_i | t) \tag{2}$$

The parameters of the latent–class RR model can be obtained by a conventional latent–class analysis provided the estimated latent–class specific probabilities, $\gamma_{j|t}$, of the conventional model can be transformed to yield the corresponding probabilities under the randomization scheme by $\lambda_{j|t} = \frac{\gamma_{j|t} - c}{e}$ (Dayton & Scheers, 1997). In general, however, direct maximization of (2) is preferable since the resulting estimates of the $\lambda_{j|t}$'s may be negative by using this transformation.

The latent–class model (2) allows considering structured latent–class specific probabilities that simplify the interpretation of the sensitive behavior under study. We consider two structural hypotheses. The first one postulates the existence of a so–called Guttman scale which assumes that there is no response error after the randomization effect is accounted for (Guttman, 1950). Thus, respondents comply or do not comply with each regulation. For three items, the matrix of item–specific probabilities $\lambda_{j|t}$ takes

[2]The programmer of the FC method changed inadvertently the randomization method such that the actual randomization constants are $c = .0671$ and $e = .7461$ for the 2002 survey.

on the following form

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

where the columns refer to the items ordered according to their response difficulty and the rows to different respondents. For example, in the context of measuring non–compliance behavior the respondent in the first row complies with all three regulation, whereas the respondent in the third row complies with the first regulation only. Since the item–specific probabilities are assumed to be known, only the latent–class probabilities π_t need to be estimated. The second hypothesis specifies that the response probabilities follow the Rasch model (Rasch, 1960). Because of its attractive interpretation, this model has received much attention in the literature. Its relationship to the latent–class model is considered in the next section.

3.2 Semi–Parametric Rasch Randomized Response Model

If there is no interaction between the item effects and membership in one of the classes, the parameters of the latent class model (1) can be constrained to the following additive form

$$\text{logit}(\lambda_{j|t}) = \theta_t - \delta_j. \tag{3}$$

The main effect of subpopulation t is captured by θ_t and the corresponding item effect by δ_j. Because θ_t is independent of the item involved, it is straightforward to interpret differences between subpopulations.

Equation (3) is a latent–class version of a Rasch model (Lindsay, Clogg, & Grego, 1991; Formann, 1992). For fixed T, the Rasch constraint can be tested via a nested log–likelihood ratio test against its unrestricted counterpart (1). The null hypothesis of additivity is rejected if there are subgroups in the population that differ in their views on how personal the items are.

Detailed studies by Lindsay et al. (1991) showed that the number of classes that can be identified under (3) cannot exceed $\frac{J}{2}$, where J is the number of binary items. The same result applies to the RR version of the semi–parametric Rasch model.

3.3 Parametric Rasch Randomized Response Model

A parametric latent–trait version of the Rasch model is obtained by modeling population heterogeneity via a continuous distribution as opposed to a finite set of groups. In some applications, this approach may appear to be more natural in representing variation in a population. By replacing the latent–class with a person–specific parameter, we obtain the RR–Rasch model:

$$\Pr(y_{ij};\theta_i,\delta_j) = c + \frac{e \times \exp(\theta_i - \delta_j)}{1+\exp(\theta_i - \delta_j)}, \tag{4}$$

where θ_i is a person (e.g., *compliance*) parameter. Selecting a parametric family Q such as the Gaussian or logistic, we can write the likelihood function for this model as

$$L(\boldsymbol{\delta}, Q) = \prod_{i=1}^{N} \int \prod_{j=1}^{J} [c + e\Pr(y_{ij};\theta_i,\delta_j)]^{y_{ij}} \; [1-(c+e\Pr(y_{ij};\theta_i,\delta_j))]^{1-y_{ij}} dQ(\theta). \tag{5}$$

A major reason for the popularity of the Rasch model stems from the result that the sum scores of the item responses $\sum_{j=1}^{J} y_{ij}$ are sufficient statistics for the person parameters of the model. This property of the model does no longer hold under the RR design because responses are not exclusively a function of the difference between the item and person

parameter but also a result of the randomization device. Although the sufficient–statistic property is lost, we favor the Rasch structure because it allows testing the hypothesis of no interaction between item and person parameters. This hypothesis is a useful benchmark for understanding how individuals differ in their behavior on the measured dimension.

3.4 Non–Pseudo Incriminators and Self–Protective Respondents

Even when respondents participate actively in the randomization process to protect their privacy, some of them may not be convinced that the protective measures are effective, and, as a consequence, may not follow the randomization scheme and give the type of answer that is requested.

It is beneficial to distinguish among different reasons for not following the instructions of the randomization scheme and to incorporate them explicitly in a statistical model to potentially distinguish between respondents who answer questions truthfully and those who do not. By asking respondents multiple questions about the same domain, one can investigate the impact of this response behavior on model inferences.[3] We distinguish the following two general types of respondents.

First, respondents who are forced to give a potentially self–incriminating answer for something they did not do, may choose not to do so. For example, in a forced–choice study reported by Edgell, Himmelfarb, and Duchan (1982) respondents were asked to say "Yes" when the outcome of a randomizing device is 0 and 1, "No" when the outcome is 8 and 9, and to answer honestly for outcomes between 2 and 7. By

[3]An alternative approach is obtained by asking the same question with different randomization probabilities (Clark & Desharnais, 1998). However, this approach requires the strong assumption that randomization probabilities do not interact with the response behavior (Ljungqvist, 1993).

fixing outcomes of the randomizing design a priori, the investigators found that about 25% of the respondents did not follow the instructions when answering a question on homosexual experiences: They answered "No" although they should have responded "Yes" according to the randomizing device. A follow–up study by Edgell, Duchan & Himmelfarb (1992) showed that this response behavior may depend on the RR method used. Using a different RR method, they could reduce substantially the percentage of respondents who did not comply with the randomization instructions. In the following, we refer to respondents who choose not to incriminate themselves when forced as "non pseudo–incriminators" (NPI, in short)[4].

Second, a certain percentage of participants may not follow the randomization scheme and give a "No"– or "Black"–response regardless of the question asked. Böckenholt and van der Heijden (2004) label this behavior as self–protective (SP) responses. These respondents choose not to reveal any information regardless of whether they are asked to give either a truthful answer or one of the responses determined by the randomization device. Reasons for this behavior can be diverse. Respondents may simply misunderstand or distrust the randomization procedure. Alternatively, they may want to hide their actual behavior by not providing any evidence that could be used to implicate them.

Importantly, SP and NPI response types lead to similar observed frequency distributions when the probability of exhibiting the sensitive behavior of interest is low. In

[4]As a variant of this behavior, respondents may prefer not to incriminate themselves when they are asked to give a truthful response but otherwise provide one of the responses that are determined by the randomization scheme. This response strategy is fairly sophisticated because it implies that respondents are willing to "incriminate" themselves when the response is provided by the randomization but, otherwise, choose to protect themselves.

this case, it may not be possible to distinguish empirically whether respondents give a negative answer regardless of the question asked or whether they do not want to incriminate themselves but otherwise give an honest answer. Either way, respondents edit their answer but the underlying motivation for this behavior (e.g., hiding vs. avoidance of self–incrimination) may be rather different.

Less extreme cases of SP and NPI response types are obtained when respondents are willing to be truthful for some of the items (which, presumably, they do not consider to be personal or sensitive) but give a "No"– or "Black"–response for more serious violations, either because they do not want to incriminate themselves, or because they want to hide their behavior. As in the complete SP and NPI cases, it may not be possible to distinguish between these reasons empirically when the probability of observing the sensitive behavior is low. We refer to the two types of edited responses as partially self–protective (PSP, in short) or partially non–pseudo incriminating (PNPI, in short).

Table 2 summarizes the hypothesized response biases for the Forced Choice and Kuk procedure. The rows of this table refer to the outcome of the randomization procedure and the table entries to the responses under each of the biases. The last column of the table contains unedited responses, that is, responses that are expected when a person follows the randomization instructions and replies truthfully. Thus, when respondents are instruction–compliant, they answer "Yes" when forced to do so or provide the respective color of the card selected from the "Yes"– or "No"–card deck. In contrast, NPI respondents give a "No"– or "Black"–response even when they are forced to say "Yes" or draw a red card from the "No"–card deck. Partially SP and NPI responses are not included in the table because they vary from item to item. For less sensitive items, partially SP or NPI respondents are instruction–compliant. However, for more

sensitive items they may give a "No" or "Black" response regardless of the true answer or when forced to incriminate themselves. Next, we present the likelihood functions for the semi–parametric RR models that allow for partial and complete SP–"No" respondents as well as "non pseudo–incriminators". Modifications for the parametric case are straightforward and are omitted.

NPI Model Letting ρ denote the probability that a person follows the randomization instruction, we interpret its complement $(1-\rho)$ as the probability that a respondent is a "non–pseudo incriminator" (NPI, in short) who does not give "Yes" or "Red" responses when forced to do so. The following likelihood function distinguishes between both groups of respondents:

$$L(\boldsymbol{\lambda}, \pi, \rho) = \prod_{i=1}^{N} (\rho \sum_{t=1}^{T} \pi_t (\prod_{j=1}^{J} [c + e\lambda_{j|t}]^{y_{ij}} [1 - (c + e\lambda_{j|t})]^{1-y_{ij}}) \tag{6}$$
$$+(1-\rho) \sum_{t=1}^{T} \pi_t (\prod_{j=1}^{J} [(1-e) + e\lambda_{j|t}]^{y_{ij}} [e\,(1-\lambda_{j|t})]^{1-y_{ij}})).$$

Note that the model part for truthful responses is the same in both mixture components. The components differ only in terms of the "forced" response part with a probability of 0 for respondents who choose not to "pseudo–incriminate" themselves.

SP-Model If respondents do not trust the RR protection, they may give a "No"– or "Black"–response regardless of the question asked. This response behavior can be captured by a RR model which assumes that a randomly sampled person answers truthfully with probability ϕ and is self–protective with probability $(1-\phi)$:

$$L(\boldsymbol{\lambda}, \pi, \phi) = \prod_{i=1}^{N} (\phi \sum_{t=1}^{T} \pi_t (\prod_{j=1}^{J} [c + e\lambda_{j|t}]^{y_{ij}} [1 - (c + e\lambda_{j|t})]^{1-y_{ij}}) \tag{7}$$
$$+(1-\phi) \prod_{j=1}^{J} \{\Pr(\text{“No”})^{y_{ij}}\ [1 - \Pr(\text{“No”})]^{1-y_{ij}}\}).$$

By decomposing a set of "No"/"Black" responses into self-protective and real ones, the estimates of the measured incidence rates under (7) are higher than under (1). The

crucial feature of the complete SP model (7) is that members of the SP–group do not provide any information about the item parameters. In the reported application, it is specified that participants who decide to give a SP response, select this response with probability 1. This assumption can be relaxed by estimating the probability of a SP response from the data.

A comparison of (6) and (7) shows that both likelihood functions become indistinguishable when the conditional probability of answering negatively ($\lambda_{j|t}$) approaches 1. Thus, although the underlying response mechanisms of (6) and (7) may be rather different, they are difficult to disentangle when most of the respondents do not exhibit the sensitive behavior of interest.

Partial NPI– and SP–Models Less extreme cases of NPI– and SP–models are obtained when respondents choose to give a "No" or "Black" response for more personal or sensitive issues but are instruction–compliant otherwise. For the following likelihood functions, it is assumed that the item ordering in terms of their personal or sensitive nature is known a priori. In our application, this assumption is satisfied because there are external criteria which can be used to assess the severity of non–compliance with the regulations and laws under study. When the items are ordered according to their compliance difficulty, the NFI likelihood function can be written as:

$$
\begin{aligned}
L(\boldsymbol{\lambda}, \pi, \rho) \;=\; & \prod_{i=1}^{N}(\rho_0 \sum_{t=1}^{T} \pi_t (\prod_{j=1}^{J} [c + e\lambda_{j|t}]^{y_{ij}} [1 - (c + e\lambda_{j|t})]^{1-y_{ij}}) \\
& + \sum_{k=1}^{J-1} \rho_k (\sum_{t=1}^{T} \pi_t (\prod_{j=1}^{k} [c + e\lambda_{j|t}]^{y_{ij}} [1 - (c + e\lambda_{j|t})]^{1-y_{ij}} \\
& \times \prod_{j=k+1}^{J} [(1-e) + e\lambda_{j|t}]^{y_{ij}} [e\,(1 - \lambda_{j|t})]^{1-y_{ij}}) \\
& + \rho_J (\sum_{t=1}^{T} \pi_t (\prod_{j=1}^{J} [(1-e) + e\lambda_{j|t}]^{y_{ij}} [e\,(1 - \lambda_{j|t})]^{1-y_{ij}}))),
\end{aligned}
\tag{8}
$$

where $\sum_{j=0}^{J} \rho_j = 1$. Similarly, for the SP model we obtain:

$$
\begin{aligned}
L(\boldsymbol{\lambda}, \pi, \phi) \;=\; & \prod_{i=1}^{N} (\phi_0 \sum_{t=1}^{T} \pi_t (\prod_{j=1}^{J} [c + e\lambda_{j|t}]^{y_{ij}} [1 - (c + e\lambda_{j|t})]^{1-y_{ij}}) \\
& + \sum_{k=1}^{J-1} \phi_k (\sum_{t=1}^{T} \pi_t (\prod_{j=1}^{k} [c + e\lambda_{j|t}]^{y_{ij}} [1 - (c + e\lambda_{j|t})]^{1-y_{ij}} \\
& \times \prod_{j=k+1}^{J} \{\Pr(\text{“No”})^{y_{ij}} \; [1 - \Pr(\text{“No”})]^{1-y_{ij}}\})) \\
& + \phi_J \prod_{j=1}^{J} \{\Pr(\text{“No”})^{y_{ij}} \; [1 - \Pr(\text{“No”})]^{1-y_{ij}}\}),
\end{aligned} \tag{9}
$$

where $\sum_{j=0}^{J} \phi_j = 1$. Because of the requirement that the ordering of the items needs to be known, we constrain the $\lambda_{j|t}$ parameters to take on the additive form given by (3). The Guttman model is not considered in this context because its deterministic response structure prevents the identification of the ρ_j or ϕ_j parameters.

For each of the response–bias models the probability of being a NPI or SP type can be made person–specific provided covariates are available that facilitate the identification of these parameters. For example, under the NPI model the probability that a randomly selected respondent follows the randomization instructions can be written as

$$\rho_i = \frac{1}{1 + \exp(-\mathbf{x}_i' \beta)},$$

where x_i refers to a set of available covariates that are specific to person i.

4 Analysis of Non–Compliance Survey Data

The previous section presented the latent–class as well as semiparametric and parametric Rasch models for the analysis of multivariate RR data. In addition, we considered partial and complete NPI and SP response biases. Table 3 provides a summary of the application of these models to the non–compliance survey data collected in 2000 and 2002. This

table reports an overall likelihood ratio χ^2–statistic and the corresponding degrees of freedom for the various models.

The fit improvement from the one– to the two–class models (e.g., 212.0 to 43.7 for the 2000 data) suggests that there is non–compliance heterogeneity in the data. Table 3 contains the fit statistics of the latent–class models that include the constraints imposed by the RR scheme. Two–class models without these constraints provide a much better fit with χ^2–statistics equal to 20.2 and 10.4 for the 2000 and 2002 survey data, respectively. However, a transformation of the class-specific probabilities under the randomization scheme leads to negative values indicating that these models are not applicable in our context (for a related discussion on this topic, see van den Hout and van der Heijden, 2002).

From Table 3, we conclude that only models which allow for response biases provide a satisfactory description. In particular, the NPI/SP–Guttman models fit best for the 2000 survey data and the parametric NPI/SP–Rasch models fit best for the 2002 survey data. Because the reported level of non–compliance is low (see Table 4 for point estimates), the NPI and SP model versions yield rather similar fit statistics in this application. The fit statistics of the partial SP and NPI models differ to a greater extent. However, because the fit improvements are small in comparison to the more parsimonious complete SP–Guttman or parametric NPI–Rasch models, we favor the latter models for subsequent analyses.

Table 4 contains the non–compliance estimates and their 95% bootstrap confidence intervals obtained both under a univariate analysis without any response bias corrections and under the Guttman and parametric Rasch models with either SP or NPI respondents. The compliance estimates under the univariate model (referred to as one–class

model in Table 4) are based on the RR–corrected marginal proportions (Chaudhuri & Mukerjee, 1988). We note that for the 2002 survey data the compliance estimate for Item A as well as the lower limits of the 95% bootstrap confidence intervals for Items A and B are negative. These results suggest that the actual incidence of non–compliance is underestimated with this approach.

The SP and NPI versions of the Guttman and parametric Rasch models arrive at similar confidence intervals for the estimated compliance statistics. However, this result is obtained under different assumptions about the underlying response process. According to the NPI models about 17% of the respondents do not incriminate themselves when asked to do so but tell the truth otherwise. In contrast, according to the SP models about 13% of the respondents give a negative response regardless of the questions asked. Both models yield non–compliance estimates that are almost twice as large as the ones obtained when no response biases are assumed. But even the response–bias corrected estimates may be too low if some or all of the SP respondents are non–compliant as well.

Importantly, the non–compliance estimates obtained by Kuk methods are substantially higher than the ones obtained by the Forced Choice methods. These differences may be a result of the two–year interval between the surveys or because of the higher privacy protection provided by the Kuk method. Since no governmental intervention took place to reduce the incidence of non–compliance (Faas, 2005; personal communication), we favor the latter possibility. But we hasten to add that this conjecture needs to tested in further research. In particular, it would be valuable to determine the degree to which respondents are aware of the differences in their privacy protection provided by the Kuk and Forced Choice methods.

5 Concluding Remarks

The investigation of sensitive topics is of great interest in many areas, but empirical progress has been hampered by a lack of statistical techniques that both facilitate the analysis of structural relationships among multiple responses and allow for the possibility that respondents may edit their responses. The proposed modeling framework addresses this gap and thus sets the stage for more efficient and informative analyses than have been possible so far. In the reported application, it was shown that the analysis of multiple RR items can provide useful insights about individual differences in compliance behavior and the incidence of response biases. Moreover, the consideration of both aspects of the RR data proved crucial in estimating the true compliance rate in the population of interest.

Clearly, RR methods may not eliminate response biases completely. The estimated incidence of edited (i.e., self–protective or non–incriminating) responses in both data sets is sufficiently large to suggest that previous RR studies in this area which did not correct for response biases underestimated substantially the true incidence of the behavior in question. In view of these results, it seems promising to investigate further possible determinants of response biases and their relationship with the behavior in question.

We also note that SP as well as NPI response types can yield comparable fits when the probability of exhibiting the sensitive behavior of interest is low. In this case, two rather different explanations may not be distinguishable on the basis of RR data alone. Respondents may either edit their responses because they want to hide their non–compliant behavior or because they do not want to be wrongly accused of something they did not do. If hiding is the main motivation, the non-compliance estimates in Table 4 are biased downwards. Covariates that measure attitudes of the respondents

towards the sensitive domain of interest may be helpful in eliminating one of these two explanations. These covariates can be incorporated easily in the proposed RR models to explain variability in the person parameter as well as membership in one of the mixture components.

Although these first steps in improving the interpretation and analysis of RR data are promising, more experimental as well as modeling work remains to be done. We suggest three avenues of future research. First, in their present form, the proposed RR can account for respondents who edit their responses in certain ways but they do not allow for possible biases at other stages of the response process. Person–specific interpretations of response categories, individual differences in recall ability, and perceptual biases (Mick, 1996) are possible factors that ultimately should be incorporated in a structural model of response behavior. Second, extensions to different data types are of much interest. Although it is straightforward to consider ordinal RR models for the analysis of rating data (Rosenberg, 1980), it is less clear how to model SP, NPI or other response types (de Jong, Steenkamp, Fox & Baumgartner, 2005). Behavioral studies are needed to provide guidance on the sources and forms of these different response biases. Third, the finding that the two surveys yielded substantially different compliance estimates suggests that the RR method itself may influence response behavior. One possible reason for the higher non–compliance estimates under the Kuk than under the Forced Method is that Kuk offers higher privacy protection. This difference may have reduced any editing of the responses and hence increased the truthfulness of the self–reports. It seems thus promising to conduct experimental studies that investigate the degree to which respondents distinguish and react positively to the different levels of privacy protection offered by these RR methods. These studies may also prove useful in validating and

further exploring the proposed specifications of the SP and NPI response types.

In conclusion, our work showed that RR data can be modeled readily with standard psychometric tools that are appropriately modified to account for response biases. While currently limited to binary data, the proposed set of models should be of much use in studies on personal and sensitive issues, especially when some of the respondents are likely to edit their responses. Although effective, privacy protection as provided by RR methods does not work to the degree as originally was perhaps hoped for. RR methods yield more valid responses but do not fully eliminate response biases. Still, as was shown by the analysis of the two survey data sets, RR data can be modelled in informative ways even if response biases are present in the data.

References

Baumgartner, H. & Steenkamp, J-B. E.M. 1992. Response styles in marketing research: A cross–national investigation. *Journal of Marketing Research*, 38, 14356.

Benitez-Silva, H., Buchinsky, M., Chan, H.-M., Cheidvasser, S. & Rust, J. 2004. How large is the bias in self–reported disability? *Journal of Applied Econometrics, 19*, 649-670.

Blair, E., Sudman, S., Bradburn, N. M., & Stocking, C. 1977. How to ask questions about drinking and sex: Response effects in measuring consumer behavior. *Journal of Marketing Research, 14*, 316–321.

Böckenholt, U. & van der Heijden, P. G. M. 2004a. *Randomized–Response Models for Measuring Noncompliance: Risk–Return Perceptions, Social Influences, and Self-Protective Responses.* Unpublished manuscript. McGill University.

, U. & van der Heijden, P. G. M. 2004b. 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,* (pp. 106-110). Florence, Italy.

Bound, J., Brown, C., & Mathiowetz, N. 2001. Measurement error in survey data. In J. J. Heckman and E. Leamer (Eds.). *Handbook of Econometrics*, 5 (pp. 3705–3843). Elsevier: Amsterdam.

Chandon, P., Morwitz, V. G. & Reinartz, W. J. 2005. Do intentions really predict behavior? Self-generated validity effects in survey research. *Journal of Marketing*, 69, 114.

Chaudhuri, A., and Mukerjee, R. 1988, *Randomized Response: Theory and Techniques.* Marcel Dekker: New York

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

Dayton, C.M. and N.J. Scheers 1997. Latent class analysis of survey data dealing with academic dishonesty. In: J. Rost and R. Langeheine (Eds.). *Applications of Latent Trait and Latent Class Models in the Social Sciences* (pp. 172–180). Waxmann Muenster: New York.

de Jong, M.G., Steenkamp, J.B.E.M., Fox, J.-P., & Baumgartner, H. 2005. *Using Item Response Theory to Measure Extreme Response Style in Marketing Research: A Global Investigation.* Unpublished manuscript. Tilburg University. The Netherlands.

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.

Edgell, S. E., Duchan, K. L. & Himmelfarb, S. 1992. An empirical test of the unrelated question randomized response techniques *Bulletin of the Psychonomic Society*, 30, 153–156.

Elffers, H., van der Heijden, P.G.M. and Hezemans, M. 2003. Explaining regulatory non-compliance: A survey study of rule transgression for two Dutch instrumental laws, applying the randomized response method. *Journal of Quantitative Criminology*, 19, 409-439.

Formann, A. K. 1992. Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association, 87*, 476-486.

Fox, J.A., and Tracy, P.E. 1986 *Randomized Response: A Method for Sensitive Surveys.* Sage: Newbury Park

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

Guttman, L. 1950. The basis for scalogram analysis. In *Measurement and Prediction, Studies in Social Psychology in World War II, Volume IV*, (pp. 6090). University Press: Princeton.

Hewitt, P. L., Flett, G. L., Sherry, S. B., Habke, M., Parkin, M., Lam, R. W., McMurtry, B., Ediger, E., Fairlie, P., & Stein, M. B. 2003. The interpersonal expression of perfection: Perfectionistic self-representation and psychological distress. *Journal of Social and Personality Psychology*, 84, 1303-1325.

Hsiao, C. & Sun, B.-H. 1998. Modeling survey response bias with an analysis of the demand for an advanced electronic device. *Journal of Econometrics, 89*, 15-39.

Hsiao, C., Sun, B. H., & Morwitz, V. G. 2002. The role of stated intentions in new product purchase forecasting. In T. B. Fomby & R. C. Hill (Eds.). *Advances in Econometrics, Vol.16*, (pp. 11–28). Elsevier Science: Oxford.

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

Lamb, C. W. & Stem, D. E. 1978. An empirical validation of the randomized response technique. *Journal of Marketing Research*, 15, 616–621.

Lara, D., Strickler, J., Olavarrieta, C.D., & Ellerston, C. 2004. Measuring induced abortion in Mexico: A comparison of four methodologies *Sociological Methods and Research, 32*, 529-558.

Lazarsfeld, P. F., & Henry, N. W. 1968. *Latent Structure Analysis.* Houghton Mifflin: Boston.

Lee, R. M. 1993. *Doing Research on Sensitive Topics.* Sage: London.

Lensvelt, G., Hox, J.J., Van der Heijden, P.G.M., and Maas, C. 2005. Meta-analysis of Randomized Response Research: 35 Years of Validation. *Sociological Methods and Research, 33*, , 319–348.

Lensvelt–Mulders, G., van der Heijden, P. G. M., Laudy, O., and van Gils, G. 2006. A validation of a computer–assisted randomized response survey for measuring fraud in social security. *Journal of the Royal Statistical Society, Series A, 169*, 305-318.

Lindsay, B., Clogg, C. C., & Grego, J. 1991. Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86, 96-107.

Ljungqvist, L. 1993. A unified approach to measures of privacy in randomized response models: A utilitarian perspective. *Journal of the American Statistical Association, 88*, 97–103.

McFadden, D. L., Bemmaor, A. C., Caro, F. G., Dominitz, J., Jun, B-H., Lewbel, A., Matzkin, R.L., Molinari, F., Schwarz, N., Willis, R. J., Winter, J. K. 2005.

Statistical Analysis of Choice Experiments and Surveys. *Marketing Letters*, 16, in press.

Mick, D. G. 1996. Are studies of dark side variables confounded by socially desirable responding? The case of materialism. *Journal of Consumer Research*, 23, 106-121.

Mittal, V. & Kamakura, W. A. 2001. Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics. *Journal of Marketing Research*, 38, 131–142

Rasch, G. 1980. *Probabilistic Models for Some Intelligence and Attainment Tests.* The University of Chicago Press: Chicago. (Original published 1960, Copenhagen: The Danish Institute of Educational Research)

Rosenberg, M. J. (1980). Data analysis by a randomized response technique for statistical disclosure control. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 311–316.

Schwarz, N. 2003. Self-reports in consumer research: The challenge of comparing cohorts and cultures. *Journal of Consumer Research*, 29, 588–594.

Tourangeau, R., Rips, L. J., & Rasinski, K. 2000. *The Psychology of Survey Response.* New York and Cambridge, Cambridge University Press: London.

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, pp. 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 data. *Sociological*

Methods and Research, 32, 310-336.

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. *Sociologial 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.

Table 1: Frequency Distribution for Health Items

Items				Surveys	
A	B	C	D	2000	2002
0	0	0	0	43	18
0	0	0	1	22	15
0	0	1	0	10	26
0	0	1	1	34	34
0	1	0	0	20	30
0	1	0	1	31	32
0	1	1	0	40	31
0	1	1	1	93	137
1	0	0	0	30	31
1	0	0	1	29	40
1	0	1	0	40	56
1	0	1	1	91	132
1	1	0	0	60	78
1	1	0	1	86	142
1	1	1	0	146	189
1	1	1	1	533	769

Note: The item labels are given in Section 2. The 2000 and 2002 surveys used Kuk and FC randomized–response methods, respectively. A "1" indicates a negative (e.g., "No"/"Black") response.

Table 2: Response Biases under Kuk and Forced Choice Instructions

Randomization Outcome	Response Types		
Forced Choice	Instruction–compliant	Non Pseudo–Incriminators	Complete SP
Forced Yes	"Yes"	"No"	"No"
Forced No	"No"	"No"	"No"
Forced Truth	"Truth"	"Truth"	"No"
Kuk			
"No"-stack & "Black"	"Black"	"Black"	"Black"
"No"-stack & "Red"	"Red"	"Black"	"Black"
"Yes"-stack &"Black"	"Black"	"Black"	"Black"
"Yes"-stack &"Red"	"Red"	"Red"	"Black"

Note: Partially self–protective respondents are instruction–compliant for less sensitive items (fourth column) and become self–protective respondents (third column) for more sensitive items. Similarly, partial NPI respondents do not incriminate themselves for more sensitive items.

Table 3: Goodness–of–fit statistics of multivariate RR models

Models	2000	2002	df
Instruction–Compliant			
1–LC	212.0	124.0	11
2–LC	43.4	35.4	6
LC–Guttman	39.9	47.8	11
LC–Rasch	43.7	35.4	9
Normal–Rasch	47.2	39.0	10
Complete Self-Protector (CSP)			
1–LC	83.0	42.3	10
2–LC	12.7	13.5	5
LC–Guttman	12.5	30.9	10
LC–Rasch	16.5	13.9	8
Normal–Rasch	16.4	14.9	9
Partial Self–Protector (PSP)			
LC–Rasch	6.2	13.2	5
Normal–Rasch	6.2	12.9	6
Non Pseudo–Incriminator (NPI)			
1–LC	105.6	46.0	10
2–LC	13.8	10.2	5
LC–Guttman	12.2	31.7	10
LC–Rasch	16.8	12.9	8
Normal–Rasch	16.7	14.3	9
Partial NPI			
LC–Rasch	12.6	6.9	5
Normal–Rasch	10.6	6.9	6

Table 4: Non–Compliance Estimates and 95% Bootstrap Confidence Intervals

2000 Survey	1–LC	LC–Guttman (NPI)	LC–Guttman (CSP)
A.	.04 (.01, .08)	.08 (.06, .10)	.08 (.06, .10)
B.	.05 (.01, .09)	.09 (.07, .12)	.10 (.07, .13)
C.	.08 (.04, .11)	.13 (.10, .17)	.13 (.10, .17)
D.	.16 (.12, .20)	.22 (.18, .27)	.24 (.19, .30)
NPI/CSP%	–	16 (10, 21)	12 (8, 17)
2002 Survey	1–LC	Normal–Rasch (NPI)	Normal–Rasch (CSP)
A.	-.00 (-.03, .02)	.04 (.01, .08)	.04 (.01, .07)
B.	.02 (-.01, .04)	.06 (.03, .10)	.06 (.03, .09)
C.	.04 (.02, .07)	.09 (.06, .13)	.09 (.06, .13)
D.	.10 (.07, .13)	.14 (.11, .19)	.15 (.11, .19)
NPI/CSP%	–	17 (8, 30)	13 (7,20)

Note: The item descriptions are given in Section 2.

NPI: Non–pseudo incriminiator.

CSP: Complete self-protective respondent.