PSYCHOMETRIKA
2007
DOI: 10.1007/s11336-005-1495-y

ITEM RANDOMIZED-RESPONSE MODELS FOR MEASURING NONCOMPLIANCE: RISK-RETURN PERCEPTIONS, SOCIAL INFLUENCES, AND SELF-PROTECTIVE RESPONSES

ULF BÖCKENHOLT

McGILL UNIVERSITY

PETER G.M. VAN DER HEIJDEN

UTRECHT UNIVERSITY

Randomized response (RR) is a well-known method for measuring sensitive behavior. Yet this method is not often applied because: (i) of its lower efficiency and the resulting need for larger sample sizes which make applications of RR costly; (ii) despite its privacy-protection mechanism the RR design may not be followed by every respondent; and (iii) the incorrect belief that RR yields estimates only of aggregate-level behavior but that these estimates cannot be linked to individual-level covariates. This paper addresses the efficiency problem by applying item randomized-response (IRR) models for the analysis of multivariate RR data. In these models, a person parameter is estimated based on multiple measures of a sensitive behavior under study which allow for more powerful analyses of individual differences than available from univariate RR data. Response behavior that does not follow the RR design is approached by introducing mixture components in the IRR models with one component consisting of respondents who answer truthfully and another component consisting of respondents who do not provide truthful responses. An analysis of data from two large-scale Dutch surveys conducted among recipients of invalidity insurance benefits shows that the willingness of a respondent to answer truthfully is related to the educational level of the respondents and the perceived clarity of the instructions. A person is more willing to comply when the expected benefits of noncompliance are minor and social control is strong.

Key words: randomized response, item response theory, cheating, concomitant variable, sensitive behavior, efficiency.

1. Introduction

Is it possible to measure noncompliance with rules and sanctions that govern public life? This paper investigates this question in the context of a recent series of surveys requested by the Dutch government to better understand and measure noncompliance behavior. The growing political interest in The Netherlands is a result of two major disasters that caused the death of many individuals. In May 2000, a firework explosion destroyed part of Enschede, a medium-sized city in the east of The Netherlands, because rules for storage of fireworks were not followed. Later in the year, on New Year's Eve, ten people died and over 130 people were injured after a fire swept through a cafe packed with teenagers because fire regulations were not followed. Of course, interest in noncompliance is not restricted to The Netherlands. For example, the IRS conducts surveys regularly to predict taxpayers' willingness to comply with tax laws.

The authors are grateful to the reviewers whose suggestions helped to improve the clarity of 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.

Requests for reprints should be sent to Ulf Böckenholt, Faculty of Management, McGill University, 1001 Sherbrooke Street West, Montreal, QC H3A 1G5, Canada. E-mail: ulf.bockenholt@mcgill.ca; or to Peter van der Heijden, Department of Methodology and Statistics, Utrecht University, P.O. Box 80.140, 3508 TC Utrecht, The Netherlands. E-mail: p.vanderheijden@fss.uu.nl

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

Recent meta-analyses have shown that RR methods can outperform significantly more direct ways of asking sensitive questions (Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005). Importantly, the relative improvements in the validity increased with the sensitivity of the topic under investigation. However, despite these positive results, RR is not used often in practical applications for a number of reasons. First, RR studies are expensive. Since the efficiency of RR estimators is low, larger sample sizes are needed to obtain estimates with a precision that is comparable to the one obtained from direct questions. Moreover, because, frequently, the true compliance rate is unknown, the extent to which the loss in efficiency is counterbalanced by a reduction in response bias cannot be assessed a priori. In fact, the RR method has been critiqued because it forces respondents to give a potentially self-incriminating answer for something they did not do, with the result that some respondents do not follow the RR instruction. For example, in a forced-choice study reported by Edgell, Himmelfarb, and Duncan (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 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 further reason that limits applications of RR methods is the lack of statistical methods for RR data. Although a number of books have been published on this topic (see Chaudhuri & Mukerjee, 1988; Fox & Tracy, 1986), much work remains to be done.

This paper will address these three issues in the following way. First, we apply appropriately modified versions of item response models for the analysis of multiple RR items (Böckenholt & van der Heijden, 2004). A similar class of models was developed independently by Fox (2005). The reported application of these models show that they are well suited to investigate individual differences in compliance. We refer to the resulting class of models as item randomized-response (IRR) models. Since the precision in estimating compliance differences is a function of the number of RR items per respondent, more precise measures of compliance can be obtained in multiple- than in single-item studies for equal sample sizes. Second, mixture versions of the IRR models are developed to allow for respondents who do not follow the RR instructions. Thus, one mixture component consists of respondents who answer RR items by following the RR design and the other component consists of respondents who do not follow the RR design by saying "no" to each RR item, irrespective of the outcome of the randomizing device. By allowing for the possibility that not all respondents may follow the RR instructions, we obtain substantially higher estimates of noncompliance than obtained with current RR methods. We also extend this model family to the case of multiple compliance domains to investigate whether respondents who are not compliant in one domain are also more likely to be less compliant in other domains. It is shown that this extension is of much importance in the reported applications. Third, to model the probability of noncompliance and factors that influence or moderate (the extent of) noncompliance, we discuss how to include covariates with respect to both the individual compliance parameter and the membership probabilities for the mixture components. This third part of our work builds on and extends previously proposed latent class and logistic regression models for RR data (Dayton & Scheers, 1997; Maddala, 1983; Scheers & Dayton, 1988; van den Hout & van der Heijden, 2002, 2004).

The remainder of the paper is structured as follows. In Section 2 we describe the data in more detail and in Section 3 we propose the IRR models and investigate their properties. Section 4 contains the results from the data analyses. We conclude the paper with several discussion points.

2. The 2002 and 2004 Compliance Surveys about Invalidity Insurance Benefits

Dutch employees must be insured under the Sickness Benefit Act, the Unemployment Insurance Act, the Health Insurance Act, and the Invalidity Insurance Act. Under each of these acts, a (previously) employed person is eligible for financial benefits provided certain conditions are met. Our focus is on noncompliance with rules that have to be followed for receiving benefits under the Invalidity Insurance Act (IIA, hereafter). In a workforce of approximately seven million people, over 800,000 draw benefits under the IIA alone. The benefit can amount to as much as 70% of the recipient's last regular income. Noncompliance with IIA rules can become a fraud if it is caused by purposeful behavior and it is not a result of ignorance about the prescribed rules.

To remain entitled to IIA benefits, recipients have to comply with regulations about extra income and health-related behavior. These regulations are made operational in simple, nonlegal terms with the objective that all recipients can understand them (Lee, 1993). There is much interest in measuring the extent of noncompliance of IIA recipients. After, in 1996, the usefulness of RR methods for measuring noncompliance in comparison to other data collection approaches was tested and established (see Figure 1, van der Heijden, Van Gils, Bouts, & Hox, 2000), in 1998 a first pilot was carried out, followed by three waves in the years 2000, 2002, and 2004. The 2006 wave is currently underway. We focus here on the results of the 2002 and 2004 surveys which used the forced choice design as an RR method. In total, 1760 and 830 IIA recipients participated in these two studies, respectively. For details on the design of the 2002 study we refer to Lensvelt-Mulders, van der Heijden, Laudy, and Van Gils (2006).

In the forced choice (FC) design adopted by both surveys, respondents were asked for each item to click on two electronic 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 instruction provided to the respondents can be found in the Appendix to this paper. The following analyses focus on six RR questions, four of which are health, and the remaining two are work related. The health questions are:

1. Have you been told by your physician about a reduction in your disability symptoms without reporting this improvement to your social welfare agency?
2. On your last spot-check by the social welfare agency, did you pretend to be in poorer health than you actually were?
3. Have you noticed personally any recovery from your disability complaints without reporting it to the social welfare agency?
4. 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?

The work-related questions are:

1. In the last 12 months have you moonlighted while receiving your IIA benefits?
2. In the last 12 months have you taken on a small job alone or together with your friends that you got paid for without informing the social welfare agency?

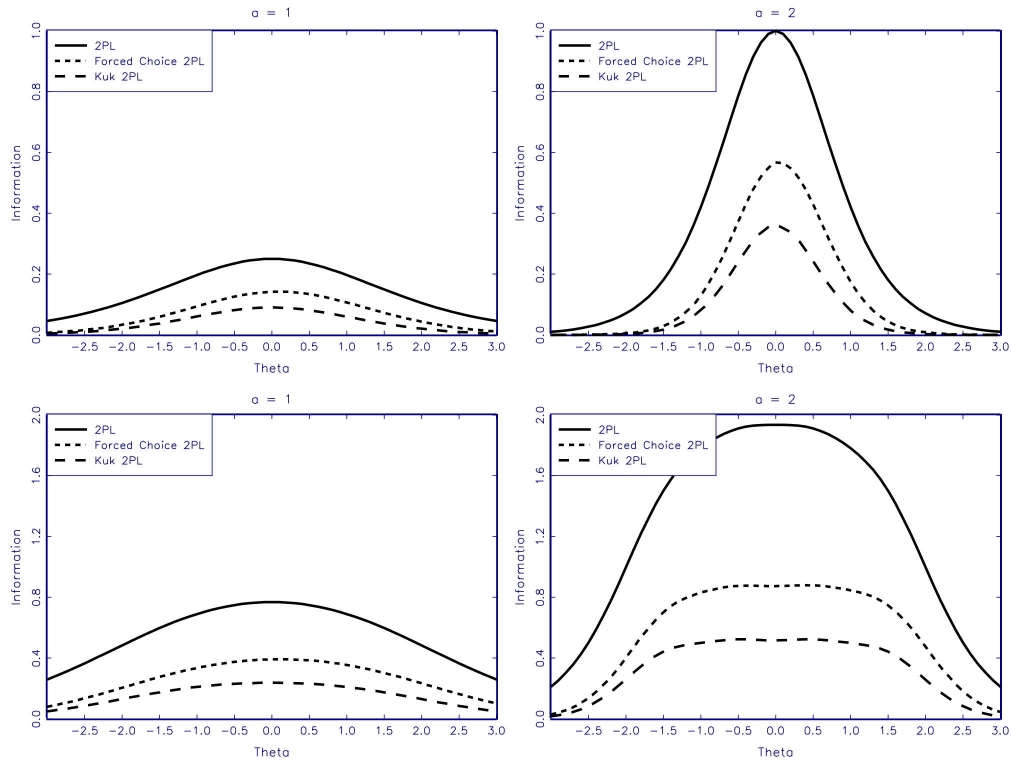

FIGURE 1.
Item (top panel) and test (bottom panel) information functions for nonrandomized and randomized two-parameter logistic (2PL) item response models.

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.

The 2002 and 2004 IIA surveys also contained questions that were hypothesized to account for individual differences in compliance behavior. Although, currently, there is no theoretical framework to predict and explain fully compliance behavior, a number of factors have been shown to account for parts of the individual differences. Most prominently, rational choice approaches (Becker, 1968; Weber, 1997) argue that a person's noncompliance behavior with the regulations is a function of the perceived risks and benefits of breaking the rules. Only when the risks outweigh the benefits, a person may choose to follow the stated regulations. Risk factors may include such factors as the likelihood of being detected, the certainty of a sanction when detected, and the severity of any sanctions. Attitude–Behavior theories (Fishbein & Aijzen, 1975; Eagley & Chaiken, 1993) are also important because they emphasize the acceptability of the rules that a person is asked to comply with and the role of social influences via norms and reactions to noncompliance by friends and neighbors in a person's decision to comply. Thus, according to this approach noncompliance may not be solely a function of a person's perception of perceived risks and expected benefits but also determined by the perceived norms about the appropriateness of the selected behavior.

Much less is known about factors that may influence a person's motivation to follow the RR instructions. As will be shown later in the next section, one of the advantages of asking multiple questions about the same domain is that one can both test whether respondents are truthful in their response behavior and study which factors may influence a person in not being compliant with the RR instructions. Using the survey data, we investigate whether such diverse factors as the clarity of the instruction, educational level of the respondents, and attitudes towards the randomizing scheme may predict whether a person is more likely to not answer truthfully.

The attitudinal and risk-return variables included in the two surveys are based on the so-called Table-of-Eleven (Elffers, van der Heijden, & Hezemans, 2003) and considered the following nine factors (with explanations given in parentheses):

(1) Acceptance (acceptability of IIA rules);
(2) Clarity (lack of knowledge about and perceived clarity of rules);
(3) Benefits (costs and benefits associated with compliance and noncompliance);
(4) Social control (anticipated reaction by family and friends in the case of noncompliance);
(5) Law abidance (general norm conformity with respect to laws and authorities);
(6) Control (subjective probability of being investigated as part of a routine inspection);
(7) Detection (subjective probability of detecting noncompliance given that a noncompliant case is checked);
(8) Sanction certainty (subjective probability that a case will be prosecuted and sanctioned, once noncompliance has been detected by the agency); and
(9) Sanction severity (degree to which a convicted rule transgressor suffers under the sanction).

The first five factors were measured on a five-point rating scale with labels ranging from "strongly agree" to "strongly disagree." For the other factors, a five-point rating scale was used with labels ranging from "very high" to "very small."

Note that factors (6) to (9) focus on the activities of the regulation-enforcing agency to induce compliance. It is of much interest to determine whether these factors are influential in a person's decision to follow regulatory laws. Factors were measured by one or two questions, the responses of which were averaged in the reported analyses.

3. Item Randomized-Response Models

Item-response models (van der Linden & Hambleton, 1997) are well suited for studying how individuals differ in their compliance behavior by ordering respondents on a latent continuum that represents their level of compliance. In this section, we discuss the necessary modifications to make these models suitable for the analysis for RR data. The resulting model class is referred to as item randomized-response (IRR) models (Böckenholt & van der Heijden, 2004; Fox, 2005). In subsequent sections, we discuss the information loss caused by the randomization scheme and model estimation issues.

Because typically the number of items is small in an RR study, the one- or two-parameter logistic model may be well suited to measure individual differences in compliance behavior. Under the two-parameter logistic (2PL) model (Birnbaum, 1968), the probability that item j is answered affirmatively by person i is written as

$$\Pr(x_{ij} = 1) = \Pr_j(\theta_i) = \frac{1}{1 + \exp(-\alpha_j(\theta_i - \gamma_j))}, \tag{1}$$

where α_j and γ_j are parameters characterizing the item-response function. When $\alpha_j = \alpha = 1$, the well-known Rasch model is obtained (Rasch, 1960). The person parameter θ_i may be specified to follow some known distribution or its distribution can be estimated from the data (Lindsay, Clogg, & Grego, 1991). Frequently, a survey consists of H item sets, each of which is measuring a different unidimensional aspect of compliance behavior. It may be of much interest to understand the relationships among these different measures. Our approach is to consider multiple θ_{hi} ($h = 1, \ldots, H$) that may be correlated in the population of test takers. If the correlation among the θ's is substantial, significant efficiency gains can be expected when the IRR models for the item bundles are estimated jointly.

Under the previously explained FC response format, respondents answer "yes" or "no" by chance with probabilities $\frac{1}{6}$ and $\frac{1}{12}$, respectively. As a result, we obtain

$$\Pr^{(\mathrm{FC})}(x_{ij} = 1) = \frac{1}{6} + \frac{3}{4}\left(\frac{1}{1 + \exp(-\alpha_j(\theta_i - \gamma_j))}\right),$$

and

$$\Pr^{(\mathrm{FC})}(x_{ij} = 0) = \frac{1}{12} + \frac{3}{4}\left(1 - \frac{1}{1 + \exp(-\alpha_j(\theta_i - \gamma_j))}\right).$$

More generally, we consider response models of the form

$$\Pr^{(\mathrm{RR})}(x_{ij} = 1) = c + \frac{e}{1 + \exp(-\alpha_j(\theta_i - \gamma_j))} = c + e \Pr{}_j(\theta_i), \tag{2}$$

where the constants c and e are determined by the randomization method. For the FC scheme we obtain $c = \frac{1}{6}$ and $e = \frac{3}{4}$. A number of other randomization methods can be shown to be special cases of this parametrization. For example, under Kuk's (1990) randomization scheme, respondents are asked to select a card from two packs. Each pack of cards corresponds to one of the response categories and contains cards of two colors only. The respondent is to report the color (red or black, say) of the card of the true response category. If the probabilities of a red card are $\frac{4}{5}$ and $\frac{1}{5}$ in the two decks, respectively, $c = \frac{1}{5}$ and $e = \frac{3}{5}$. For Warner's (1965) method, with two questions A and B, with one being the negation of the other and the probability of answering question A is .7, we obtain $c = .3$ and $e = .4$.

The considered family of IRR models is related to the well-known three-parameter logistic model,

$$\Pr(x_{ij} = 1) = \epsilon + \frac{1 - \epsilon}{1 + \exp(-\alpha_j(\theta_i - \gamma_j))}, \tag{3}$$

where ϵ is a so-called guessing parameter (van der Linden & Hambleton, 1997). This model is used in educational testing applications to account for the possibility that (low-ability) respondents may not know the correct answer to a question but guess it with a probability of success equal to the value of ϵ. There is no guessing in the RR context but instead the randomization procedure introduces "Yes" or "No" answers with known probabilities that are captured by the constants c and e.

3.1. Test and Item Information Loss under Randomization

Assuming truthful responses, we can judge the information loss of an item or a test under different randomization schemes by transforming the item's response functions into information functions (Birnbaum, 1968). Information functions allow quantifying the contribution of single or multiple items for estimating a person parameter θ. The test information function which is defined as the sum of item information functions can be written as

$$I(\theta) = \sum_{j=1}^{J} \frac{[(\delta/\delta\theta)(c + e\Pr_j(\theta))]^2}{[c + e\Pr_j(\theta)][1 - (c + e\Pr_j(\theta))]},$$

where $\Pr_j(\theta)$ is given by the 2PL model (1). In general, the information function for any item score is inversely related to the squared length of the asymptotic confidence interval for estimating

θ from this score. Since information functions provide a straightforward approach to assess the information loss caused by a randomization scheme, it is instructive to compare them across different randomization schemes.

Consider the top panels of Figure 1, which for θ values between ± 3 display the item information functions of the Kuk model (with $c = \frac{1}{5}, e = \frac{3}{5}$), the FC model (with $c = \frac{1}{6}, e = \frac{3}{4}$), and their nonrandomized counterpart (with $c = 0, e = 1$). The top left panel is obtained for item information functions with location parameter $\gamma_1 = 0$ and item discrimination parameter $\alpha_1 = 1$, and the top right panel is obtained for the parameter pair ($\gamma_1 = 0,\ \alpha_1 = 2$). In addition to demonstrating the informational benefit of a more discriminating item, the plots illustrate the strong ordering of the information functions: The Kuk method provides much less information about the person parameter θ than the FC method which in turn is less informative than the nonrandomized IRT model. For example, two items are needed under the FC method and three under the Kuk approach to obtain comparable precision in estimating θ as given by a single item in the nonrandomized case.

Not surprisingly, the loss in information becomes more substantial when considering multiple items. As an illustration, consider the test functions displayed in the bottom panel of Figure 1. Here the item location parameters are specified as ($\gamma_1 = -1.5, \gamma_2 = -.5, \gamma_3 = .5, \gamma_4 = 1$). The item discrimination parameter in the bottom left and right panels is $\alpha = 1$ and $\alpha = 2$, respectively. We note that the FC method is substantially more informative than the Kuk method. Within the range of the specified item locations, the gain is more than 50% indicating that the privacy protection under the Kuk method is substantially higher than under the FC approach. Clearly, the item and test information functions simplify the comparison of different randomization methods and provide a convenient approach toward computing the number of items needed to obtain a desired level of precision in estimating θ.

3.2. Likelihood Functions for Item Randomized-Response Models

A baseline RR model assumes that the answers to a set of RR items are independent. Thus respondents are homogenous in their compliance behavior and have a fixed probability of answering each item. For multiple items, and under random sampling of the respondents, the likelihood function can be written as

$$L = \prod_{i=1}^{n} \prod_{j=1}^{J} [c + e \Pr{}_j]^{x_{ij}} \ [1 - (c + e \Pr{}_j)]^{1-x_{ij}}, \tag{4}$$

where $\Pr_j$ is the probability that a person answers affirmatively to item j. In contrast, the IRR models allow for individual differences in the response behavior yielding the following likelihood function:

$$L = \prod_{i=1}^{n} \int \prod_{j=1}^{J} [c + e \Pr{}_j(\theta)]^{x_{ij}} \ [1 - (c + e \Pr{}_j(\theta))]^{1-x_{ij}} f(\theta; \mu, \sigma)\, d\theta, \tag{5}$$

where $f(\theta; \mu, \sigma)$ is the normal density with mean μ and standard deviation σ. Since the mean μ of the population distribution cannot be estimated independently of the item locations, it is convenient to set $\mu = 0$. It is worthwhile stressing that the normal distribution assumption may not always be appropriate in RR studies. Especially, when the number of items is large, it is useful to consider other distributional forms to capture noncompliance variability in the population of interest. For the reported applications, we also applied semiparametric versions (Lindsay et al., 1991) of (5). However, we found that because of the small number of items there was little power in testing the normality assumption against alternative specifications.

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, they may not follow the randomization scheme and provide a truthful answer. If the number of items is sufficiently large ($J \geq 40$), both global and local person-fit statistics (Emons, Sijtsma, & Meijer, 2005) can be developed for IRR models that allow identifying such respondents. However, these methods are of little use for a smaller number of items because they lack power for a satisfactory detection rate. For this reason, we follow a different approach by formulating a specific hypothesis about a response bias in RR data and incorporating it in the IRR models. This hypothesis was motivated by an analysis of different RR data sets which all showed that IRR models underestimated severely the observed number of "No" responses.

The next section examines this response bias in detail and discusses its implementation for the single and multiple domains under study. To explain individual differences, both in being compliant with the regulatory laws and in exhibiting this response bias, we also consider covariates as a further model extension.

3.2.1. Self-Protective Response Behavior. The notion that aggregate- but not individual-level information can be inferred from RRs may neither be intuitive nor obvious to most survey participants. Thus, even when respondents are told that their privacy is protected, not all of them may be convinced that this is indeed the case. As a result, it should be expected that a certain percentage of participants do not trust the randomization scheme and give a "No" response regardless of the question asked. In the following, we refer to this behavior as self-protective (SP)- "No" responses. It is straightforward to account for SP- "No" responses by extending the likelihood function (5) as follows:

$$L = \prod_{i=1}^{n} \left(\pi_i \int \prod_{j=1}^{J} \{[c + e \Pr{}_j(\theta)]^{x_{ij}} \; [1 - (c + e \Pr{}_j(\theta))]^{1-x_{ij}}\} f(\theta; \mu, \sigma)\, d\theta \right.$$
$$\left. + (1 - \pi_i) \prod_{j=1}^{J} \{\Pr(\text{“No”})^{x_{ij}} \; [1 - \Pr(\text{“No”})]^{1-x_{ij}}\} \right), \tag{6}$$

where π denotes the probability of a randomly sampled person to answer the questions according to the randomization mechanism. By decomposing the "No" responses into SP and real ones, the estimates of the underlying noncompliance rates under (6) are higher than under (5).

In the reported application, we specified that participants who decided to give an SP- "No" response, select this response with probability 1. The crucial assumption of the mixture-IRR model (6) is that members of the SP- "No" group do not provide any information about the items' location and discrimination parameters. This assumption is restrictive and thus easily testable in RR data sets. We note that our response-bias hypothesis is a special case of the so-called π^* model (Rudas, Clogg, & Lindsay, 1994) which provides an estimate of the proportion of respondents that are not described by the postulated model:

$$L = \prod_{i=1}^{n} \left(\pi_i \int \prod_{j=1}^{J} \{[c + e \Pr{}_j(\theta)]^{x_{ij}} \; [1 - (c + e \Pr{}_j(\theta))]^{1-x_{ij}}\} f(\theta; \mu, \sigma)\, d\theta + (1 - \pi_i)\Psi \right),$$

where Ψ is an unspecified probability distribution (see also Dayton, 2003). Although not reported here, we considered this special version of a π^* model as an alternative and more general specification to (6) in the data analyses. In all cases, the main source of misfit resulted from one response pattern only, which consisted exclusively of "No" responses.

3.2.2. Multiple Item Sets. A further model extension concerns the analysis of multiple item bundles, each of which is measuring a different aspect of the behavior under study. To investigate the relationship among different domains (e.g., compliance with health and work regulations), we consider multiple θ_{hi}'s ($h = 1, \ldots, H$) that may be correlated in the population of interest. If the correlation among the θ_{hi}'s is substantial, significant efficiency gains can be expected when the item response models for the item bundles are estimated jointly. A convenient assumption is that the H-dimensional vector $\boldsymbol{\theta}_i$ follows a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, leading to the following multivariate version of (5):

$$L = \prod_{i=1}^{n} \int \cdots \int \prod_{h=1}^{H} \prod_{j=1}^{J_h} [c + e\,\Pr(\theta_h)]^{x_{ijh}} \, [1 - (c + e\,\Pr(\theta_h))]^{1-x_{ijh}} f(\boldsymbol{\theta}; \boldsymbol{\Sigma})\, d\boldsymbol{\theta}. \tag{7}$$

As in the single-response case, we assume that not all participants respond truthfully to the questions. The choice to answer truthfully may be domain-specific. For some domains, a person may give an SP-"No" response, but for other domains the person may answer the questions as instructed. Thus, for H domains there are potentially 2^H response classes, one class consisting of truthful respondents and the other classes consisting of SP-"No" respondents for at least one of the domains. Let z_h be an indicator variable with value 0 when questions to a domain are answered truthfully and value 1 when SP-"No" responses are given to a domain, then the likelihood function can be written as

$$\begin{aligned} L = \prod_{i=1}^{n} \sum_{z_1=0}^{1} \sum_{z_2=0}^{1} \cdots \sum_{z_H=0}^{1} \pi_{z_1 z_2 \ldots z_H i} \Bigg\{ \int \cdots \int \\ &\times \prod_{j=1}^{J_1} \{[c + e\,\Pr{}_j(\theta_1)]^{x_{ij1} z_1} \, [1 - (c + e\,\Pr{}_j(\theta_1))]^{(1-x_{ij1}) z_1} \\ &\times \Pr(\text{“no”})^{x_{ij1}(1-z_1)} \, [1 - \Pr(\text{“no”})]^{(1-x_{ij1})(1-z_1)}\} \\ &\times \prod_{j=1}^{J_2} \{[c + e\,\Pr{}_j(\theta_2)]^{x_{ij2} z_2} \, [1 - (c + e\,\Pr{}_j(\theta_2))]^{(1-x_{ij2}) z_2} \\ &\times \Pr(\text{“no”})^{x_{ij2}(1-z_2)} \, [1 - \Pr(\text{“no”})]^{(1-x_{ij2})(1-z_2)}\} \\ &\quad \cdots \\ &\times \prod_{j=1}^{J_H} \{[c + e\,\Pr{}_j(\theta_H)]^{x_{ijH} z_H} \, [1 - (c + e\,\Pr{}_j(\theta_H))]^{(1-x_{ijH}) z_H} \\ &\times \Pr(\text{“no”})^{x_{ijH}(1-z_H)} \, [1 - \Pr(\text{“no”})]^{(1-x_{ijH})(1-z_H)}\} f(\boldsymbol{\theta}; \boldsymbol{\Sigma})\, d\boldsymbol{\theta} \Bigg\}, \end{aligned} \tag{8}$$

where J_h represents the number of items in the hth item bundle. A special case of (8) is obtained when respondents either answer truthfully or select the SP-"No" category for all domains. In this case, the likelihood function (8) simplifies to

$$\begin{aligned} L = \prod_{i=1}^{n} \pi_i \Bigg\{ \int \cdots \int \prod_{h=1}^{H} \prod_{j=1}^{J_h} [c + e\,\Pr{}_j(\theta_h)]^{x_{ijh}} \, [1 - (c + e\,\Pr{}_j(\theta_h))]^{1-x_{ijh}} f(\boldsymbol{\theta}; \boldsymbol{\Sigma})\, d\boldsymbol{\theta} \Bigg\} \\ + (1 - \pi_i) \left\{ \prod_{h=1}^{H} \prod_{j=1}^{J_h} \Pr(\text{“no”})^{x_{ijh}} \, [1 - \Pr(\text{“no”})]^{1-x_{ijh}} \right\}. \end{aligned} \tag{9}$$

3.3. Estimation

Maximum marginal likelihood methods in combination with Gauss–Hermite quadrature are used for the estimation of the mixture IRR models (Bock & Aitkin, 1981). In the reported application, model parameters are estimated by a quasi-Newton method that approximates the inverse Hessian according to the Broyden–Fletcher–Goldfarb–Shanno update (see Gill, Murray, & Wright, 1981). The algorithm utilizes the partial derivatives of the log-likelihood function with respect to all parameters and estimates the Hessian in the form of the crossproduct of the Jacobian of the gradient.

In the application, large sample tests of fit are reported based on the likelihood-ratio (LR) χ^2-statistic (referred to as G^2) which compares observed and expected frequencies of the RR responses. The applicability of these tests is limited when continuous covariates are part of the model. In this case, we report nested model tests based on the deviances of the IRR model with and without covariates (for more details, see De Boeck & Wilson, 2004, p. 56).

4. Results from the 2002 and 2004 IIA Surveys

Table 1 reports the goodness-of-fit statistics obtained from fitting the IRR models to the work and health items for the 2002 and 2004 IIA surveys. We also include the fit statistics obtained when considering the health items only. Because a minimum of three items are needed to identify an IRR model without covariates, no separate IRR models are estimated for the two work items.

The first set of fitted models are based on the baseline RR assumptions represented by (4) and serve as a benchmark for the (mixture-) IRR models. The second part of Table 1 is obtained by fitting a Rasch version of (5) to the health item set and by fitting a Rasch version of (7) to both item sets simultaneously. The fit statistics reported in the third part of Table 1 are obtained from model (6) for the health domain and model (8) for the work and health domains.

The homogeneous-compliance models require the estimation of two and six-item location parameters when applied to the four health items and the four health and two work items, respectively. None of the reported fits are satisfactory, indicating that the assumption of no individual differences does not agree with the data. This result is supported by the fit improvement obtained from the IRR models (5) and (7), that allow for heterogenous compliance behavior

TABLE 1.
Fit statistics of RR models for work and health items.

Survey year	Health items	Health and work items
	G^2 (df)	G^2 (df)
1. Homogeneous compliance		
2002	124.0 (11)	282.4 (57)
2004	56.2 (11)	184.4 (57)
2. Heterogenous compliance		
2002	39.0 (10)	100.6 (54)
2004	23.8 (10)	95.6 (54)
3. Heterogenous compliance and SP-“No” sayers		
2002	14.9 (9)	54.2 (53)
2004	10.8 (9)	63.5 (53)
2002 and 2004	29.4 (24)	131.0 (114)

TABLE 2.
Parameter estimates (and standard errors) of RR models for 2002 and 2004 health and work items.

Parameter	Health ($h=1$)	Health ($h=1$) and Work ($h=2$)		
$\hat{\gamma}_{h1}$	5.64 (1.03)	5.55 (1.01)		3.40 (2.56)
$\hat{\gamma}_{h2}$	4.86 (.92)	4.85 (.90)		5.09 (3.71)
$\hat{\gamma}_{h3}$	4.09 (.81)	4.13 (.81)		—
$\hat{\gamma}_{h4}$	3.33 (.64)	3.35 (.64)		—
$\hat{\sigma}_h$	2.51 (.52)	2.50 (.52)		3.34 (2.45)
$\ln \frac{\hat{\pi}_{11}}{\hat{\pi}_{00}}$	−1.85 (.20)		−1.96 (.18)	
$\ln \frac{\hat{\pi}_{10}}{\hat{\pi}_{00}}$			−3.72 (1.40)	
$\ln \frac{\hat{\pi}_{01}}{\hat{\pi}_{00}}$			−2.67 (1.66)	

without requiring item-specific discrimination parameters. With one additional parameter, the variance of the normal distribution, σ^2, for the health items and three additional parameters for the bivariate covariance matrix of the health and work items, these IRR models provide major fit improvements. However, despite the better fit, these models do not describe the data satisfactorily. As shown by a residual analysis of the data, the main reason for the misfit is that the outcome of consistent "No" responses to the items is greatly underestimated by these models. Thus, more respondents than expected under the IRR models give exclusively "No" responses when asked questions about their compliance with the health and work regulations. Models (6) and (8) can address the problem of extra-"No" responses. The models' parsimonious representation appears to be in good agreement with the 2002 and 2004 IIA survey data as indicated by the fit statistics in Table 1.

As a further step, we fitted mixture-IRR model versions (6) and (8) that constrain the model parameters to be the same for both survey years. As indicated by the goodness-of-fit statistics in Table 1, the uni- and two-dimensional IRR models describe the data well. We conclude that there is no significant change in the item parameters, in the association between the work and health domains, and the incidence of SP respondents over the two-year period. Table 2 reports the resulting parameter estimates. The second column of Table 2 contains the parameters of the health items which are ordered as expected. The population standard deviation of the person parameters is estimated to be 2.51 (.52). About $1/(1+\exp(1.85)) = 14\%$ of the respondents are categorized as SP-"No" sayers. The remaining columns of Table 2 contain the estimates of the two-dimensional IRR mixture model. Although the parameter estimates of the health items are similar to the ones obtained from the univariate IRR model, the large standard errors of the work-item parameter estimates indicate that this model part is only weakly identified. The estimated joint probabilities are $\hat{\pi}_{00} = .81$, $\hat{\pi}_{01} = .06$, $\hat{\pi}_{10} = .02$, and $\hat{\pi}_{11} = .11$, indicating that about 13% and 17% of the respondents are SP-"No" sayers for the health and work items, respectively. Because of the small estimated probabilities for $\hat{\pi}_{01}$ and $\hat{\pi}_{10}$, the fit of the simpler mixture-IRR model (9) (which sets these two parameters equal to 0) is almost the same as the one obtained under its more complex counterpart (8) with $G^2 = 132.9$ ($df = 116$). Thus, when participants give an SP-"No" response, most of them appear to do it regardless of the domain under study.

The estimated correlation between the θ-parameters of the two domains is substantial with $\hat{\rho}_{12} = .58\ (.08)$ but still distinguishable from 1. Thus, not surprisingly, a significant fit deterioration is obtained when estimating a one-dimensional model (6) with different discrimination parameters for both item sets with $G^2 = 163.4$ ($df = 118$), demonstrating that individual differences for both domains are separable.

TABLE 3.
Noncompliance estimates and 95% bootstrap confidence intervals.

Domain	Items	No bias correction	Bias correction [model (9)]
Health	1	.002 (.000, .015)	.033 (.010, .050)
	2	.014 (.000, .034)	.053 (.033, .075)
	3	.048 (.027, .070)	.083 (.055, .112)
	4	.085 (.063, .107)	.130 (.087, .159)
Work	1	.030 (.009, .052)	.074 (.050, .096)
	2	.110 (.086, .133)	.159 (.104, .190)

Figure 2 displays the test information functions for the estimated work and health items. It is clear that the four health items differentiate better among the respondents and over a wider noncompliance range than the two work items. However, these differences are mitigated to some extent by the higher discrimination parameter of the work items.

Most importantly, the mixture-IRR models provide more accurate estimates about the noncompliance rate in the population of interest than RR methods that do not allow for response biases. Table 3 contains the noncompliance estimates and their 95% bootstrap confidence intervals obtained under both the homogeneous-compliance model without a response bias correction and under model (9) with the response bias correction. For example, for the health domain, the bias-uncorrected noncompliance percentages for the four items are estimated as 0.2%, 1.4%, 4.8%, and 8.5%, respectively. In contrast, under model (9) the corresponding estimates are 3.3%, 5.3%, 8.3%, and 13.0%. These differences are substantial and demonstrate the value of the proposed approach for the analysis of RR data. By not taking into account possible response biases, the actual incidence of noncompliance is severely underestimated. Equally important, these estimates do not include the mixture component consisting of the SP respondents. Thus, it

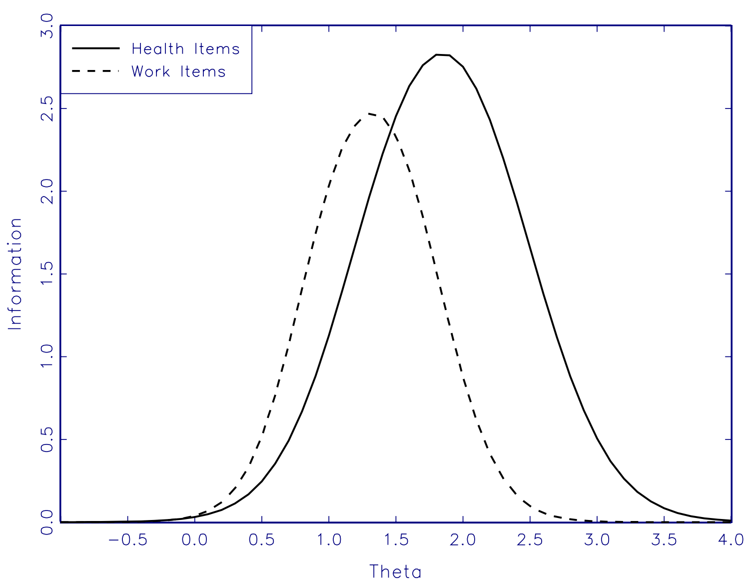

FIGURE 2.
IRR-test information functions of work and health items.

TABLE 4.
Estimated compliance percentages for health and work domains.

Counts	Work			
Health	0	1	2	Total
0	71	7	3	81
1	7	2	1	11
2	3	1	1	5
3	1	1	1	2
4	0	0	0	1
Total	82	11	6	100

is possible that even the response-bias corrected estimates are too low if some or all of the SP respondents are noncompliant as well.

Table 4 displays the joint distribution of the number of infringements for the health and work domains estimated under (9). About 71% of the respondents report no compliance violations. The estimated levels of compliance for the Health and Work domains are 81% and 82%, respectively. These figures are useful in monitoring the effectiveness of interventions aimed at increasing the overall and domain-specific compliance level in the population of interest.

4.1. Covariates

Table 1 demonstrates that respondents differ both in their compliance behavior and in their willingness to follow the randomized response format. It is important to understand the sources of these individual differences. In this section we investigate whether attitudinal, risk-return, and demographic items can account for the variability in the IRR person parameters and predict membership in the SP-"No" sayer class for extended versions of models (6) and (9) that allow for covariates.

Both θ_{i1} and θ_{i2} can be expressed as a function of covariates with $\theta_{ih} = \sum_f x_{if}\beta_f + \epsilon_{ih} (h = 1, 2)$. Random effects that are not captured by the available covariates are represented by ϵ_{ih} which are assumed to be normally distributed with mean 0. In addition, the probability of being in the "Truth"-sayer group can be expressed as a logistic function of the covariates.

Our analyses thus proceeded in two steps. First, we parametrized the $\boldsymbol{\theta}_i$'s as a linear function of the nine factors listed in Section 2 and tested for both time/survey and domain effects. Subsequently, we removed the nonsignificant covariates from the mixture-IRR models and estimated a model in which both the θ- and π-parameters are expressed as a function of covariates. These analyses were conducted for the health, the work, and the combined health and work items for both survey years.

4.1.1. Modeling Individual Differences in Noncompliance. The first step of the analyses showed that there is no significant time and domain effect in the relationship between θ and the covariates. Specifically, the values of the LR tests comparing the two-mixture-IRR models for the 2002 and 2004 survey data were $G^2 = 15.3$ ($df = 13$) for the work items and $G^2 = 18.1$ ($df = 15$) for the health items, indicating that both the work and health item parameters did not vary significantly over the two-year period. Moreover, an LR test of the null hypothesis that there is no domain effect for the nine factors yielded nonsignificant test statistics of $G^2 = 15.9$ ($df = 9$) for the 2002 survey year and of $G^2 = 13.4$ ($df = 9$) for the 2004 survey year. Finally, we tested whether the parameter estimates of a model that includes both domains can be set

TABLE 5.
Effects of covariates (and their standard errors) on θ parameters and descriptive statistics for the combined 2002 and 2004 IIA survey data.

Items	Work and health		Means	Std.
Acceptance	−.32	(.12)	2.2	.6
Clarity	−.01	(.12)	2.6	1.0
Benefits	.80	(.14)	2.6	1.1
Social control	−.39	(.12)	3.6	1.0
Law abidance	−.41	(.12)	2.5	.9
Control	−.18	(.11)	2.9	1.1
Detection	−.15	(.13)	2.7	1.0
Sanction certainty	−.15	(.11)	3.4	.9
Sanction severity	−.21	(.11)	2.4	1.1

equal for the two survey years. This hypothesis could also not be rejected with an LR test statistic of $G^2 = 26.7$ ($df = 19$). The estimated regression effects of this final model, as well as descriptive statistics of the covariates, are listed in Table 5. We note that on average respondents tend toward selecting the middle response category of the covariates. Four of the nine items predict individual variation in θ: Acceptance, Benefits, Law Abidance, and Social Control. However, and more importantly, none of the induced compliance factors are significant although their signs are in the expected direction. The average correlation among the four factors is .13 suggesting that the common variance among the factors is weak at best. A model without these four induced compliance factors did not fit significantly worse as indicated by an LR test statistic of $G^2 = 9.1$ ($df = 4$). Thus, only cost-benefit considerations and social norms appear to play a significant role in a person's decision to act in accordance with the regulatory health and work laws.

4.1.2. Modeling Response Biases. In the analysis of the covariates' effects on the individual-specific π-parameters, we conjectured that the choice of a person to select the "No" response category is determined by both a person's perception that the randomization scheme protects one's privacy ("Trust") and the perceived clarity of the RR instructions ("Instructions"). Respondents who did not feel protected or did not understand the instructions were expected to give an SP-"No" response more frequently. Because we found in preliminary analyses that the perceived clarity of the instructions covaried negatively with the educational level of a person, we included "Education" as a third variable. Finally, we added the remaining attitudinal and risk-return variables to test whether they can predict a person's decision to respond to the questions truthfully. Table 6 displays the results obtained under a joint analysis of the item sets for the two survey years.

As expected, both "Instruction" and "Education" are significant predictors of self-protective responses. Respondents who found the instructions to be clear are more likely to state the truth. Out of the set of attitudinal and risk-return factors, an understanding of the regulations ("Clarity") and the subjective probability that a case will be prosecuted and sanctioned when detected ("Sanction Certainty") as well as the severity of the sanction are significant.

Overall, the applications of the mixture-IRR models yielded a number of important results. First, none of the factors that induce compliance by enforcing the law are important in predicting individual differences. This result suggests that perceptions of the likelihood of external control and the severity of punishments do not account for individual differences in following regulatory laws. Instead, a person is more likely to comply when: (1) the regulations are stated clearly; (2)

TABLE 6.
Effects of covariates (and their standard errors) on θ- and π-parameters for the combined 2002 and 2004 IIA survey data.

Items	Work and health	
θ-Effects; noncompliance with IIA regulations		
Acceptance	−.31	(.11)
Benefits	.76	(.13)
Law abidance	−.39	(.11)
Social control	−.40	(.13)
π-Effects; adherence to RR instructions		
Trust	.19	(.11)
Instruction	.46	(.09)
Education	.54	(.14)
Clarity	.24	(.11)
Detection	−.20	(.11)
Control	−.21	(.12)
Sanction certainty	−.27	(.11)
Sanction severity	−.24	(.12)

there is strong social control; (3) the person's general beliefs are consistent with law abidance; and, most importantly, (4) the expected benefits of noncompliance are minor. These results are consistent with rational choice and attitude-behavior theories that emphasize the importance of both personal benefits and social influences.

Trust in the randomization scheme had a positive but nonsignificant effect on a person's decision to state the truth. A more important factor proved to be the clarity of the randomization instructions. Some respondents, especially those with a lower educational level found it difficult to understand the RR instruction. Because respondents who understand the instruction are less likely to be SP-"No" sayers, this result suggests that it is useful to invest in a better explanation of the RR technique, especially when the target group includes respondents with low levels of education.

It is also important to note that the perception of sanction severity and certainty appear to be better predictors of a person's propensity to respond truthfully than of a person's compliance with work and health regulations. The negative sign of these variables indicates that respondents are more likely to be classified as SP respondents when they estimate the probability to be high that detected noncompliance will be prosecuted and sanctioned severely. To some extent, this result can be explained by the forced "Yes" response of the RR method that requires respondents to incriminate themselves even if they did follow the regulations: Respondents who are concerned about possible sanctions for violating the regulations may have found it difficult to follow the instructions by stating to be noncompliant, and, instead, selected the "No" response.

5. Concluding Remarks

This paper started with the question as to whether it is possible to measure noncompliance. The presented application suggests that the answer is affirmative provided that two conditions are satisfied. First, a sufficiently large number of items needs to be available to obtain a precise estimate of the individual compliance parameter vector $\boldsymbol{\theta}$. The required number of items is a function of the privacy protection provided by the RR scheme and can be computed using the test information function. However, we stress that there are ethical considerations in keeping the number of items low, since the RR scheme provides less protection when responses to different items are correlated. The second requirement is that respondents follow the RR instructions and answer truthfully. If some respondents do not conform to the RR scheme, it becomes necessary to identify them in order to reduce the impact of their responses on the estimates of noncompliance behavior for the domain under study.

This paper proposed mixture-IRR models with concomitant variables to facilitate the simultaneous classification and measurement of respondents in multiple domains. The model framework proved useful in the analysis of the 2002 and 2004 IIA Dutch surveys and provided valuable insights about both the degree of compliance and possible reasons for answering truthfully as well as complying with the IIA regulations. The obtained estimates for noncompliance were substantially higher than the ones obtained in univariate analyses because the mixture-IRR model can adjust for respondents who are systematic "No" sayers. Mixture-IRR models are parsimonious and can be extended in a number of ways to accommodate more diverse response behavior. However, despite their simple form, the models were effective in describing the important features of our data.

The analyses offered new insights about compliance behavior for the IIA regulations. Most noteworthy, it was shown that induced compliance factors as sanctions and control mechanisms were not effective in predicting individual compliance differences. Instead, social control and personal benefits played an important role in support of attitude-behavior and risk-return frameworks. The relationship between the covariates and the individual-difference parameter were found to be stable over a period of two years and to be invariant for the two domains under study. In addition, we identified systematic predictors of response biases in both surveys: Clarity of the RR instructions and an understanding of the regulations mattered strongly. We also found that respondents who were concerned about sanction certainty and severity appeared to be more likely to give a self-protective "No" response. Clearly, despite the apparent usefulness of RR methods in understanding noncompliance, they did not work to the degree as originally was perhaps hoped for.

We conclude that RR methods can yield more valid responses but that they may not fully eliminate response biases. Thus, methods for the analysis of RR data that do not correct for response biases may underestimate substantially the true incidence of the behavior in question. Based on the obtained results, we believe that the proposed mixture-IRR approach has promise in the analysis of RR data when a domain of interest can be described by multiple items. In addition to allowing more powerful inferences about individual differences, multiple items also facilitate the identification of possible response biases. Both advantages may be crucial in estimating the compliance rate in the population of interest and in analyzing relationships between sensitive behavior and explanatory variables.

Appendix

The following (translated) instruction was used in the 2002 web survey. (The original instruction was in Dutch and can be obtained from the second author.)

"We would like to ask you some questions about your allowance. From previous research we know that many people find it really hard to answer questions about allowances, because they consider them a violation of their privacy. Some people fear that an honest answer might even have negative consequences for their allowance. But we do not intend to embarrass anyone. Therefore Utrecht University has developed a method to ask these questions in such a way that your privacy is absolutely protected. You are about to answer the following questions with the help of two dice. With these dice you can throw any number between 2 and 12. Your answer depends on the number you have thrown. In this way your privacy is guaranteed, because nobody, neither the interviewer, nor the researchers, nor the social welfare authorities will ever know the number you have thrown, and thus they can never know why you gave the answer you did.

Now how does this work? On your screen you see the dice rolling. By pushing the "enter" button you will stop the dice from rolling. You can then directly see the number you have thrown. If you push the "enter" button again, the question will appear on your screen.

If you throw 2, 3, or 4 you always push button 1 (= yes). If you throw 11 or 12 you always push button 2 (= no). If you throw 5, 6, 7, 8, 9, or 10 you always answer truthfully. You answer "yes" by pushing button 1 or "no" by pushing button 2.

Even if you find this technique with the dice a bit strange, it is fun to use and it is useful since it guarantees your privacy. Because nobody but you knows what you threw, nobody knows why you pushed button 1 or button 2. Therefore your true answer really remains a secret. The method is still useful for the researchers of Utrecht University because they can estimate both the number of people that pushed button 1 because of the number they threw, and the number of people that pushed button 1 because they had to answer truthfully.

Now follow three exercise questions to acquaint yourself with this method. Please first throw the dice (the virtual dice appear automatically on the screen). If you throw 2, 3, or 4 please push button 1 (= yes). If you throw 11 or 12 please push button 2 (= no). If you throw 5, 6, 7, 8, 9, or 10 please answer the following question truthfully: Have you ever used public transportation without a valid ticket during the last four weeks? Push button 1 for yes, push button 2 for no. (The other exercise questions were: (1) Have you read the paper today? (2) Have you driven through a red traffic light during the last week?)

You just answered the practice questions. Maybe you threw 2, 3, or 4 and therefore had to push button 1, while your true answer would have been "no." On the other hand, you may have thrown 11 or 12 and had to push button 2 while your true answer would have been "yes." From previous research we know that people find it strange to answer incorrectly or even dishonestly. You do not need to worry about this. With this dice method you are being honest when you answer according to the rules. It is like a game, when you follow the rules of the game you are playing it honestly."

References

Becker, G.S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, *76*, 169–217.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F.M. Lord, & M.R. Novick, *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46*, 443–459.

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

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

Dayton, C.M. (2003). Applications and computational strategies for the two-point mixture index of fit. *British Journal of Mathematical and Statistical Psychology*, *56*, 1–13.

Dayton, C.M., & Scheers, N.J. (1997). Latent class analysis of survey data dealing with academic dishonesty. In J. Rost, & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 172–180). New York: Waxmann Muenster.

Eagley, A.H., & Chaiken, S. (1993). *The psychology of attitudes*. Fort Worth: Harcourt, Brace, Jovanovich.

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.

Elffers, H., van der Heijden, P.G.M., & 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.

Emons, W.H.M., Sijtsma, K., & Meijer, R.R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods, 10*, 101–119.

Fishbein, M., & Aijzen, I. (1975). *Belief, attitude, intention and behavior.* Reading, MA: Addison Wesley.

Fox, J.A., & Tracy, P.E. (1986). *Randomized response: A method for sensitive surveys*. Newbury Park, CA: Sage.

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

Gill, P.E., Murray, W., & Wright, M.H. (1981). *Practical optimization*. New York: Academic Press.

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

Landsheer, J.A., van der Heijden, P.G.M., & Van Gils, G. (1999). Trust and understanding, two psychological aspects of randomized response. A study of a method for improving the estimate of social security fraud. *Quality and Quantity, 33*, 1–12.

Lee, R.M. (1993). *Doing research on sensitive topics*. London: Sage.

Lensvelt-Mulders, G.J.L.M., Hox, J.J., van der Heijden, P.G.M., & Maas, C. (2005). Meta-analysis of randomized response research: 35 years of validation. *Sociological Methods and Research, 33*, 319–348.

Lensvelt-Mulders, G.J.L.M., van der Heijden, P.G.M., Laudy, O., & Van Gils, G. (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, 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.

Maddala, G.S. (1983). *Limited dependent and qualitative variables in econometrics*. New York: Cambridge University Press.

Rudas, T., Clogg, C.C., & Lindsay, B.G. (1994). A new index of fit based on mixture methods for the analysis of contingency tables. *Journal of the Royal Statistical Society, Series B, 56*, 623–639.

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

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 data. *Sociological Methods and Research, 32*, 310–336.

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. *Sociological Methods and Research, 28*, 505–537.

van der Linden, W.J., & Hambleton, R.K. (Eds.) (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press. (Original published 1960, Copenhagen: The Danish Institute of Educational Research.)

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

Weber, E.U. (1997). The utility of measuring and modeling perceived risk. In A.A.J. Marley (Ed.), *Choice, decision, and measurement: Essays in honor of R. Duncan Luce* (pp. 45–57). Mahwah, NJ: Erlbaum.

Manuscript received 3 MAY 2006
Final version received 5 SEP 2006