Sample survey on sensitive topics: Investigating respondents' understanding and trust in alternative versions of the randomized response technique

Abstract

In social science research, survey respondents hesitate to answer sensitive questions. This explains why traditional self-report surveys often suffer from high levels of non-response and dishonest answers. To overcome these problems, an adjusted questioning technique is necessary. This article examines one such adjusted questioning technique: the randomized response technique. However, in order to obtain reliable and valid data, respondents need to understand and trust this technique. Respondents' understanding and trust are assessed in two online variants of the randomized response technique: (a) forced response and (b) unrelated question. Results show that understanding was significantly higher in the forced-response condition. Respondents' trust, however, was low in both conditions.

1. Introduction

Asking questions is a common strategy to obtain data for research and policy making. Often surveys are used to gather such data efficiently. However, when the topic is sensitive in nature, respondents may refuse to cooperate in the survey (which leads to "unit non-response"), refuse to answer specific questions (which leads to "item non-response"), and tend to answer dishonestly (Chaudhuri & Mukerjee, 1988; Tourangeau & Yan, 2007). These three sources of error can negatively influence data quality and thus jeopardize the usefulness of the data for both research and policy making. Although such errors cannot be totally

avoided, carefully considering the mode of administration, the assurance of confidentiality and anonymity, and the order and wording of the questions may help to minimize the likelihood of such errors (Bradburn, Sudman, & Wansink, 2004). Another way to deal with sensitive topics is to use alternatives to the usual direct questioning (DQ) technique. The randomized response (RR) technique discussed in this article is one such alternative. However, because this technique is not very common, respondents' understanding and trust in the technique is crucial to obtain valid and reliable data (Fox & Tracy, 1980).

This article aims to compare respondents' understanding and trust in two variants of the RR technique: (a) forced response and (b) unrelated question. Section 2 discusses sensitive questions in general and their possible consequences in survey research. Section 3 introduces the RR technique and its variants, formulating specific issues for further study. Sections 4 and 5 present the design and outcome of an empirical study aimed at comparing the two variants of the RR technique. Finally, Sections 6 and 7 round up the article with discussions and conclusions.

2. Sensitive Topics

Research in cognitive psychology suggests that, in answering survey questions, respondents first assess the accuracy of the response formulated by them without revealing it to the researcher (Cannell, Miller, & Oksenberg, 1981; Tourangeau, Rips, & Rasinski, 2000). A second evaluation based on other goals takes place after assurance that the intended response would fulfil the objectives of the question. At this stage the respondent takes the sensitive nature of the question into account (Cannell, Miller, & Oksenberg, 1981). This could lead to either an honest answer or three negative reactions: (a) unit non-response, (b) item non-response, and (c) dishonest answers. Dishonest answers can be of two types: socially desirable behaviour might be over-reported, such as library card ownership, seat-belt usage, charitable giving, and voting (Bradburn, Sudman, & Wansink, 2004); socially undesirable behaviour, on the other hand, might be under-reported, such as illegal drug usage and alcohol consumption (Bradburn, Sudman, & Wansink, 2004). The problem with systematic non-response and misreporting is that these lead to an over- or under-estimation of the behaviour under study, which then suggests false relationships between variables and leads to wrong conclusions (Rasinski, Willis, Baldwin, Yeh, & Lee, 1999; Tourangeau & Yan, 2007).

As the likelihood of inaccurate responses is high for sensitive questions, it is important that the researcher should know beforehand if the topic under study is sensitive. However, characterizing a topic as sensitive is not a straightforward process (Barnet, 1998); there is limited theoretical conceptualisation. Instead, researchers use a commonsense interpretation, assuming it to be a self-evident matter (Barnett, 1998). A discussion about strategies to determine whether a question is sensitive or not can be found in Barnett (1998). She distinguishes between two commonly used strategies: (a) a post hoc interpretation and (b) an interpretation in relation to some criteria. The former is a subjective evaluation where the

researcher and/or the respondents decide whether the topic under investigation is sensitive (Westall, 2011). Nonetheless, one can never be sure that respondents share the researcher's interpretation of sensitivity. Additionally, there might be differences in perceptions of sensitivity among respondents.

In contrast to the former, the latter strategy is more objective. Barnett (1998) distinguishes between two such objective criteria: empirical and theoretical criteria. Scholars defining sensitivity in relation to an empirical criterion use, for example, the degree of employees' concern if supervisors would be aware of their responses (e.g., Giles & Feild, 1978) or a significant difference between the answers of the respondents who answered anonymously and those who answered with their identities disclosed (e.g., Fuller, 1974). Tourangeau and Yan (2007) use theoretical criteria to define sensitivity. They state that the sensitivity of a question can be determined from an assessment of the following features of a topic: (a) intrusiveness, (b) threat of disclosure, and (c) social undesirability.

First, a topic is sensitive in an intrusive way when it is about a taboo, a topic inappropriate to talk about, such as income and religion (Tourangeau & Yan, 2007). A topic is also intrusive when respondents are ashamed of it, for example, rape and homosexuality.

Second, there is a threat of disclosure when respondents worry about the consequences of an honest response beyond the research project. In face-to-face interviewing, there might be a personal consequence for a teenager admitting to drinking alcohol when the parents are overhearing the conversation. As Tourangeau and Yan (2007) argue, this same question might not be threatening when asked in the presence of peers. This does not mean that the question asked in the presence of peers is not sensitive. Although the threat has disappeared, the question still asks for a specific behaviour that is disapproved in society, but might be perceived as "cool" in the peer group.

Third, a question is sensitive when it asks respondents to admit to have violated a social norm, thus asking for confirmation of socially undesirable behaviour or attitudes (e.g., racism, criminal acts, and neglecting to vote) (Tourangeau & Yan, 2007). Of course, some topics can be sensitive in multiple ways. Barnett (1998) highlights the importance of this type of definition where sensitivity is defined in relation to the context of the question.

Yet, all theoretical criteria aside, it is the respondent who has to answer the survey question. Whether or not a researcher has defined a topic as sensitive, if the respondent believes it is, the perceived sensitivity will influence the answer. Given the likelihood of inaccurate data, a researcher needs to consider the potential sensitivity of the questions for the specific population before administering the survey. When the researcher believes the questions might be sensitive, measures need to be taken. The remainder of this article will use examples under the assumption that the topics and questions are sensitive.

The more sensitive a question, the greater is the number of non-response or dishonest responses (Chaudhuri & Mukerjee, 1988; Dalton & Metzger, 1992; Tourangeau, Rips, & Rasinski, 2000). Barnett (1998) lists several strategies to minimize these types of response error: (a) providing more explicit anonymity guarantees, (b) adjusting the questionnaire format and wording of the questions, (c) adopting another mode of administration such as via computers, and (d) considering alternative methods of data collection, such as focus group discussions. One particular way to guarantee anonymity is to apply the randomized response (RR) technique. The next section will make clear that RR might be a good alternative to the traditional direct questioning (DQ) technique while dealing with sensitive topics.

3. Randomized Response: Basic Principle and Two Variants

The RR technique, originally developed by Warner (1965), aims to eliminate or at least minimize non-response and dishonest answering by survey respondents. This is accomplished by separating the response from the respondent by introducing a controlled measure of chance or uncertainty, which amounts to randomization of the answering process (explained below with an example). This protects the identity of the respondents, at the cost of introducing a degree of uncertainty into the responses.

An example (modified from Fox & Tracy, 1980) will make the basic principle clear. A researcher is interested in the prevalence of date rape (or acquaintance rape) among female students at a university and brings a group of 100 female students together. Then, the researcher asks the students to raise a hand if one has been the victim of date rape. Probably no one will respond due to the sensitive nature of the question. Next, the researcher asks all students to flip a coin and raise their hand if either the coin indicated "heads" or if they have been the victim of date rape. Suppose that 53 students raise their hands. Following probability theory, it can be expected that 50 of them have raised their hands because the coin indicated "heads." So, the other 3 reflect those whose coin indicated "tails," but who are date rape victims. With this data, it can be estimated that 3 of the 50 students who have obtained "tails" are date rape victims (i.e., 6 per cent of the population). At this point the respondents are fully protected since a raised hand has two interpretations. Neither the researcher, nor the other participants will ever know whether a student obtained "heads" or is a victim.

Over the years, several alternative variants of this basic technique have been developed. In what follows only two modern variants will be described: (a) unrelated question and (b) forced response. Next, the various possibilities and applications of this technique are discussed.

3.1. Unrelated Question

In the unrelated question variant of the RR technique (Greenberg, Abul-Ela, Simmons, & Horvitz, 1969) respondents are presented with two questions:

A.   Are   you   a   member   of   sensitive   group   A?
B. Are you a member of non-sensitive group B?

The selection probability of both questions (p and $1 - p$) is known to the researcher and with a randomizing device the respondent is instructed to answer question A or question B (Greenberg et al., 1969). For example, after rolling two dice, a respondent can be instructed to answer question A when the combined results is 5, 6, 7, 8, 9, or 10. Alternatively, when the combined result is 2, 3, 4, 11, or 12, the respondent needs to answer question B.

For the non-sensitive question B, there are two options. The first option is to ask for a non-sensitive trait with unknown population proportion, such as "Do you read the newspaper daily?" Then, two independent samples are necessary since the population proportion of the non-sensitive trait has to be estimated as well as the sensitive trait under investigation (Greenberg et al., 1969). In order to obtain good population estimates, these samples need to be randomly selected. However, this random sampling has nothing to do with the randomization of the answers in the RR technique. This first option is inefficient because it demands a large sample to obtain the same confidence intervals as with the direct questioning (DQ) technique. Therefore, the second option for question B is to ask for a non-sensitive trait with known population proportion, such as "Are you born in August?" (the expected population proportion is 31/365 or 8.5 per cent). In this case, only one sample is necessary (Fox & Tracy, 1980; Scheers, 1992). Although the design is more efficient than the first option, large samples are still necessary if we aim to be as efficient as in using the DQ technique.

A problem occurs when more than one sensitive trait is studied because the group of non-sensitive statements with known population proportion is limited. Usually question B asks for demographic characteristics, but when too much personal information is asked, respondents could think identification is possible (Lensvelt-Mulders, 2003; Tracy & Fox, 1981).

3.2. Forced Response

Another variant was developed by Boruch (1971, cited in Peeters, 2005). The idea of two questions is discarded in the forced response variant; a single sensitive question with a more complex randomizing device takes its place. Each respondent is given two dice they need to roll. When the combined result is 2, 3, or 4, the respondent is asked to answer "Yes," irrespective of their honest answer. If, however, the combined result is 11 or 12, the respondent is asked to answer "No," again irrespective of their honest answer. A result of 5, 6, 7, 8, 9, or 10 then requires the respondent to answer the question honestly (Peeters, 2005). The probability of each of the combinations is known to the researcher. For the researcher it is unclear whether a "Yes" response indicates the possession of the sensitive trait or it is merely a forced "Yes." This variant is as efficient as the unrelated question variant, with known population proportion (Lensvelt-Mulders, 2003; van der Heijden, Hox, & Elffers, 2002). This

means that the sample size still needs to be twice as large as the sample in the DQ technique, to obtain the same confidence intervals and power of statistical analysis, but it is much better than the original Warner model which needs eight times that (van der Heijden, Hox, & Elffers, 2002).

The forced response variant also suffers from a deficit. It might lead to another source of measurement error and biased estimates when respondents do not follow the instructions. For example, an "innocent" marketer is asked whether he has engaged in price fixing. After rolling the dice, the marketer is instructed to answer "Yes." Since this has the perception of self-incrimination and dishonesty, it might tempt the innocent respondent to answer "No," a response tendency called *cheating* (Boeije & Lensvelt-Mulders, 2002; van der Heijden, Hox, & Elffers, 2002; van der Heijden, van Gils, Bouts, & Hox, 2000). However, this does not need to be an insurmountable problem, since methods were recently developed to detect cheaters and deal with them in the analysis (e.g., Böckenholt, Barlas, & van der Heijden, 2009; Cruyff, Böckenholt, van den Hout, & van der Heijden, 2008). Still, compared with the unrelated question variant, the cognitive effort expected from the respondent is higher.

3.3. Developments of the Randomized Response Technique

The RR technique has been developed over the years. A first development is the possibility of going beyond dichotomous questions. The basic RR technique and its variants discussed above all imply a Yes/No response to the sensitive question. In social science research, one is not only interested in the prevalence of sensitive behaviour, one is also interested in its frequency. Frequency questions such as "How often do you pray to God?" might yet be more sensitive than the prevalence question "Do you pray to God?" (Scheers, 1992). Therefore, several alternatives have been developed to estimate quantitative and categorical characteristics with more than two possible answers (e.g., Liu & Chow, 1976; Peeters, 2005).

Second, as with traditional survey data (i.e., through direct questioning, DQ), it is possible to correlate data obtained through the RR technique with both other RR data and DQ data. A chi-square test of independence between two categorical variables of which one or both are collected through RR is a first option (van den Hout & van der Heijden, 2004; van der Heijden, Hox, & Elffers, 2002). The calculation of odds ratios is a second possibility (van den Hout & van der Heijden, 2002). Third, more complex log-linear models can be fitted, such as an adjusted version of the logistic regression with several DQ independent variables and an RR dependent variable (van der Heijden, Hox, & Elffers, 2002). Fourth, the RR technique can be described as a special case of a latent class model (Böckenholt, Barlas, & van der Heijden, 2009; van den Hout & van der Heijden, 2004). Recently, more advanced models were developed using item-response theory to estimate individual differences in the sensitive behaviour under study (Böckenholt, Barlas, & van der Heijden, 2009; Fox, 2005). Analyses with RR data are thus possible, but fitting complex causal models is difficult and requires advanced levels of statistical skills. However, it is (for the moment) impossible to have both

an RR dependent variable and several RR independent variables. This is an important deficiency of the technique.

A third development concerns the flexibility in the application of the RR technique. The modalities of the technique can be adjusted to the wishes of the researcher and the study population; face-to-face interviewing (Elffers, van der Heijden, & Hezemans, 2003; Lensvelt-Mulders, van der Heijden, Laudy, & van Gils, 2006; van der Heijden, Hox, & Elffers, 2000), telephone interviewing (Stem & Steinhorst, 1984), mail questionnaire (Robertson & Rymon, 2001; Stem & Steinhorst, 1984), and completion on paper in large groups (Burton & Near, 1995; Donovan, Dwight, & Hurtz, 2003) are all possibilities, as are electronic versions such as computer assisted randomized response surveys with or without the presence of the researcher (Coutts & Jann, 2011; Lensvelt-Mulders et al., 2006; Peeters, 2005; van der Heijden et al., 2000). Moreover, numerous possible randomizing devices are available: coloured balls, spinners, dice, coins, digits from a telephone number, and so on (Scheers, 1992). Respondents might actually welcome this diversity, because it might give the survey project something interesting compared to the traditional questionnaires.

3.4. Application of the Randomized Response Technique

Application of the RR technique requires larger sample sizes and involves complex statistical analysis. This can only be justified if the RR technique yields responses which are more "valid" than the responses that would be obtained through an ordinary survey using the DQ technique.

However, sensitive topics often relate to hidden behaviour. Consequently, the validity of both RR and DQ studies on sensitive topics is hard to establish. Lensvelt-Mulders, Hox, van der Heijden, and Maas (2005) distinguish between two types of validation studies: (a) individual validation studies and (b) comparative studies. In individual validation studies, the true score of each respondent is known to the researcher and can be compared with the observed score. In comparative studies, one or more techniques (e.g., DQ) are compared with the RR technique. If the sensitive traits are socially undesirable, thus prone to underreporting, then a higher proportion or frequency estimate could be considered more valid.

Some authors have compared RR and DQ in studies of police integrity (Peeters, 2005), tax evasion (Holbrook & Krosnick, 2010; Larkins, Hume, & Garcha, 1997), and academic cheating (Burton & Near, 1995), but could not conclude if RR was more valid. On the other hand, some authors have found significantly higher estimates (interpreted as more valid) with RR methods compared to DQ on topics such as criminal arrests (Tracy & Fox, 1981), job application faking (Donovan, Dwight, & Hurtz, 2003), abortion (Lara, Strickler, Diaz Olavarrieta, & Ellertson, 2004), unethical behaviour in accounting (Gibson & Frakes, 1997), and theft by employees (Wimbush & Dalton, 1997). Since the results are mixed, Lensvelt-Mulders et al. (2005) conducted a meta-analysis. They conclude that all data collection

methods will lead to some amount of misreporting of the sensitive behaviour under study. However, the RR techniques obtain the lowest level of misreporting; this is an indication of more valid results than DQ. Although reducing unit non-response and item non-response is a second objective of RR techniques, none of the studies cited above discussed this issue.

3.5. Respondents' Understanding and Trust

The results of the RR validation studies are thus not unambiguously positive. Landsheer, van der Heijden, and van Gils (1999) suspect that there are two important preconditions for the RR method to be successful: (a) the degree of respondents' understanding of the technique and (b) their trust in it. They state that both understanding and trust are important to obtain accurate data. However, this issue is rarely investigated and it is unclear how both preconditions relate to each other.

As becomes clear from the limited empirical work that has been done, it is not self-evident that respondents understand the questionnaire instructions and the mechanisms for protection of their anonymity since these require additional cognitive effort (Landsheer, van der Heijden, and van Gils, 1999; Lensvelt-Mulders et al., 2005). Respondents that do not understand the instruction might unintentionally fail to comply with them. Additionally, respondents that to not understand how the RR method protects their anonymity might refuse to follow the instructions. Landsheer, van der Heijden, and van Gils (1999) state that respondents are less likely to trust the method and comply with it when they do not understand it. However, some studies show that respondents' trust is not self-evident either (Coutts & Jann, 2011; Lensvelt-Mulders & Boeije, 2007). Coutts and Jann (2011) found that most respondents understood the RR instructions, but that only a minority of them trusted the guarantees concerning their anonymity. Based on their semi-structured qualitative interviews, Lensvelt-Mulders and Boeije (2007) came to the same results: most respondents found the instructions clear, but they did not trust that the randomizing device could protect their privacy. These results show that both preconditions are not self-evident. Additionally, understanding does not imply trust automatically. Although there are no studies to confirm or reject this, theoretically it is possible that there is trust without understanding. Yet, both understanding and trust are necessary.

Due to this potential lack of understanding and trust, data collected with the RR technique might still have poor quality. The next question is then whether the additional costs of a larger sample and the analytical complexities can be justified. Online surveys are a good alternative to reduce the financial costs. Additionally, it has been shown that self-administration without the presence of a researcher improves honesty in answering questions on sensitive topics (Bowling, 2005). However, Landsheer et al. (1999) point to the presence of a researcher as a crucial condition to have respondents understand and trust the RR method. In fact, most RR projects applied a data collection method where a researcher was physically available for respondents; online RR surveys are rather sparse (e.g., Coutts & Jann, 2011; Holbrook &

Krosnick, 2010). As the absence of a researcher might impact respondents' understanding and trust, it is important to empirically explore understanding and trust in an online survey environment where a researcher is unable to provide additional information on request. However, of the online RR applications, only Coutts and Jann (2011) asked respondents for their understanding and trust. They compared the forced response variant of RR with both DQ and the unmatched count technique (which is an alternative to direct questioning; more information about the technique can be found in Coutts & Jann, 2011). As discussed above, the forced response variant of RR might require higher cognitive effort by respondents to comply with the forced "Yes" and "No" than the unrelated question variant of RR. This might explain why respondents' trust in RR was that low in Coutts and Jann's (2011) study.

4. Empirical Study to Examine Understanding and Trust

An empirical study was designed to compare the two variants of RR, forced response and unrelated question, with regard to respondents' understanding and trust. The study compared the two variants of RR in an online environment. As the forced response variant might require higher cognitive effort by respondents, it is expected that understanding and trust will be significantly lower there than in the unrelated question variant of RR. Additionally, it is expected that differences in trust will lead to a significantly higher drop-out in the forced response than in the unrelated-question condition. This leads to the following hypotheses:

- *Hypothesis 1*. Understanding is higher in the unrelated question variant of RR than in the forced response variant
- *Hypothesis 2*. Trust is higher in the unrelated question variant of RR than in the forced response variant
- *Hypothesis 3*. Drop-out in the RR part of the survey is higher in the forced response variant of RR than in the unrelated question variant

4.1. Data Collection

The data used in this study were collected in 2011. A Web survey on unethical behaviour at work was distributed to friends, family, and acquaintances by e-mail and through the social network site, Facebook. Respondents were asked to further distribute the survey in their networks. The welcome page introduced unethical behaviour as the topic of interest. Respondents were assured total anonymity because a special method would be used. It was then explained that in some cases they would be asked not to answer the question on unethical behaviour, depending on the throw of two dice. It was clarified that their answers would still be valuable because probability theory would help to determine the right amount of violations, but that these violations could never be linked to an individual respondent. The introduction was kept as short as possible since experience shows that respondents do not bother to read long introductions.

A total of 272 respondents participated; 75.4 per cent of them completed the survey, while 24.6 per cent dropped out before the end. A total of 214 respondents gave demographic

information. The youngest person was 21 years old, the oldest 63 years, with a mean of 32.2 years. The majority of the respondents were female (63.6 per cent) and had completed higher studies (86.4 per cent). This sample is by no means representative of the working population, but since the aim of the study is to compare the two variants of RR, representativeness was not a primary concern.

4.2. Experimental Conditions

There are two experimental conditions: forced response and unrelated question. First, respondents were asked whether the last digit of their personal mobile phone number was even or odd. Based on their answer, respondents were randomly assigned to one of the conditions: forced response for even numbers (123 respondents) and unrelated question for odd numbers (149 respondents). As the age variable was not normally distributed (Kolmogorov-Smirnov $Z = 3.451$, $p = 0.000$), a Mann-Whitney U test shows no significant difference between the respondents in both conditions based on their age (Z-score = -1.493, $p = 0.135$). Additionally, a chi-square test shows no significant difference between both conditions on gender ($\chi^2 = 0.035$, df = 1, $p = 0.851$), but indicates an overrepresentation of lower educated employees in the unrelated question group (χ^2: 8.67, df = 2, $p = 0.013$). Weight score were added to the model to correct for this overrepresentation.

Two online dice were used as the randomizing device in both conditions. Respondents were instructed to copy the Web link to the dice, paste it in another window and throw the two dice only once. It was expected that they would then be confident that the result of their throw would not be controlled nor saved by the Web survey program. When respondents in the forced-response condition rolled a combined result of 2, 3, or 4, they had to answer "Yes"; when the result was 11 or 12, they had to answer "No." A result of 5, 6, 7, 8, 9, or 10 required them to answer honestly. The selection probability to answer honestly was thus 27/36 = 0.75 (in a roll of two dice, there are 36 outcomes altogether, of which a sum of 5, 6, 7, 8, 9, or 10 is obtained in 27 cases). In the unrelated-question condition, respondents were presented two questions with only one "Yes/No" possibility. If they rolled 2, 3, 4, 11, or 12, they had to answer the non-sensitive B-question (probability = 9/36 = 0.25). Otherwise, they had to answer the sensitive A-question. Again, the selection probability of the sensitive question was 0.75.

4.3. Measures

Understanding was defined as "insight into the protection that the methods have to offer and knowing what to do at every phase of the [answering] process" (Landsheer, van der Heijden, & van Gils, 1999, pp. 3-4), while trust was defined as "the confidence of being protected by the use of Randomized Response methods" (p. 3). For the measures of understanding and trust, inspiration was found in the studies of Landsheer, van der Heijden, and van Gils (1999) and Westall (2011). Understanding, on the one hand, was measured with four items ($\alpha = 0.72$). For each item, respondents could indicate whether they agreed or not with (1) strongly

disagree, (2) disagree, (3) somewhat disagree, (4) somewhat agree, (5) agree, and (6) strongly agree. Trust, on the other hand, was measured with five items (α = 0.70) on the same answering scale. The items, their means, and standard deviations are presented in Table 1. Neither the understanding scale (Kolmogorov-Smirnov Z = 2.031, p = 0.001), nor the trust scale (Kolmogorov-Smirnov Z= 2.012, p = 0.001) is normally distributed. Consequently the non-parametric Mann-Whitney U test was used for comparing between the experimental conditions.

5. Results

Table 1 presents the individual items for understanding and trust in the two experimental conditions. In general, respondents seem to have a good understanding of both the forced response and the unrelated question instructions. However, it is clear that for three items the mean is higher in the forced-response condition. The means of the trust items indicate, in contrast, that trust in the RR variants is not that high. As this trust level cannot be compared to trust in direct questioning (DQ), it is still uncertain whether trust issues are problematic for the RR technique. There is nevertheless room for improvement. Additionally, the items have more or less the same score in both experimental conditions.

Table 1. *Items of Understanding and Trust With Their Means (and Standard Deviations)*

Understanding Items	Forced Response	Unrelated Question
The instructions of the questionnaire were clear	5.36 (0.828)	4.72 (1.254)
It is clear that this procedures guarantees secrecy about my real activities	4.54 (1.206)	4.57 (1.255)
It was clear when I had to answer what	5.33 (0.831)	4.54 (1.350)
I am confident that I followed the instructions correctly	5.27 (0.868)	5.00 (1.232)
Trust Items		
This seems more like a game than like real research (-)	4.13 (1.393)	3.97 (1.412)
This questioning method stimulated my tendency to answer less seriously (-)	2.40 (1.025)	2.75 (1.394)
I am confident that my answers are	4.83	4.81

secret	(1.098)	(1.062)
I have the feeling that the researcher is misleading me (-)	3.11 (1.340)	3.20 (1.430)
I do not trust this questioning method (-)	3.19 (1.307)	3.06 (1.533)
Note. (-) negatively worded item		

Combining the items in a scale, a Mann-Whitney U test indicates that there is significantly higher understanding in the forced-response condition (Z-score = -3.262, p = 0.001). In contrast, it seems that there is no significant difference in trust between the two experimental conditions (Z -score: -0.100, p = 0.920). Both hypotheses 1 and 2 are thus rejected. In total 67 respondents dropped out the survey, but 57 of them did in the RR part; the remaining 10 dropped out in the part about demographic information or the understanding and trust questions. In the forced-response condition, 18.7 per cent of the respondents dropped out in the RR part; that was 22.8 per cent in the unrelated-question condition. A chi-square test indicates that this small difference is however not significant (χ^2 = 0.690, df = 1, p = 0.406), which also rejects hypothesis 3.

6. Discussion

Although RR instructions place a cognitive burden on respondents, understanding was high in both experimental conditions. The respondents in this project were mainly highly educated, thus the higher cognitive demands might not be a barrier for them. These results are in line with Coutts and Jann's (2011) Web survey on RR. However, contrary to expectation, understanding was significantly higher in the forced-response condition than in the unrelated-question condition. Hypothesis 1 needs to be rejected. One possible explanation might be that in the unrelated question variant two questions were presented (A and B), but there was only one possibility to answer. The Yes/No answer was presented below question B. The instructions clarified that respondents had to use this "Yes" or "No" irrespective of the question that needed to be answered. Probably this might not have been clear enough or respondents did not read the instructions carefully. In that respect, a forced "Yes" or "No" is easier to understand. Consequently, the presence of a researcher might not be crucial to obtain understanding in an online forced response survey with highly educated respondents, but perhaps the understanding scores would have been higher if respondents would have had the opportunity to ask for additional information. However, Landsheer, van der Heijden, and van Gils (1999) pointed to the necessity of a researcher being physically present to help respondents attain better understanding. This might be particularly true for children or mentally disabled respondents, but RR applications with these groups is lacking. Future research should test whether online variants of the RR technique could also be used for those types of respondents and how the instructions should be presented in order to be

understandable. This would significantly reduce the costs of the survey and it might encourage honest answering on sensitive questions (Bowling, 2005).

As was already stated by Coutts and Jann (2011), respondents' understanding does not guarantee their trust, since these items had low scores in both experimental conditions. There was also no significant difference in trust between respondents under the two conditions. Hypothesis 2 also needs to be rejected. Lensvelt-Mulders and Boeije (2007) obtained the same results in face-to-face qualitative interviewing, so it is uncertain whether the online environment is responsible for this lack of trust. Unfortunately in their Web survey comparison of trust, Coutts and Jann (2011) did only assess respondents' trust in the RR technique and the unmatched count technique, not in the DQ technique. Further research is needed to compare these trust levels with those in the protection mechanisms of DQ surveys, both in online environments as paper-and-pencil administrations with or without the presence of a researcher.

Some comments of respondents at the end of the survey point to possible explanations for this lack of trust. First, one respondent indicated that it was difficult to copy the dice-rolling link and paste it in another window. Possibly it might have been unclear that this procedure was introduced to show respondents that the results dice rolling could not be saved by the survey program. Indeed, another respondent indicated that he would have preferred to roll with real dice instead of electronic ones. Another randomizing device seems necessary, one that is not electronic, easy by hand and allows a fixed selection probability. Second, a respondent indicated he did not believe his answers were secret because his IP address could have been saved. Of course, this was not the case. Thus, it needs to be clarified in the instructions that respondents' IP addresses would not be saved, but it is unclear whether respondents will believe that. Third, there were only three RR questions. Some respondents indicated that they had to answer honestly (forced response) or had to answer the sensitive question (unrelated question) in each of the three RR trials. To show respondents that the dice are truly random, more RR questions would have increased the chances of a forced response or a response on the non-sensitive question. However, in that case, it becomes difficult to find non-sensitive questions for the unrelated question variant.

The third hypothesis also needs to be rejected; there is no significant difference in drop-out between the experimental conditions. Even if there is less trust in the method, respondents keep participating. It is, however, uncertain to what degree they comply with the instructions and answer honestly. Data quality is thus still uncertain.

7. Conclusion

The aim of this study was to compare respondents' understanding and trust in two online variants of the RR technique: (a) forced response and (b) unrelated question. The results indicate that more experimental research is necessary to design an optimal survey with

sensitive questions. Respondents' understanding and trust need to be assessed in both online and non-online surveys, with and without the presence of a researcher, for a diverse population and for RR and DQ techniques. This study shows that for highly educated respondents an online forced response survey might be a good solution if the instructions are clearly written, if there are enough RR trials, and if there is the possibility of a non-electronic randomizing device, so trust scores would increase. However, the online environment did decrease the sampling cost, but is not yet a solution for the statistical complexities in the analysis of RR data. Therefore, other alternatives to DQ (e.g., unmatched count technique) should be assessed as well.

Acknowledgements

References

Barnett, J. (1998). Sensitive questions and response effects: An evaluation. *Journal of Managerial Psychology*, *13*(1 and 2), 63-76.

Böckenholt, U., Barlas, S., & van der Heijden, P. G. M. (2009). Do randomized-response designs eliminate response bias? An empirical study of non-compliance behaviour.*Journal of Applied Econometrics*, *24*(3), 377-392.

Boeije, H., & Lensvelt-Mulders, G. (2002). Honest by chance: A qualitative interview study to clarify respondents' (non-)compliance with computer-assisted randomized response.*Bulletin de Méthodologie Sociologique*, *75*(1), 24-39.

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health*, *27*(3), 281-291.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design for market research, political polls, and social and health questionnaires*. San Francisco, CA: Jossey-Bass.

Burton, B. K., & Near, J. P. (1995). Estimating the incidence of wrongdoing and whistle-blowing: Results of a study using randomized response technique. *Journal of Business Ethics*, *14*(1), 17-30.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 389-437). San Francisco, CA: Jossey-Bass.

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

Coutts, E., & Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT).*Sociological Methods & Research*, *40*(1), 169-193.

Cruyff, M. J. L. F, Böckenholt, U., van den Hout, A., & van der Heijden, P. G. M. (2008). Accounting for self-protective responses in randomized response data from a social security survey using the zero-inflated Poisson model. *The Annals of Applied Statistics*,*2*(1), 316-331.

Dalton, D. R., & Metzger, M. B. (1992). Towards candor, cooperation, & privacy in applied business ethics research: The randomized response technique (RRT). *Business Ethics Quarterly*, *2*(2), 207-221.

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance*, *16*(1), 81-106.

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*(4), 409-439.

Fox, J. A., & Tracy, P. E. (1980). The randomized response approach. Applicability to criminal justice research and evaluation. *Evaluation Review*, *4*(5), 601-622.

Fox, J.-P. (2005). Randomized item response theory models. *Journal of Educational and Behavioral statistics*, *30*(2), 1-24.

Fuller, C. (1974). Effect of anonymity on return rate and response bias in a mail survey.*Journal of Applied Psychology*, *59*(3), 292-296.

Gibson, A. M., & Frakes, A. H. (1997). Truth or consequence: A study of critical issues and decision making in accounting. *Journal of Business Ethics*, *16*(2), 161-171.

Giles, W. F., & Feild, H. S. (1978). Effects of amount, format, and location of demographic information on questionnaire return rate and response bias of sensitive and nonsensitive items. *Personnel Psychology*, *31*(3), 549-559.

Greenberg, B. G., Abul-Ela, A. A., Simmons, W. R., & Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, *64*(326), 520-539.

Holbrook, A. L., & Krosnick, J. A. (2010). Measuring voter turnout using the randomized response technique: Evidence calling into question the method's validity. *Public Opinion Quarterly*, *74*(2), 328-343.

Landsheer, J. A., van der Heijden, P., & 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), 1-12.

Lara, D., Strickler, J., Diaz Olavarrieta, C., & Ellertson, C. (2004). Measuring induced abortion in Mexico: A comparison of four methodologies. *Sociological Methods Research*, *32*(4), 529-558.

Larkins, E. R., Hume, E. C., & Garcha, B. S. (1997). The validity of the randomized response method in tax ethics research. *Journal of Applied Business Research*, *13*(3), 25-32.

Lensvelt-Mulders, G. (2003). *Randomized response technieken als instrument voor het onderzoek van sociaal gevoelige onderwerpen* [Randomized response techniques as instrument for research on socially sensitive topics]. In A. E. Bronner, P. Dekker, J. C. Hoekstra, E. De Leeuw, T. Poiez, K. de Ruyter, & A. Smidst (Eds.), *Ontwikkelingen in het marktonderzoek: Jaarboek 2003* [Developments in marketing research: Annual report 2003] (pp. 59-74). Haarlem, Netherlands: De Vrieseborch.

Lensvelt-Mulders, G., & Boeije, H. (2007). Evaluating compliance with a computer assisted randomized response technique: A qualitative study into the origins of lying and cheating. *Computers in Human Behaviour*, *23*(1), 591-608.

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research: Thirty-five years of validation.*Sociological Methods Research*, *33*(3), 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 (Statistics in Society)*, *169*(2), 305-318.

Liu, P. T., & Chow, L. P. (1976). A new discrete quantitative randomized response model.*Journal of the American Statistical Association*, *71*(353), 72-73.

Peeters, C. F. W. (2005). *Measuring politically sensitive behaviour: Using probability theory in the form of randomized response to estimate prevalence and incidence of misbehaviour in the public sphere: A test on integrity violations*. Vrije Universiteit Amsterdam.

Rasinski, K. A., Willis, G. B., Baldwin, A. K., Yeh, W., & Lee, L. (1999). Methods of data collection, perceptions of risks and losses, and motivation to give truthful answers to sensitive survey questions. *Applied Cognitive Psychology*, *13*(5), 465-481.

Robertson, D. C., & Rymon, T. (2001). Purchasing agents' deceptive behavior: A randomized response technique study. *Business Ethics Quarterly*, *11*(3), 455-479.

Scheers, N. J. (1992). A review of randomized response techniques. *Measurement and Evaluation in Counseling & Development*, *25*(1), 27-41.

Stem, D. E., & Steinhorst, R. K. (1984). Telephone interview and mail questionnaire applications of the randomized response model. *Journal of the American Statistical Association*, *79*(387), 555-564.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*,*133*(5), 859-883.

Tracy, P. E., & Fox, J. A. (1981). The validity of randomized response for sensitive measurements. *American Sociological Association*, *46*(2), 187-200.

van den Hout, A., & van der Heijden, P. G. M. (2002). Randomized response, statistical disclosure control and misclassification: A Review. *International Statistical Review*,*70*(2), 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 & Research*, *32*(3), 384-410.

van der Heijden, P. G. M., Hox, J. J., & Elffers, H. (2002). *Het meten van regelnaleving: Een voorstudie in opdracht van het expertisecentrum rechtshandhaving van het ministerie van justitie* [The measurement of rule compliance: A preliminarily study commissioned by the centre of expertise of law enforcement of the ministry of justice]. Leiden, Netherlands: Nederlands Studiecentrum Criminaliteit en Rechtshandhaving.

van der Heijden, P. G. M., van Gils, G., Bouts, J., & Hox, J. J. (2000). A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning: Eliciting sensitive information in the context of welfare and unemployment benefit. *Sociological Methods Research*, *28*(4), 505-537.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*(306), 63-69.

Westall, W. (2011). *Comparing respondents' trust and understanding of two variants of the randomized response method*. Unpublished master's dissertation, Katholieke Universiteit Leuven, Belgium.

Wimbush, J. C., & Dalton, D. R. (1997). Base rate for employee theft: Convergence of multiple methods. *Journal of Applied Psychology*, *82*(5), 756-763.