Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Published as:

Coutts, Elisabeth, Ben Jann (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.

ETH Zurich Sociology Working Paper No. 3

Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT)

Elisabeth Coutts and Ben Jann

January 2008

ETH Zurich, Chair of Sociology

SEW E 21, Scheuchzerstrasse 70
8092 Zurich, Switzerland

Tel. +41 44 632 55 56
Fax +41 44 632 10 54
info@soz.gess.ethz.ch
www.socio.ethz.ch

Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT)[*]

Elisabeth Coutts and Ben Jann

ETH Zurich

January 2008

Comments are welcome. Please direct your correspondence to

Elisabeth Coutts
ETH Zurich, SEW E 21
Scheuchzerstrasse 70
CH-8092 Zurich, Switzerland
E-mail: coutts@soz.gess.ethz.ch

or

Ben Jann
ETH Zurich, SEW E 28
Scheuchzerstrasse 70
CH-8092 Zurich, Switzerland
E-mail: jann@soz.gess.ethz.ch

[*] We thank Jürg Schellenberg and Fabian Schenkel, who helped develop the questionnaire and program the survey, and Frauke Kreuter for her helpful comments. We also thank Globalpark GmbH and Respondi AG for providing free access to the "Sozioland" respondent pool. This research was supported by the German Research Foundation (DFG) Priority Program Number 1292 on "Survey Methodology".

Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT)

Abstract

Gaining valid answers to so-called sensitive questions is an age-old problem in survey research. Various techniques have been developed to guarantee anonymity and minimize the respondent's feelings of jeopardy. Two such techniques are the randomized response technique (RRT) and the unmatched count technique (UCT). In this study we evaluate the effectiveness of different implementations of the RRT (using a forced-response design) in a computer-assisted setting and also compare the use of the RRT to that of the UCT. The techniques are evaluated according to various quality criteria, such as the prevalence estimates they provide, the ease of their use, and respondent trust in the techniques. Our results indicate that the RRTs are problematic with respect to several domains, such as the limited trust they inspire and non-response, and that the RRT estimates are unreliable due to a strong false "no" bias, especially for the more sensitive questions. The UCT, however, performed well compared to the RRTs on all the evaluated measures. The UCT estimates also had more face validity than the RRT estimates. We conclude that the UCT is a promising alternative to RRT in self-administered surveys and that future research should be directed towards evaluating and improving the technique.

Keywords: sensitive questions, online survey, randomized response technique, unmatched count technique, item count technique, methodological experiment

I. Introduction

Gaining valid answers to so-called sensitive questions, i.e. questions pertaining to private, socially frowned upon or illegal behavior, is an age-old problem in survey research. Making valid estimates of the prevalence or frequency of sensitive behavior is hindered by the fact that people typically underreport such behavior, while overreporting socially desirable behaviors (Barnett 1998; Lee 1993; Rasinski et al. 1999; Singer, von Thurn and Miller 1995; Tourangeau, Rips and Rasinski 2000: Chapter 9; Tourangeau and Yan 2007). There is evidence that such bias stems from several sources, including the sensitivity of the topic being asked about, question format, data collection mode, respondent and interviewer characteristics, and interviewer behavior.

Researchers have tried to combat this response bias, or systematic over- or underreporting depending on the desirability of the behavior in question, in a variety of ways. Several of these methods are geared toward providing the respondent greater perceived confidentiality. Various dejeopardizing techniques have been developed toward that end. Lee (1993; 82-90) uses this term to describe a variety of statistically based techniques designed to guarantee anonymity and minimize the respondent's feelings of jeopardy when asked to admit to a behavior that is stigmatized or incriminating. Two such techniques are the randomized response technique (RRT, introduced by Warner in 1965), and the unmatched count technique (UCT, also called the item count technique, the unmatched block design, or block total response, see Raghavarao and Federer 1979).

The Randomized Response Technique

The RRT technique has been implemented in various forms (Folsom et al. 1973; Fox and Tracy 1986; Greenberg et al. 1969; Kuk 1990; Scheers and Dayton 1987). However, all of these forms rely on the pairing of an unthreatening question or contingency with the sensitive question of interest. A randomizing device is used to determine whether the respondent will answer the sensitive question, a direction known only to the respondent. For example, in a variant of Boruch's (1971) forced response method, a respondent may be asked to flip a coin to determine

whether to automatically answer "yes" (heads) or instead answer a sensitive question (tails). Since only the respondent knows whether he or she has flipped heads or tails, a "yes" answer cannot be interpreted as an admission of guilt. However, the proportion of the sample that has engaged in the behavior of interest can be calculated with knowledge of the properties of the randomizing device.

There are many indications in the literature that the randomized response technique (RRT) leads to more accurate estimates of both the prevalence of socially undesirable behavior and the frequency with which it is performed by a single person than asking the sensitive question directly. Mostly, the use of the RRT has resulted in increased reporting of sensitive or frowned-upon behaviors as varied as child abuse, drug use, abortion, employee theft, welfare fraud and premature sign-offs on audits in comparison to the reporting of the same deeds in response to a direct question (Goodstadt and Gruson 1975; Lara et al. 2004; Reckers, Wheeler and Wong-On-Wing 1997; Stem and Steinhorst 1984; Tracy and Fox 1981; van der Heijden et al. 2000; Zdep and Rhodes 1976). Of course, a higher reported prevalence or frequency does not necessarily imply a more accurate prevalence or frequency estimate. However, the comparison of answers gained with the RRT with objective outside information on behavior indicates that, while the RRT estimates of socially undesirable behavior are still too low relative to the actual prevalence, the RRT provides more accurate estimates than data gained through direct questioning (Tracy and Fox 1981; van der Heijden et al. 2000).

Research on the technique is now about forty years old (see Fox and Tracy's 1986 book summarizing the various techniques and issues with their use, and the recent meta-analysis by Lensvelt-Mulders, Hox et al. 2005). It has been tested in various implementations and settings. The techniques described above can be employed in different survey modes (face-to-face interviews, telephone surveys, mail questionnaires, etc.). However, their effectiveness may vary from mode to mode (as well as among specific implementations of the techniques). Much current research on these techniques therefore focuses on so-called mode effects, or the differing use of the methods across modes of administration as a potential determinant of the validity of responses (for a summary see Tourangeau, Rips and Rasinski 2000: Chapter 10).

Most of our knowledge about the method comes from its use in face-to-face interviews. This may be because there are some indications that the technique may be less useful in self-administered modes. Two important factors need to be clarified before the technique can be used effectively in self-administered modes. The first is a definition of the situations in which the costs of the technique (the respondent's time and effort, the larger required sample sizes) are likely to be worth the gain in perceived anonymity and therefore in answer validity. The second is knowledge about which of the technique's many implementations are best understood and most trusted by respondents. This factor is especially critical in situations in which no interviewer is present to answer questions about the technique.

For example, the RRT may lead to less improvement in answer quality in self-administered survey modes that are already perceived by the respondent to be minimally jeapordizing. These situations are ones in which both the risk of having a very incriminating answer connected to one's person are low and the risk of loss based on that connection are low (Nathan and Sirken 1988). The largest differences between estimates based on direct questioning and those based on the RRT have been therefore obtained for very sensitive behaviors (Himmelfarb and Lickteig 1982; Lensvelt-Mulders et al. 2005) under non-anonymous survey conditions or conditions in which the costs of the behavior's being revealed are high. The benefits of the technique for less sensitive topics and in situations in which respondents have reason to believe that their anonymity has been guaranteed appear to be lower.

For example, surveys administered in typical group experimental settings are more anonymous than mail surveys with respondent tracking. A typical finding in studies comparing direct-question versus RRT version of a mail survey is higher estimates of the sensitive behaviors with the RRT (Armacost et al. 1991; Buchman and Tracy 1982; Houston and Tran 2001). Similar differences have been reported for self-administered surveys conducted in non-anonymous situations in which the sensitive behavior may be sanctioned (such a survey on illegal behavior conducted in an office setting a la Dalton, Wimbush and Daily 1994; Wimbush and Dalton 1997; Wimbush, Shepard and Markham 1997). When anonymity is guaranteed, for instance with a self-administered paper-and-pencil questionnaire that is delivered to a collection box, few

differences are found between estimates derived from direct questioning and those derived from the RRT (Burton and Near 1995; Stem and Steinhorst 1984, Experiment 1).

Reckers et al. (1997) attempted to separate the de-biasing effects of anonymity and the RRT and anonymity. As one might expect, the lowest estimates of the prevalence of sensitive behavior were obtained with direct questioning under non-anonymous conditions in which a self-administered questionnaire was to be submitted with the respondent's name. While the respondents' not having their names directly connected to their survey answers already provided a strong boost to the prevalence estimates, the use of the randomized response technique had an additional effect (these estimates were higher still).

Another difficulty with implementing the RRT in a self-administered setting is the choice of an appropriate randomizing device. The randomizing device must be implemented in a way that makes the protection offered by the technique clear to the respondent. A novel, overly complex or abstract randomizing device may lead respondents to doubt the value of the randomizing procedure or, worse, to feel they are being tricked by the researcher into providing information under false pretenses. The result of the randomizing device must be perceived as truly random and unknown to the researcher in order to convince respondents that their "yes" answers can not be interpreted as an admission of guilt.

Convincing survey participants of this point is easiest if they control the randomizing procedure themselves. Employing such randomizing devices (especially those that make the process as transparent and controllable as possible, such as the colored balls that Tracy and Fox's 1981 respondents were asked to draw from a container) is least problematic in a face-to-face setting. However, several randomizing devices have also been employed in self-administered contexts. These are usually items assumed to be easily available to the respondent (a phone book or paper money, for example) or items sent to the respondent in advance. The devices most commonly employed in self-administered settings are serial numbers on a bill (Buchman and Tracy 1982; Houston and Tran 2001; Larkins, Hume and Garcha 1997; Robertson and Rymon 2001), the flip of a coin (Bailey, Hasselback and Karcher 2001; List et al. 2001), sheets of random numbers prepared by the researchers (Burton and Near 1995), numbers selected from the phone book

(Armacost et al. 1991), a cardboard spinner (Stem and Steinhorst 1984, Experiment 2), and the respondent's month of birth (Musch, Bröder and Klauer 2001).

Despite the higher and likely more accurate prevalence estimates obtained with the RRT in non-anonymous self-administered contexts, response rates are also typically lower for RRT than direct-questioning versions of a questionnaire (Buchman and Tracy 1982; Houston and Tran 2001). The authors of these studies suggest various reasons for the lower response rates in the RRT condition, for example that the technique is too complex and time-consuming, not having a randomizing device handy, and frustration with having to answer irrelevant questions. The focus group participants in Droitcour et al.'s (1991) qualitative study of the RRT technique even had strong questions about the method's seriousness.

Even among potential respondents who do decide to participate in an RRT survey, a substantial fraction likely do not follow the RRT instructions (Böckenholt and van der Heijden 2007; Lensvelt-Mulders and Boeije 2007; Musch, Bröder and Klauer 2001). This may occur because respondents do not believe that the technique protects their anonymity, understand that it provides useful answers or trust that a "yes" answer will not be construed as an admission of guilt. Several studies have found that a substantial proportion of respondents are not confident that the RRT actually protects their anonymity (Abernathy, Greenberg and Horwitz 1970; Krotki and Fox 1973; Landsheer, Van der Heijden and van Gils 1999; Soeken and Macready 1982), and that this confidence varies systematically with the sociodemographic characteristics of the respondent (Landsheer, Van der Heijden and van Gils 1999). Some authors hypothesize that lack of trust and non-compliance should be highest among those who have the most to lose and the least use for the anonymity provided by using the technique, i.e. those who have not committed the sensitive behavior (Fox and Tracy 1986; Nathan and Sirken 1988; Tourangeau and Yan 2007).

The Unmatched Count Technique

The UCT represents a similar approach, in that it does not allow the researcher to make conclusions about the respondents' behavior on the basis of their answers. With the UCT, the

respondents are asked directly about their own sensitive behavior at the same time as they are asked about a number of neutral or socially desirable behaviors. Estimation of the prevalence of the sensitive behavior requires an estimate of the aggregate prevalence of the other behaviors. This method therefore usually requires two samples: a reference sample that answers questions only about unthreatening behaviors and a sample that answers a sensitive question as well. For example, two lists of activities may be constructed. These lists are identical except for the fact that one list is longer by one behavior, namely the sensitive behavior of interest. Respondents are asked to report only the number of activities in which they have participated, but not which ones. Subtracting the average number of behaviors in the reference group from the average number of behaviors in the sensitive-question group provides an estimate of the frequency of the sensitive behavior, while preserving the anonymity of those in the sensitive-question group. For a more detailed description, see Dalton et al. (1994).

Various studies point to the effectiveness of the UCT at providing higher estimates of such sensitive behaviors as employee misconduct, shoplifting, hate crime victimization, and risky sexual behaviors (Dalton, Wimbush and Daily 1994; LaBrie and Earleywine 2000; Recker Rayburn, Earleywine and Davison 2003; Tsuchiya, Hirai and Ono 2007; Wimbush and Dalton 1997). We are not aware of any mail-survey studies in which the UCT was used, but several studies have examined the use of the UCT in other self-administered surveys. The same phenomenon is observed with the UCT as with the RRT: Less anonymous survey procedures and more sensitive topics are associated with larger differences between prevalence estimates based on direct questioning and those based on the UCT.

Some researchers have demonstrated the benefit of using the UCT in non-anonymous situations. Dalton et al.(1994) and Wimbush and Dalton (1997) estimated more illicit professional behaviors with the UCT than with direct questioning in a workplace survey. Tsychiya et al. (2007) compared the estimates gained with direct questioning and the UCT in a survey of online-access-panel members. They observed no significant differences between the two for a non-sensitive item, but found a significantly higher prevalence for a sensitive behavior using the UCT. Studies employing self-administered questionnaires in student populations have provided mixed results. Some have found a clear benefit to using the UCT (Recker Rayburn, Earleywine

and Davison 2003), further studies have found no effect (Ahart and Sackett 2004), and others have reported effects only for more sensitive behaviors (Anderson et al. 2007; LaBrie and Earleywine 2000). It is worth noting that one of these surveys was conducted outside the classroom in a presumably anonymous situation (Ahart and Sackett 2004), while another was conducted in class (Anderson et al. 2007) and others were conducted in class for extra credit, which presumably required that the respondents provide their names to the experimenter (LaBrie and Earleywine 2000; Recker Rayburn, Earleywine and Davison 2003).

The UCT has an important advantage over the RRT in that no randomizing device is required. This presumably both increases respondent trust in the technique and makes it less time-consuming. However, understanding of the technique may remain an issue. For example, Droitcour et al. (1991) report that a substantial fraction of their respondents used the technique incorrectly and/or reported little understanding of how it could lead to valid estimates. Tsychiya et al. (2007) found a larger benefit to using the UCT with more educated respondents (although it is unclear if this reflects greater understanding of the technique or more frequent commission of the sensitive behavior in question).

The Current Study

According to Lensvelt-Mulders et al. (2005: 323), a "thorough look at the literature on RRTs reveals that 35 years of research have not led to a consensus or a description of best practices." This is even truer of the use of RRTs in self-administered modes in general, and computer-assisted modes in particular. While many previous studies have shown that dejeopardizing techniques lead to less socially desirable responses than direct questioning, we are aware of only one study comparing computer-administered direct questioning to a computer-administered RRT (Musch, Bröder and Klauer 2001). That study examined the effect of the probability of answering a sensitive question in the use of the RRT.

There are, however, numerous other factors that may affect how dejeopardizing techniques are used by respondents in a computer-administered setting. This study is an exploration of the effectiveness of different implementations of RRT in such a setting. The RRT implementations

studied differ in the nature of the randomizing device employed and the amount of respondent control over that device. These different techniques are then applied to questions of varying sensitivity.

We also compare the use of the RRT to that of the UCT. We are aware of only one other study directly comparing the use of the two techniques (Wimbush and Dalton 1997). This study found that they produced similar estimates in a group-administration context. This comparison was, however, limited to the size of the estimates. We therefore compare the RRT implementations and the UCT along several dimensions. One is the benefit of using the technique, i.e. the difference between estimates obtained with the technique and those based on direct questioning. We also examine ease of use, and the degree to which respondents trust and believe they understand the technique in question. We do so in the non-anonymous self-administered context of an online access panel survey[1].

II. Method

Measurement techniques

Various measurement techniques are employed in our study to estimate the prevalence rates for six sensitive behaviors. In addition to the baseline method of direct questioning (DQ), we implemented five variants of the randomized response technique (RRT) and the unmatched count technique (UCT).

The RRTs all employed a forced-response design where the probability of being directed to answer the sensitive question was one half. Respondents were instructed to apply the randomization device and then, depending on the outcome, either to answer the sensitive question truthfully or automatically provide a "yes" answer. The probability of both events was 50 percent.

[1] See Postoaca (2006) for a definition of and more information about an online access panel.

All RRT respondents were instructed to generate six randomizing-device outcomes before they viewed a screen with the sensitive questions (except in the case of RRT Variant 2, see below). This procedure was intended to maximize compliance with the RRT instructions, since respondents would have to ignore the results they had already generated in order not to comply with the technique. The following variants were used:

1. Manual coin toss: Respondents were instructed to get a coin, toss the coin six times, and note the results of those tosses (heads or tails) one after the other on a sheet of paper. Then the rules for answering the six sensitive questions (answer the first question honestly if the first result is "heads"; simply answer with "yes" if the result is "tails"; and so on) appeared at the top of a new screen, along with a detailed example for their use. The sensitive questions followed on the same page. The basic rule ("Depending of the result of your Xth coin toss, please answer the question either ...", where X stands for the number of the question) was again displayed below each question.

2. Electronic coin toss: Respondents were instructed to press the "Toss Coin" button and answer accordingly (answer honestly if "heads", simply answer "yes" if "tails") for each of the questions that followed on the same screen. Pressing the button next to a question displayed the result of the toss ("heads" or "tails") and the relevant instruction ("Answer the question honestly" or "Simply answer with 'yes'"). The buttons were programmed in such a way that the respondents could press them as many times as they liked to convince themselves that random results were being generated. Results of a previous study with an electronic coin toss indicate that many respondents prefer it to a manual one (Lensvelt-Mulders, van der Heijden 2006).

3. Banknote serial numbers: Respondents were instructed to get two Euro bills and write the last three digits of their serial numbers one after the other on a sheet of paper. Then the rules for answering the six questions as a function of the parity of the numbers (answer the first question honestly if the first number is even; simply answer with "yes" if the

number is odd; and so on) and a detailed example were provided on a new screen. As in (1), the rules were repeated for each question.

4. Telephone numbers: The same as (3), except that the respondents were instructed to use the last three digits of two telephone numbers of their choice.

5. Banknote serial numbers, with the option to use telephone numbers instead: Similar to (3), but with the instruction to use telephone numbers if no banknotes were available.

The unmatched count technique (UCT) was implemented using six sets of statements, one set for each sensitive behavior. Each set contained 5 neutral statements and possibly also contained a statement on the sensitive behavior. In experimental group 1, the sensitive behaviors were omitted from sets 1, 2, and 4; in group 2 they were omitted from sets 3, 5, and 6. In other words, every respondent provided answers both for sets that contained sensitive statements and for sets that did not. This setup was intended to make the logic of the method as clear as possible to the survey participants. The respondents were instructed to provide for every set the number of statements they would agree with and an example was given (see the appendix for list of the UCT sets).

All the RRT variants and the UCT were introduced with the following text: “In order to insure that your answers remain absolutely anonymous, we ask you to carry out the following procedure. In doing so, please adhere strictly to the instructions, otherwise the explanatory power of all the data collected will be compromised.” Located between the instructions at the top of the screen and the sensitive questions, the RRTs contained an additional statement explaining that, “Since we do not know the results of your coin tosses [the serial numbers of your bank notes, etc.], we cannot know which kind of answer you provided. We can, however, calculate a frequency for the entire group containing all our respondents with the aid of probability calculus.” For the UCT this statement was: “We cannot know which of the individual statements apply to you. We can, however, calculate a frequency for the entire group containing all of our respondents with the aid of probability calculus.”

Data collection

The survey was implemented using the Unipark online research platform by Globalpark GmbH (see www.unipark.de). The 2,075 respondents were recruited from the German "Sozioland" access panel by Respondi AG (see www.sozioland.de) between August 1 and September 30, 2007 with an e-mail invitation. The sample used in this survey is by no means representative of either the general population or the internet-using population, which however is not critical because we are primarily interested in differences among experimental groups (Gosling et al. 2004; Reips 2002). Female respondents are overrepresented (65% female, 35% male). Furthermore, the respondents are relatively young (about 60% are below age 30) and well educated (60% have some higher education) compared to the general population.

Participation in the study was nominally anonymous because we were not able to track response to the questionnaire. However, although we do not know the identities of the survey participants, respondents may have had reason to believe that such knowledge was possible. Their identities are known to the panel organizers and they may have feared that their answers would be connected to those identities by "Socioland."

Respondents were randomly assigned to one of ten experimental groups (three groups who were questioned directly, one group for each of the five RRT variants tested, and two groups for the unmatched count technique) once they had activated the questionnaire. Table 1 provides an overview of the number of observations in each of the experimental groups. Since respondents were assigned to the experimental conditions on the fly at the time they started filling out the questionnaire, group sizes are subject to some random variation.

[Table 1 about here.]

Respondents were invited by e-mail to participate in a survey on "Security and Everyday Offenses". The questionnaire began with a set of basic demographic questions, followed by some questions on the respondent's living conditions and neighborhoods and an item battery measuring personality traits. Respondents then saw a text explaining that they were about to

answer questions that some might consider personal and assured that their responses would be treated confidentially.

The sensitive questions of interest, which addressed various illegal or frowned-upon behaviors, were then posed using one of the seven techniques outlined above. The questionnaire continued with some additional questions on the respondents' attitudes towards the sensitive behaviors and concluded with an evaluation of the respondents' perceptions of the technique they had used. The median time required to complete the questionnaire was 6.9 minutes.

Sensitive questions

Six behaviors of varying sensitivity were evaluated in this study. The questions were:

1. "Have you ever received too much change and knowingly kept it?" (Keeping too much change)
2. "Have you ever knowingly used public transportation without buying a ticket?" (Freeriding)
3. "Have you ever deliberately taken an article from a store without paying for it? (Shoplifing)
4. "Have you used marihuana in the past month?" (Marihuana use)
5. "Have you ever driven a car although your blood alcohol was almost certainly over the legal limit?" (driving under influence, DUI)
6. "Have you ever cheated on your partner?" (Infidelity)

The sensitivity levels of the first two behaviors, keeping too much change and freeriding, were expected to be relatively low. We assume that many people have engaged in such behaviors at least once in their lives and that most people would not worry about admitting to having done so in a direct question, since both behaviors can easily be legitimized. In the first case, it could be argued that it is the seller's own fault and the obligation for the customer to return the extra change is only weak. The second behavior could be legitimized by situational factors (for example, that one had no change to buy ticket, that the ticket machine was defective, that one

would otherwise have missed the bus, and so on). Shoplifting seems more sensitive since it clearly violates a social norm, could be fined, and is usually hard to legitimize. We think that adult respondents would be very reluctant to admit occasional stealing. However, most people who have at least once engaged in shoplifting probably did so in their teens and may excuse their behavior as youthful folly. We therefore assume that the sensitivity of question three is only moderate.

Using marihuana may be quite accepted nowadays. We nonetheless believe that question 4 is one of the most sensitive items in our study because it refers to current behavior that cannot be excused as "one of the silly things I once tried." Furthermore, although many people may not consider consuming marihuana unethical, many people may be reluctant to be considered habitual users. Driving under influence, the fifth item, also seems rather sensitive to us, since such behavior endangers others and is legally prosecuted. However, as in most cases above, the question aks whether the respondent has ever engaged in such behavior at least once. Long-ago commission may be excused as a one-time mistake. The last item, infidelity, although highly sensitive in presence of the respondent's partner or acquaintances, is probably only moderately sensitive in the context of a nominally anonymous interview.

III. Results

We evaluated the techniques according to various quality criteria, namely the estimates they provided, the ease of their use and respondent trust in the techniques. Before looking at the prevalence estimates obtained with the different measurement methods, we analyze the other indicators of how well the techniques worked in our survey. The indicators are (1) whether respondents felt they had understood the procedures, (2) whether respondents were convinced that the techniques guaranteed the anonymity of their answers, (3) the time required to read the instructions and answer the sensitive questions, and (4) the amount of non-response induced by the techniques. The results for these indicators are summarized in Table 2.

[Table 2 about here.]

(1) The instructions provided in the survey seem to have been clear to most respondents. The proportion of respondents who believed that they had completely understood the instructions lies between 80% and 93% depending on method.[2] However, note that even a single respondent who does not understand the instructions is a potential source of bias. This is especially true for the RRTs, which require all respondents to follow a relatively complex procedure exactly. The understanding rates of around 80% to 85% percent for the manual coin toss, banknote, and telephone number RRTs are therefore clearly unsatisfactory. Significantly better rates are achieved for the electronic coin toss RRT (93%) and the unmatched count method (92%).

(2) A second important aspect determining the success of the different methods is how many respondents believed in the protection offered by the procedures. If respondents remain suspicious in using the methods, they will behave self-protectively and either provide biased answers or refuse to respond to the questions (Tourangeau and Yan 2007). Table 1 displays the proportion of respondents who believed that the technique they used guaranteed the anonymity of their answers.[3] The results are disillusioning, with trust rates ranging from 15% for the electronic coin toss RRT to 29% for the unmatched count technique. However, also note that an additional approximately 20% of the respondents indicated that they did not worry about anonymity.

The lower level of trust in the electronic coin toss RRT compared to the other RRTs makes sense because, technically, the outcomes of the electronic randomization device could have been tracked and stored on the project computer (the difference is significant on the 10 percent level; Fisher's exact $p = 0.07$). Furthermore, it is interesting to see that the trust rate was higher for the unmatched count technique than for the RRTs (statistically significant with $p < 0.01$ even if the

[2] The question was: "In this survey, we used a special technique to ask you questions about some personal topics. Do you feel that you completely understood the instructions provided for of the method?" Possible answers were "yes", "no", and "don't know". Table 2 reports the proportion of "yes" answers among all answers.

[3] The question was: "Do you believe that, as we explained, we can indeed not arrive at your answers to these questions due to the use of this special method?" Possible answers were "yes", "no", "cannot say", and "have not thought about it". Table 2 reports the proportion of "yes" answers among all answers.

electronic coin toss RRT is omitted). Our interpretation of this result is that the instructions of the UCT are easier to understand than the instructions of the RRT, and that more complicated instructions make respondents more skeptical. This assertion is supported by the positive association between the "understood" and "trust" variables (phi = 0.13, $p < 0.001$).

(3) Table 2 also contains information on the time required to read the instructions and answer the sensitive questions as a function of the measurement technique used. Median times are reported (the median is preferred here over the arithmetic mean because there are large outliers, probably due to interruptions while completing the questionnaire). Clearly, direct questioning is the fastest method, with a median response time of 28 seconds for the six questions (about 5 seconds per question). Answer times increase by a factor of 5 to 6 in the case of the manual RRTs, because respondents have to get paper and pen and possibly a coin or banknotes. As expected, the electronic coin toss RRT is faster (97 seconds) than the manual RRTs, as is the unmatched count technique (116 seconds), but the answer times are still inflated by a factor of 3.5 to 4 compared to the direct questions.

(4) Finally, Table 2 reports non-response rates for the different techniques. Non-response is measured here as the proportion of respondents, who did not answer any of the six questions after having read the instructions (including respondents who answered some, but not all six questions somewhat increases the non-response rates but does not alter the pattern observed). The results are very clear: The methods that require respondents to engage in a mode shift (i.e. take a pen and paper, toss a coin, etc.) induce more non-response. The amount of non-response for these RRTs was between 5 and 9% (or 6.5 to 12% if respondents with at least one unanswered question are counted), whereas it was virtually nonexistent for the other methods (direct questions, electronic coin toss RRT, and UCT). The banknote and phone number forms of the RRT differed only in the amount of physical work and interruption they required of the respondents, either retrieving some banknotes from a wallet or other location or simply thinking of two phone numbers. Non-response was higher in the banknote condition, supporting the contention that mode shift was responsible.

To summarize, the manual RRTs (manual coin toss, banknotes, and telephone numbers) were problematic with respect to several domains. Many respondents did not understand the procedures and both answer times and levels of non-response were considerable. The electronic coin toss RRT, although easier to use and better understood by the respondents, induced less trust because, in principle, it would have been possible for the researchers to find out whether the innocuous or the sensitive question had been answered. The unmatched count technique (UCT), however, performed well compared to the RRTs on all of these measures.

We now turn to an analysis of the prevalence estimates for the sensitive behaviors. The estimates based on the different techniques are listed in Table 3 and also depicted in Figure 1. The true rates of the sensitive behaviors are unknown for this sample, so we cannot say which method provided the most valid results. There are nevertheless several interesting conclusions that can be drawn from the results. For example, it is immediately evident from Table 3 that the RRTs must have been used incorrectly by at least some respondents, since strongly negative estimates are observed (in brackets).[4] These negative estimates indicate that survey participants were responding with a "no" when proper use of the technique would have meant answering with an automatic "yes". Apparently, some respondents were reluctant to give an automatic "yes" answer, possibly because they feared that it could falsely be construed as an admission of guilt (Edgell, Himmelfarb and Duchan 1982; Lensvelt-Mulders and Boeije 2007; Nathan and Sirken 1988). The RRTs therefore strongly underestimate the rates of the behaviors in question.

[Table 3 about here.]

[Figure 1 about here.]

Assuming that the rate of affirmative answers in the "answer the sensitive question" condition of the RRT is at least as high the number of "yes" answers to the direct questions (which seems reasonable for these items), we can compute an estimate for the lower bound of the proportion of

[4] A rate, of course, cannot be negative. However, if the true prevalence is close to zero, the RRT can occasionally result in negative estimates due to random variation, even if the procedure is correctly applied by all respondents.

respondents who answered "no" although they were instructed to give an automatic "yes" answer. In our RRT designs, the expected value for the proportion of observed "yes" answers can be written as

$$\lambda = p \cdot \pi_x + (1-p) \cdot \pi_y \tag{1}$$

where p is the probability of being directed to the sensitive question ($p = 0.5$ in our design), π_x is the (unknown) probability of answering "yes" to the sensitive question, and π_y is the probability of answering "yes" to the innocuous question. In our case, π_y equals one because the "innocuous question" is a direct "yes" response. Solving the equation and substituting in the observed "yes" proportion $\hat{\lambda}$ (and setting $p = 0.5$ and $\pi_y = 1$), we obtain an estimate for π_x, namely:

$$\hat{\pi}_x = \frac{1}{p}\left(\hat{\lambda} - (1-p) \cdot \pi_y\right) = 2\left(\hat{\lambda} - 0.5\right) \tag{2}$$

These estimates are reported in the first panel of Table 3. We can slightly modify equation (1) and rewrite it as

$$\lambda = p \cdot \pi_x + (1-p) \cdot (1-\gamma) \cdot \pi_y \tag{3}$$

where γ is the probability of the respondent's disregarding the instructions and giving a "no" answer although an automatic "yes" answer would have been indicated according to the RRT instructions. Ideally, if all respondents follow the instructions, γ is zero. A proportion greater than zero is a real problem for RRT because it translates directly into the RRT prevalence estimate (the bias of the estimate is $-\gamma$ in our design).

If we substitute reasonable values for π_x, we can compute estimates for γ. In particular, if we assume that π_x is at least as high as the observed rate based on direct questions, denoted by $\hat{\pi}_x^{\mathrm{DQ}}$, the following relation holds for γ:

$$\hat{\gamma} \geq 1 - \frac{1}{(1-p)\cdot\pi_y}\cdot\left(\hat{\lambda} - p\cdot\hat{\pi}_x^{\mathrm{DQ}}\right) = 1 - 2\left(\hat{\lambda} - 0.5\cdot\hat{\pi}_x^{\mathrm{DQ}}\right) \qquad (4)$$

The right-hand side of (4) is simply the difference between the prevalence estimate based on direct questions and the RRT prevalence estimate in our design. The last row of Table 3 contains these lower bound estimates for the proportion of false "no" answers. It is evident that for at least some of the sensitive questions, the proportion must have been high. Furthermore, the proportion of false "no" answers is lowest for the apparently least sensitive items (see Table 4). This heavily biases the RRT estimates, but the exact amount of bias remains unknown.[5]

[Table 4 about here.]

While the RRT estimates seem to be unreliable due to strong false "no" biases (interestingly, the electronic coin toss RRT seems to be the least biased), the unmatched count technique (UCT) provides more reasonable estimates (the UCT estimates are computed as the mean differences between the counts for the two experimental groups). It seems noteworthy that in the case of marihuana consumption, the UCT yields a much higher estimate than the direct question. This makes sense given that the marihuana question is the only question that refers to current behavior ("within the past month" as opposed to "ever") and that it is probably the only question whose answers could be confirmed (through blood tests). However, this result should not be overinterpreted since standard errors are large for the UCT. For the marihuana question, the UCT standard error is 11.3, so the 95% confidence interval for the prevalence estimate ranges between 10 and 55 percent. The high sampling variance makes interpretation of the other prevalence

[5] Note that we can compute the absolute minimum for γ by setting π_x to zero, which yields $1 - 2\hat{\lambda}$ (and happens to be equal to the negative of the RRT prevalence estimate) in our design. This absolute minimum of false "no" answers is 31% for the marihuana question.

estimates difficult as well, but overall the UCT estimates seem to have more face validity than the RRT estimates.

IV. Conclusions

Our results indicate that the unmatched count technique (UCT) is superior to any of the implemented randomized response techniques (RRTs) along several dimensions. The procedure's instructions were generally better understood and more respondents believed that the technique guaranteed the anonymity of their answers. Furthermore, response times were shorter than for most RRT variants and non-response was almost nonexistent, compared to rates of up to 9 percent for RRT. Also, the prevalence estimates obtained by the UCT did not suffer from the strong negative biases observed for the RRT estimates. A drawback of the UCT is, however, the high sampling variance. UCT estimates are relatively inefficient compared to forced-choice RRT estimates, so that large samples are required to obtain precise estimates.

The often negative prevalence estimates obtained with the various versions of the RRT indicate that non-compliance with RRT instructions was frequent in our study. Similar results have been reported in other studies, especially those in which a forced-choice method directs respondents to provide an automatic "yes" answer (Lensvelt-Mulders and Boeije 2007; van der Heijden et al. 2000). The amount of noncompliance with instructions for forced-choice techniques has been found to increase with the sensitivity of the question (Edgell, Himmelfarb and Duchan 1982). Respondents clearly feel as if they are being asked to answer the sensitive question with "yes", versus simply being asked to react to the outcome of a randomizing device. Edgell and coauthors (1982) report that, "Despite [a] favorable endorsement of the randomizing device and the RRT procedure, some subjects indicated that they did not like to be directed to give embarrassing answers" (p. 97). There are two ways to address this issue. One is in using a (potentially less efficient) RRT implementation that poses an unthreatening alternate question to the respondents. Another is to address the respondents' reluctance directly. As Lensvelt-Mulders and Boeije (2007) write, "[t]o avoid cheating in a forced response questionnaire, it is necessary to

acknowledge the fact that being forced to answer contrary to one's own truth is difficult and sometimes even painful" (p. 604).

Interestingly, the RRT method that provided the highest prevalence estimates was the electronic coin toss method, in which the outcome of the randomizing device was computer generated. Respondents reported less trust in the technique than in the other RRT techniques, but apparently adhered more closely to the RRT instructions when using it. The same thought that may have led to the lack of trust they expressed, i.e. that the outcome of the electronic coin flip could be recorded and used to determine which question had been answered, also seemed to have disciplined respondents into giving an automatic "yes" answer when it was called for. That said, the electronic coin toss RRT method did not provide significantly higher prevalence estimates than direct questioning for an item and often provided lower estimates, so we may surmise that few participants used it to report true transgressions if they were directed to answer the sensitive question honestly. Despite the higher prevalence derived with the electronic coin-toss method than with the other RRT methods, it does not seem to be a useful method for this setting because it provides no benefit over direct questioning.

In contrast, many respondents seem not to have complied with the other RRT methods, neither providing an automatic "yes" answer when directed to do so nor answering the sensitive question honestly. While the necessity of time-consuming random outcome generation in RRT imposes strong costs on the respondent (as evinced by the high reaction times), the low level of understanding of the technique might lead respondents to ask why they should go to the greater effort required to answer the RRT questions. The result is higher non-response for RRT techniques that make use of a manual randomization device. These results are a strong argument against using the RRT in situations in which its use can not be explained in detail and practiced in advance, along with those in which someone can monitor that the device is actually being employed (even if the outcome is unknown to that person).

Based on our results, the UCT seems a more promising approach for a self-administered setting. However, much work remains to be done on determining optimal implementations of the technique. For example, there is some debate on the optimal prevalence of the non-target items

(Droitcour et al. 1991; Tsuchiya, Hirai and Ono 2007)[6]. Another direction for further research is the determination of an optimal length of item lists. All other things being equal, longer lists offer more protection to the respondent, but this gain is potentially offset by the memory load longer lists impose. It remains to be seen if the actual protection provided by longer lists translates into a sufficient increase in the perceived protection to justify the reduction in accuracy and efficiency (Tsuchiya, Hirai and Ono 2007).

Perhaps the most promising avenue for research on either the UCT or the RRT is the analysis of subgroups for whom the techniques are likely to be most effective. For example, work continues on the role of education for trust, understanding and correct use of the techniques, with the general finding that trust, understanding and compliance with the instructions increase with education (Landsheer, Van der Heijden and van Gils 1999; Tsuchiya, Hirai and Ono 2007). Other analyses have focussed on the power of further sociodemographic characteristics to predict the effectiveness of the techniques (Droitcour et al. 1991; Landsheer, Van der Heijden and van Gils 1999; Tracy and Fox 1981; Tsuchiya, Hirai and Ono 2007). Finally, some authors have suggested that answers among the 'guilty' are most likely to show the expected benefit of the techniques, since those who have not committed the sensitive behaviors will be reluctant to provide the false "yes" answer that is sometimes required (see above). There is mixed evidence with regard to this hypothesis (Droitcour et al. 1991; Landsheer, Van der Heijden and van Gils 1999; Tracy and Fox 1981), likely as a result of varying implementations of the techniques.

It is also worth noting that models have been developed to allow analysis of both RRT and UCT data with covariates. The models for the RRT have been described elsewhere (for example, Scheers and Dayton 1988; van den Hout and Kooiman 2006) . The analysis of the UCT data is possible with a count model (poisson, negative-binomial) with a treatment variable (0 = the set does not include the sensitive item, 1 = the set includes the sensitive item). The model includes the covariates and their interactions with the treatment variable. The interactions represent the

[6] It is clear that anonymity will be compromised if all of the non-sensitive items are generally agreed with by the UCT control group. If most respondents were in agreement with all of the non-target statements, someone who had committed the sensitive behavior would likely have to respond with the full number of statements, which would be a clear indication of having engaged in that behavior.

effects of interest. However, as noted above, rather large sample sizes will be required for these analyses. This remains the main impediment to effective use of the technique.

V. Appendix

Statements used in the unmatched count technique (UCT). The statements on the sensitive behaviors, which were included in only one group's list, are in brackets. The order of the statements was as listed.

Set 1: "I have been to Spain", "I have read the book 'Perfume'", "I have one or more cats as pets", "I am an active, playing member of a soccer club", ["I have received too much change and knowingly kept it at least once"], "I have a brother"

Set 2: "I have a cell phone", "I have more than one sister", "I read a newspaper every day or nearly every day", "I use an electronic planner to schedule my appointments", "I own a digital camera", ["I have knowingly used public transportation without having a valid ticket at least once"]

Set 3: "I have ordered an item online at least once", "I own a dog", "I have my own car", "My shoe size is larger than 10 ½", "I have lived in another country for longer than a month", ["I have deliberately taken an article from a store without paying for it at least once"]

Set 4: "I have been to America", "I have one or more children", "I drink coffee every morning or nearly every morning", "I wear glasses to correct my vision", ["I have used marijuana in the past month"], "I have a television set in my bedroom"

Set 5: "I work on a computer every day or nearly every day", "I am self employed", "I live in an apartment building", "I can drive a motorcycle", ["I have driven a car although my blood alcohol was almost certainly over the legal limit at least once"], "In my free time I go jogging at least once a week"

Set 6: “I have a dishwasher in my kitchen”, “I go shopping more than once a week”, “In my free time I listen to music at least once a week”, [“I have cheated on a partner at least once”], “I eat only vegetarian dishes”, “In winter I go skiing or snowboarding at least once”

VI. Literature

Abernathy, James R., Bernard G. Greenberg, and Daniel G. Horwitz. 1970. "Estimates of induced abortion in urban North Carolina." *Demography* 7:19-29.

Ahart, Alison M., and Paul R. Sackett. 2004. "A new method of examining relationships between individual difference measures and sensitive behavior criteria: Evaluating the unmatched count technique." *Organizational Research Methods* 7:101-114.

Anderson, Drew A., Angela M. Simmons, Suzanne M. Milnes, and Mitchell Earleywine. 2007. "Effect of response format on endorsement of eating disordered attitudes and behaviors." *International Journal of Eating Disorders* 40:90-93.

Armacost, Robert L., Jamshid C. Hosseini, Sara A. Morris, and Kathleen A. Rehbein. 1991. "An empirical comparison of direct questioning, scenario, and randomized response methods for obtaining sensitive business information." *Decision Sciences* 22:1073-1090.

Bailey, Charles D., James R. Hasselback, and Julia N. Karcher. 2001. "Research misconduct in accounting literature: A survey of the most prolific researchers' actions and beliefs." *Abacus* 37:26-46.

Barnett, Julie. 1998. "Sensitive questions and response effects: An evaluation." *Journal of Managerial Psychology* 13:63-76.

Böckenholt, Ulf, and Peter G. M. van der Heijden. 2007. "Item randomized-response models for measuring noncompliance: Risk-return perceptions, social influences, and self-protective responses." *Psychometrika* 72:245-262.

Boruch, Robert F. 1971. "Assuring confidentiality of responses in social research: A systematic analysis." *American Psychologist* 26:413-430.

Buchman, Thomas A., and John A. Tracy. 1982. "Obtaining responses to sensitive questions: Conventional questionnaire versus randomized response technique." *Journal of Accounting Research* 20:263-271.

Burton, Brian K., and Janet P. Near. 1995. "Estimating the incidence of wrongdoing and whistle-blowing: Results of a study using randomized-response technique." *Journal of Business Ethics* 14:17-30.

Dalton, Dan R., James C. Wimbush, and Catherine M. Daily. 1994. "Using the unmatched count technique (UCT) to estimate base rates for sensitive behavior." *Personnel Psychology* 47:817-828.

Droitcour, Judith, Rachael A. Caspar, Michael L. Hubbard, Teresa L. Parsely, Wendy Vischer, and Trena M. Ezzati. 1991. "The item count technique as a method of indirect questioning: A review of its development and a case study application." Pp. 185-210 in *Measurement errors in surveys*, edited by Paul Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman. New York: Wiley.

Edgell, Stephen E., Samuel Himmelfarb, and Karen L. Duchan. 1982. "Validity of forced responses in a randomized-response model." *Sociological Methods & Research* 11:89-100.

Folsom, Ralph E., Bernard G. Greenberg, Daniel G. Horvitz, and James R. Abernathy. 1973. "The two alternate questions randomized response model for human surveys." *Journal of the American Statistical Association* 68:525-530.

Fox, James Alan, and Paul E. Tracy. 1986. *Randomized response: A method for sensitive surveys*. London: Sage.

Goodstadt, Michael S., and Valerie Gruson. 1975. "The randomized response technique: A test of drug use." *Journal of the American Statistical Association* 70:814-818.

Gosling, Samuel D., Simine Vazire, Sanjay Srivastava, and Oliver P. John. 2004. "Should we trust web-based studies? A comparative analysis of six preconceptions about Internet questionnaires." *American Psychologist* 59:93-104.

Greenberg, Bernard G., Abdel-Latif A. Abul-Ela, Walt R. Simmons, and Daniel G. Horvitz. 1969. "The unrelated question randomized response model for human surveys." *Journal of the American Statistical Association* 64:520-539.

Himmelfarb, Samuel, and Carl Lickteig. 1982. "Social desirability and the randomized-response technique." *Journal of Personality and Social Psychology* 43:710-717.

Houston, Jodie, and Alfred Tran. 2001. "A survey of tax evasion using the randomized response technique." *Advances in Taxation* 13:69-94.

Krotki, Karol J., and Bonnie Fox. 1973. "The randomized-response technique, the interview, and the self-administered questionnaire: An empirical comparison of fertility reports." *Proceedings of the Social Statistics Section, American Statistical Association*:367-371.

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

LaBrie, Joseph W., and Mitchell. Earleywine. 2000. "Sexual risk behaviors and alcohol: Higher base rates revealed using the unmatched-count technique." *Journal of Sex Research* 37:321-326.

Landsheer, Johannes A., Peter Van der Heijden, and Ger van Gils. 1999. "Trust and understanding, two psychological aspects of randomized response: A study of a method for improving the estimate of social security fraud." *Quality & Quantity* 33:1-12.

Lara, Diana, Jennifer Strickler, Claudia D. Olavarrieta, and Charlotte Ellertson. 2004. "Measuring induced abortion in Mexico: A comparison of four methodologies." *Sociological Methods & Research* 32:529-558.

Larkins, Ernest R., Evelyn C. Hume, and Bikramjit S. Garcha. 1997. "The validity of the randomized response method in tax ethics research." *Journal of Applied Business Research* 13:25-32.

Lee, Raymond M. 1993. *Doing Research on Sensitive Issues*. London: Sage.

Lensvelt-Mulders, Gerty J. L. M., and Hennie R. Boeije. 2007. "Evaluating compliance with a computer assisted randomized response technique: a qualitative study into the origins of lying and cheating." *Computers in Human Behavior* 23:591-608.

Lensvelt-Mulders, Gerty J. L. M., Joop J. Hox, Peter G. M. van der Heijden, and Cora J. M. Maas. 2005. "Meta-analysis of randomized response research " *Sociological Methods & Research* 33:319-348.

List, John A., Charles D. Bailey, Patricia J. Euzent, and Thomas L. Martin. 2001. "Academic economists behaving badly? A survey on three areas of unethical behavior." *Economic Inquiry* 39:162-170.

Musch, Jochen, Arndt Bröder, and Karl Christoph Klauer. 2001. "Improving survey research on the World-Wide Web using the randomized response technique." Pp. 179-192 in *Dimensions of Internet Science*, edited by Ulf-Dietrich Reips and Michael Bosnjak. Lengerich, Germany: Pabst Science Publishers.

Nathan, Gad, and Monroe G. Sirken. 1988. "Cognitive aspects of randomized response." *Proceedings of the Section on Survey Research Methods, American Statistical Association*:173-178.

Postoaca, Andrei. 2006. *The Anonymous Elect: Market Research Through Online Access Panels*. Berlin: Springer.

Raghavarao, Damaraju, and Walter T. Federer. 1979. "Block total response as an alternative to the randomized response method in surveys." *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 41:40-45.

Rasinski, K. A., G. B. Willis, A. K. Baldwin, W. C. Yeh, and L. Lee. 1999. "Methods of data collection, perceptions of risks and losses, and motivation to give truthful answers to sensitive survey questions." *Applied Cognitive Psychology* 13:465-484.

Recker Rayburn, Nadine, Mitchell Earleywine, and Gerald C. Davison. 2003. "Base rates of hate crime victimization among college students." *Journal of Interpersonal Violence* 18:1209-1221.

Reckers, Philip M. J., Stephen W. Wheeler, and Bernard Wong-On-Wing. 1997. "A comparative examination of auditor premature sign-offs using the direct and the randomized response methods." *Auditing: A Journal of Practice & Theory* 16:69-78.

Reips, Ulf-Dietrich. 2002. "Standards for Internet-based experimenting." *Experimental Psychology* 49:243-256.

Robertson, Diana C., and Talia Rymon. 2001. "Purchasing agents' deceptive behavior: A randomized response technique study." *Business Ethics Quarterly* 11:455-479.

Scheers, N. J., and C. Mitchell Dayton. 1987. "Improved estimation of academic cheating behavior using the randomized-response technique." *Research in Higher Education* 26:61-69.

—. 1988. "Covariate randomized response models." *Journal of the American Statistical Association* 83:969-974.

Singer, Eleanor, Dawn R. von Thurn, and Esther Miller. 1995. "Confidentiality assurances and response: A quantitative review of the experimental literature." *Public Opinion Quarterly* 59:66-77.

Soeken, Karen L., and George. B. Macready. 1982. "Respondents perceived protection when using randomized response." *Psychological Bulletin* 92:487-489.

Stem, Donald E., and R. Kirk Steinhorst. 1984. "Telephone interview and mail questionnaire applications of the randomized response model." *Journal of the American Statistical Association* 79:555-564.

Tourangeau, R., Lance J. Rips, and Kenneth Rasinski. 2000. *The psychology of survey response*. Cambridge: Cambridge University Press.

Tourangeau, Roger, and Ting Yan. 2007. "Sensitive questions in surveys." *Psychological Bulletin* 133:859-883.

Tracy, Paul E., and James Allan Fox. 1981. "The validity of randomized response for sensitive measurements." *American Sociological Review* 46:187-200.

Tsuchiya, Takahiro, Yoko Hirai, and Shiguro Ono. 2007. "A study of the properties of the item count technique." *Public Opinion Quarterly* 71:253-272.

van den Hout, Ardo, and Peter Kooiman. 2006. "Estimating the linear regression model with categorical covariates subject to randomized response." *Computational Statistics & Data Analysis* 50:3311-3323.
van der Heijden, Peter G. M., Ger van Gils, Jan Bouts, and Joop J. Hox. 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* 28:505-537.
Warner, Stanley L. 1965. "Randomized-response: A survey technique for eliminating evasive answer bias." *Journal of the American Statistical Association* 60:63-69.
Wimbush, James C., and Dan R. Dalton. 1997. "Base rate for employee theft: convergence of multiple methods." *Journal of Applied Psychology* 82:756-763.
Wimbush, James C., J. M. Shepard, and S. E. Markham. 1997. "An empirical examination of the multi-dimensionality of ethical climate in organizations." *Journal of Business Ethics* 16:67-77.
Zdep, S. M., and Isabelle N. Rhodes. 1976. "Making the randomized response technique work." *Public Opinion Quarterly* 40:531-537.

Table 1. Sizes of the experimental groups

Group	Count	Percent
Direct questioning 1	193	9.30
Direct questioning 2	232	11.18
Direct questioning 3	218	10.51
RRT: Manual coin toss	185	8.92
RRT: Electronic coin toss	201	9.69
RRT: Banknotes	194	9.35
RRT: Phone numbers	218	10.51
RRT: Banknotes or phone numbers	236	11.37
Unmatched count 1	210	10.12
Unmatched count 2	188	9.06
Total	2075	100.00

Table 2. Quality measures for the different techniques

Experimental condition	N	Understood (in %)	Trust (in %)	Time (in sec.)	Non-Response (%)
Direct questioning	643	n.a.	n.a.	28	0.0
RRT: Manual coin toss	185	85.7	21.1	175	4.9
RRT: Electronic coin toss	201	92.9	14.7	97	0.5
RRT: Banknotes	194	82.3	20.6	169	8.8
RRT: Phone numbers	218	84.5	18.4	159	6.4
RRT: Banknotes or phone numbers	236	79.5	22.3	166	5.5
Unmatched count	398	91.8	28.6	116	0.3

Legend: Understood: percentage of respondents who felt they had completely understood the instructions; Trust: percentage of respondents who believed that the technique guaranteed the anonymity of their answers; Time: time spent reading the instructions and answering the six sensitive questions (median in seconds); Non-Response: percentage of respondents who did not answer any of the six questions.

Table 3. Prevalence estimates based on the different techniques (in percents; standard errors in parentheses)

Experimental condition	Keeping too much change	Free-riding	Shop-lifting	Marihuana use	DUI	Infidelity
Direct questioning	56.1	61.8	23.4	4.7	29.0	26.2
	(2.0)	(1.9)	(1.7)	(0.8)	(1.8)	(1.7)
RRT: Manual coin toss	61.4	45.1	4.0	[-31.4]	5.7	9.1
	(6.0)	(6.7)	(7.6)	(7.2)	(7.5)	(7.5)
RRT: Electronic coin toss	59.0	67.8	22.0	[-7.0]	8.0	20.0
	(5.7)	(5.2)	(6.9)	(7.1)	(7.0)	(6.9)
RRT: Banknotes	58.2	54.3	[-4.0]	[-44.5]	1.1	[-4.5]
	(6.1)	(6.3)	(7.6)	(6.8)	(7.5)	(7.5)
RRT: Phone numbers	59.6	59.6	15.8	[-38.7]	2.5	[-3.0]
	(5.6)	(5.6)	(6.9)	(6.5)	(7.1)	(7.1)
RRT: Banknotes or phone numbers	54.1	55.2	6.3	[-35.1]	[-6.3]	0.5
	(5.6)	(5.6)	(6.7)	(6.3)	(6.7)	(6.7)
Unmatched count	43.5	76.5	17.5	32.5	19.0	35.9
	(11.1)	(10.1)	(10.3)	(11.3)	(9.3)	(9.1)
RRTs combined	58.3	56.7	9.2	[-31.1]	1.9	4.4
	(2.6)	(2.6)	(3.2)	(3.1)	(3.2)	(3.2)
Lower bound for proportion of false "no" answers in RRTs	0.0	5.1	14.2	35.7	27.0	21.8

Legend: Keeping too much change: whether respondent once received too much change and knowingly kept it; Freeriding: whether respondent once knowingly used public transportation without buying a ticket; Shoplifting: whether respondent once deliberately took an article from a store without paying for it; Marihuana use: whether respondent used marihuana within the past month; DUI: whether respondent once drove a car although the blood alcohol was almost certainly over the legal limit; Infidelity: whether respondent once cheated on a partner.

Table 4. Perceived question sensitivity

	Keeping too much change	Free-riding	Shop-lifting	Marihuana use	DUI	Infidelity
Percentage agreeing with "It is alright to ..."	43.0	24.7	3.1	33.2	2.5	12.7
Percentage believing that many people in their neighborhoods had done.	59.2	46.2	13.5	8.5	32.8	27.4
Percentage believing that it would be uncomfortable for most people to admit that they had done.	27.8	25.7	81.2	56.3	53.8	82.8
Total sensitivity score	12.7	16.1	70.3	40.3	39.2	55.9

Legend: Total sensitivity score is calculated by adding the proportions of respondents who think that the behavior is not alright, that not many neighbors have done it, and that admitting it would be uncomfortable for most.

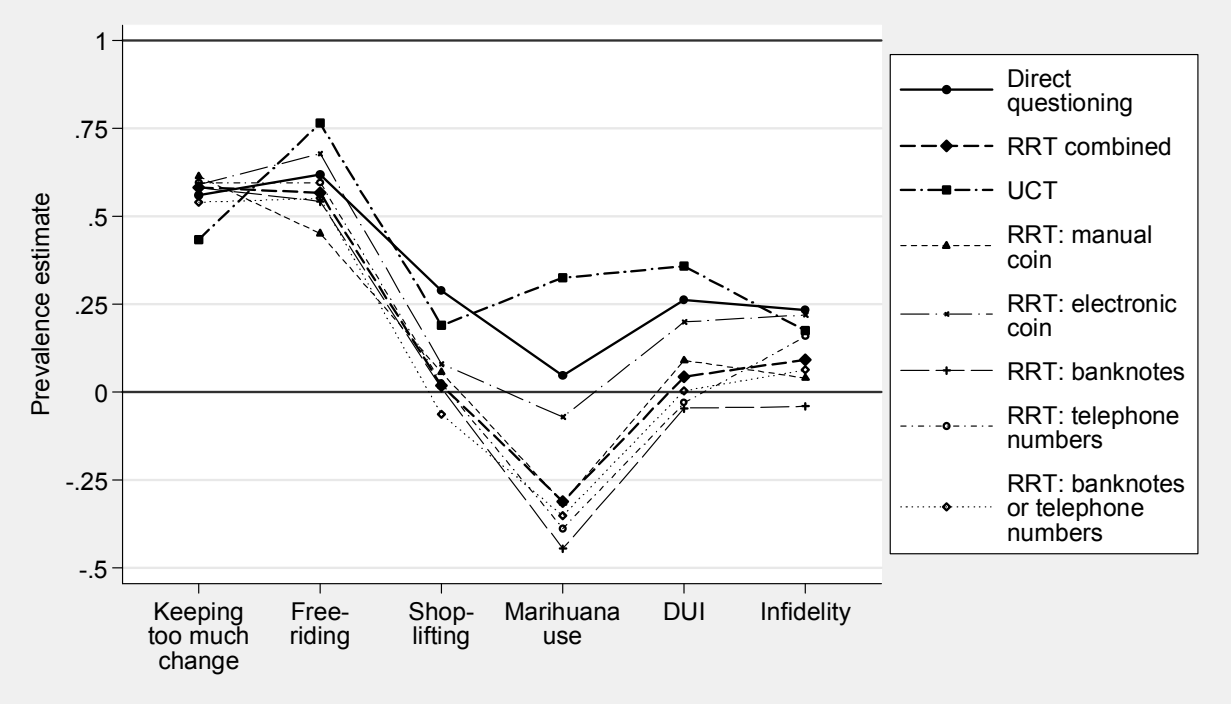

Figure 1. Prevalence estimates for the sensitive behaviors based on the different techniques. The behaviors are ordered according to the total sensitivity scores in Table 4.