Paradoxical Effects of Randomized Response Techniques

Leslie K. John, George Loewenstein, Alessandro Acquisti and Joachim Vosgerau

Invited Resubmission at Journal of Marketing Research

Abstract

Marketers, psychologists, sociologists, and policy makers have a variety of reasons for asking personal and even intrusive questions to consumers. For marketers, sensitive questions of interest pertain to social issues, transformative consumer behavior, and to taboo issues. By adding random noise to individual responses, randomized response techniques (RRTs) enhance privacy protection and, ideally, encourage disclosure of sensitive information. We show however, that RRTs can lead to paradoxical prevalence estimates: estimates that are lower (Experiments 1-5) and less valid (Experiments 1 and 5) than direct questioning. We provide evidence that these effects occur in part because the noise introduced by RRTs makes respondents concerned that innocuous responses will be interpreted as admissions. Specifically, Experiments 2 and 3 show that the paradox is alleviated by manipulations that reduce apprehension over response misconstrual; Experiments 4 and 5 test whether reduced self-protective responding is greatest when concern over response ambiguity is heightened because the stakes of responding affirmatively are high (Experiment 4) or because people have not, in fact, engaged in the target behavior (Experiments 5A&B).

Keywords: randomized response technique, survey research, privacy, disclosure

> "Was Judge Irwin ever broke -- bad broke?" I asked it quick and sharp, for if you ask something quick and sharp out of a clear sky you may get an answer you never would get otherwise.
>
> (Robert Penn Warren, All the King's Men, 1946)

Marketers, psychologists, sociologists, and policy makers have a variety of reasons for asking personal and even intrusive questions to consumers. For marketers, sensitive questions of interest pertain to social issues (prevalence of responsible behavior regarding public health issues or environmental issues), transformative consumer behavior (tobacco, alcohol, and other drug consumption, gambling, financial behavior), and to taboo issues (demand for adult entertainment; DeJong, Pieters, Fox, 2010). Given the goal of obtaining truthful responses to sensitive queries, is it best to ask questions directly – "quick and sharp" – or is it better to use a more elaborate technique that guarantees the respondent's privacy? Proponents of randomized response techniques (RRTs) would recommend the latter strategy; however, this paper suggests that the more straight-forward approach can often yield more valid responses.

THE RANDOMIZED RESPONSE TECHNIQUE (RRT)

Although RRTs take many different forms (Böckenholt & Van der Heijden, 2007; Campbell & Joiner, 1973; Martin G. de Jong, Pieters, & Fox, 2010; Park & Park, 1987; Pollock & Bek, 1976; Scheers, 1992; Tracy & Fox, 1980; Warner, 1965), they all add noise to individual responses, making it easier for individuals to admit to sensitive behaviors, thoughts, and feelings. For example, in the coin flip technique (Dawes & Moore, 1978; Warner, 1965) – one of the most common forms of the RRT – the interviewee is asked a sensitive question with response options "yes" and "no." Prior to answering the question, the interviewee is asked to flip a coin and to answer the question based on the outcome of the coin flip. If he flips 'heads,' he is instructed to respond 'yes,' *regardless* of whether he has actually engaged in the given behavior; if he flips 'tails,' he is instructed to answer the question truthfully. Since the interviewer cannot see the outcome of the coin flip, she cannot tell whether a given 'yes' response denotes an affirmative admission or a coin flip that has come up heads (or both). By correcting for the (known) probability of answering the focal question (i.e. in the coin flip technique, flipping tails), however, the researcher can deduce the population-wide prevalence of the behavior. In principle, therefore, the RRT can be used to estimate with greater accuracy the prevalence of behaviors that people are uncomfortable disclosing.

Two types of studies have been used to assess the effectiveness of RRTs (Tourangeau & Yan, 2007). 'Comparative studies' contrast prevalence estimates obtained using RRTs with those obtained via direct questioning (DQ); given the assumption that people tend to under-report the behavior in question because they are embarrassed admitting to it, the method that produces the higher estimate is presumed to be more valid. Consistent with this prediction, de Jong and colleagues recently introduced a polytomous item RRT model that increased reporting of sexual proclivities relative to DQ (Martin G. de Jong, et al., 2010) (for example, 14.9% of participants in the RRT condition indicated interest in using sex toys, compared to only 9.0% in the DQ condition). Individual validation studies, in contrast, compare prevalence rates of sensitive behaviors that are known to the researcher (e.g., registered voter records) to prevalence estimates obtained using the RRT (e.g., "Are you a registered voter?"). Reinmuth & Geurts (1975) underscored the importance of individual validation studies by noting that "for the methods of randomized response sampling to become useful to the marketing researcher, procedures must be designed for externally verifying the results." Yet to date, fewer than ten validation studies have been published (this manuscript would add three to that tally).

In an influential meta-analysis of 32 comparative studies and six individual validation studies published between 1965 and 2000, prevalence estimates gleaned using RRTs were typically higher than DQ prevalence estimates or closer to the true prevalence rates (Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005). The authors concluded that "…using randomized response questions in surveys on sensitive topics significantly improves the data quality of the surveys (p. 25, ibid)."

However, the results of this meta-analysis should be interpreted with caution. First, the authors identified 70 comparative and individual validation studies in the literature, but included only 38 studies (54%) that satisfied the authors' inclusion criteria and were accessible to the authors. Second, for studies in which boasting was expected (e.g., questions like "How often do you donate blood?"), RRT prevalence estimates were often found to be lower than DQ estimates. In these cases, the lower estimate (RRT) was taken to be the more valid estimate, contrary to the criterion applied to the other studies where the method that produces the higher estimate is presumed to be more valid[1]. The problem with this recoding is that a) it assumes that boasting is more likely for DQ than RRT; however, there is no research suggesting this to be the case, and b) it turns studies that potentially *disconfirm* the effectiveness of the RRT into studies *confirming* the effectiveness of the RRT. Third, and probably most problematic, a potential file-drawer bias—the disproportionate likelihood that studies finding null or conflicting results are not published (Ioannidis, 2008; Pashler & Harris, 2012; Rosenthal, 1979; Sterling, 1959)[2]—was not addressed;

[1] Since the authors did not identify the 32 comparative studies used in their meta-analysis, it is not clear whether such recoding was applied only to the Himmelfarb and Lickteig (1982) data or to all comparative studies included in the meta-analysis.

[2] Recent evidence suggests that such publication bias may be relatively common: in a poll of 2,155 behavioral scientists, 58% indicated that they had omitted study conditions in a paper; 67% indicated that they had selectively

neither were researchers in the field contacted for unpublished studies, nor was the file-drawer bias statistically accounted for. Given our own experience doing classroom demonstrations of RRT (which almost always produced lower estimates of sensitive behaviors with RRTs than with DQ), the RRT may generally perform worse than suggested by the meta-analysis.

PARADOXICAL RRT EFFECTS

Indeed, many researchers have found RRTs to produce the same or lower admission rates than point-blank questions and known prevalence estimates (Akers, Massey, Clarke, & Lauer, 1983; Begin & Boivin, 1980; Beldt, Daniel, & Garcha, 1982; Duffy & Waterton, 1988; Goode & Heine, 1978; Locander, Sudman, & Bradburn, 1976; Tamhane, 1981; Tracy & Fox, 1980; Williams & Suen, 1994; Wiseman, Moriarty, & Schafer, 1975) . For example, a national survey conducted by the Australian Bureau of Statistics to estimate the prevalence of drug use concluded that the RRT “did not significantly increase the number of affirmative responses to the controversial question, and was rather time-consuming” (Goode & Heine, 1978). Similarly, Weissman, Steer, and Lipton (1986) asked respondents whether they had used each of four illicit drugs (cocaine, heroin, PCP, and LSD) and found that drug usage prevalence estimates were equivalent across inquiry methods (RRT vs. DQ). Tracy and Fox (1980) conducted a validation study which used either the RRT or DQ to elicit from criminals self-reports of the number of times they had been arrested. Comparing the RRT and DQ responses to actual arrest records, the researchers found that the RRT produced higher admissions than the DQ for those who had been arrested only once, but had the opposite effect, yielding substantially *lower* prevalence estimates for individuals who had been arrested more than once.

reported dependent measures in a paper; and 50% indicated to having selectively reported studies that ‘worked’(L. K. John, Loewenstein, & Prelec, 2012).

Even more disconcerting, some RRT studies have yielded impossible admission rates. One nationally representative survey of marijuana use, for example, found RRT prevalence estimates not only to be consistently and considerably lower than DQ estimates (Brewer, 1981), but also to be *negative* – a “dead giveaway” that some respondents failed to respond ‘yes’ when instructed by the randomizer. Such impossible admission rates were also found by Holbrook and Krosnick (2010). In two nationally representative samples (N = 966 and N = 6,094), the RRT yielded estimates of 111.6% and 102.0% of voters having voted in the 2000 presidential election and the 2002 house of representatives election, respectively.

NON-ADHERENCE TO RRT INSTRUCTIONS AND SELF-PROTECTIVE BEHAVIOR

Why can the RRT produce paradoxical estimates—prevalence estimates that are lower than DQ or even impossible (negative or in excess of 100%)? Three reasons have been suggested (Campbell, 1987; Clark and Desharnais, 1998). First, participants may not understand the RRT instructions and answer questions with ‘no’ when the random device requires them to answer ‘yes’. Böckenholt and van der Heijden (2007) provided support for this hypothesis in a study on cheating in health insurance benefit claims. Clarity of instructions and participants’ education level were negatively related to the amount of ‘no’ answers when a ‘yes’ answer was required.

A second possibility is that, although participants may understand the instructions correctly, they do not want to give self-incriminating answers for something they did not do. So, the same noise that protects respondents’ privacy in the RRT could create apprehension: respondents who flip heads may fail to check ‘yes’ due to concern that their response might be interpreted as an affirmative admission. Böckenholt and van der Heijden (2004) call this self-protective behavior. Respondents who believed sanctions for cheating to be likely and severe were less likely to adhere to RRT instructions, but they were no more likely to have actually cheated on their health benefit claims (Böckenholt & Van der Heijden, 2007). These results appear strange, as one would assume that sanction severity and certainty would affect both non-adherence to RRT instructions and actual cheating. It is therefore not clear to what extent prevalence estimates for non-adherence were accurately discerned from prevalence estimates for actual cheating.

A related concern giving rise to self-protective behavior may be that it is impossible to unambiguously interpret individual affirmative responses in the RRT as admissions – a fact known to participants when responding – *psychologically,* the procedure might not feel so safe. For example, Brewer et al. (1981) surmised that RRTs may "introduce a sinister element into the proceedings and […] put people on their guard." In this vein, the elaborate instructions typically accompanying RRTs could lead respondents to infer that the behaviors in question are undesirable, and therefore to deny having engaged in them.

Non-adherence to RRT instructions—whether unintentionally due to actual misunderstanding or intentionally in the form of self-protective behavior—invalidates RRT prevalence estimates and can produce nonsensical admission rates (i.e. negative or greater than 100%). Several researchers have tried to assess the seriousness of this problem. Edgell, Himmelfarb, and Duchen (1982) surreptitiously recorded the outcome of the randomizer, and found 25% of respondents to answer 'no' when the randomizer had instructed them to answer 'yes.' Two alternative methods have been used to estimate the extent of non-adherence post-hoc without compromising participants' privacy.

Clark and Desharnais (1998) suggested using different probabilities for forced 'yes' and truthful answers. For example, the random device instructs one group of participants to answer questions truthfully in 75% of cases and to answer 'yes' in 25% of cases, whereas the other half of participants is instructed to answer questions truthfully in 25% of cases and to answer ‘yes’ in 75% of cases. By comparing the prevalence of ‘yes’ responses of both groups, the true prevalence of the behavior in question can be estimated, as well as the rate of non-adherence to RRT instructions. Using this methodology, Ostapczuk, Moshagen, Zhao, and Musch (2009) estimated a non-adherence rate of 20.2% among Chinese students asked about cheating in exams, and Ostapczuk, Musch, and Moshagen (2011) estimated a 38.9–55.2% non-adherence rate in a German patient sample asked about compliance with doctor-prescribed medicine intake. The Clark and Desharnais (1998) methodology works best when the two probabilities (e.g., 25% and 75%) are as far apart as possible; however, the further apart they are, the less participants’ privacy is protected (Clark and Desharnais, 1998). Furthermore, very large samples are required to yield reliable estimates of non-adherence rates.

The second post-hoc approach of estimating non-adherence rates uses latent class models e.g., (Böckenholt & Van der Heijden, 2007; Cruyff, Böckenholt, van den Hout, & Van der Heijden, 2008; Cruyff, van den Hout, Van der Heijden, & Böckenholt, 2007). Multiple items per construct allow for estimating the prevalence of latent, non-observable groups. Each respondent’s probability of belonging to a latent group of non-adherents is estimated, yielding a sample estimate of the overall prevalence of non-adherents. Using this multi-item latent class approach, Böckenholt and van der Heijden (2007) estimated 13-17% of non-adherence in their study on cheating in Dutch health insurance benefit claims. In two studies on the prevalence of sexual attitudes and behaviors, DeJong, Pieters, and Fox (2010) estimated a non-adherence rate of 10.3% in a Dutch sample, and DeJong, Pieters, and Stremersch (2012) estimated rates ranging from 5% (Brazil, Japan), over 20% (Netherlands), to 29% (India).

Estimating non-adherence rates with latent class models requires multiple constructs (i.e., sensitive topics) to be assessed with multiple items per construct, where the randomized response technique is applied to each item. Surveys using this methodology can thus become cumbersome to administer. Latent class models, like the methodology by Clark and Desharnais (1998) also require large sample sizes to estimate non-adherence rates.

To summarize: non-adherence to RRT instructions is a serious problem; post-hoc prevalence estimates range from 5-55%. However, whether and to what extent such non-adherence is driven by unintentional misunderstanding of instructions or intentional self-protective behavior is an open question.

In this paper, we investigate whether and why non-adherence in the form of self-protective behavior occurs in the RRT, and test a modification of the RRT procedure to reduce self-protective behavior. In the experiments reported herein, we consistently document a paradoxical effect of RRTs – estimates that are lower than DQ estimates (Experiments 1-5), less valid than DQ estimates (Experiments 1 and 5), or impossible (i.e., negative, Experiments 3-5). We start with a simple demonstration of the paradox using a validation methodology (Experiment 1) and then, in four subsequent experiments, provide empirical evidence for why the effect occurs. Experiments 2 and 3 show that the paradox is alleviated by manipulations that reduce apprehension over response ambiguity – in Experiment 2, by framing the target behavior as socially desirable; and in Experiment 3, by revising the RRT response option labels to subtly communicate that affirmative answers will not be construed as personal admissions. Experiments 4 and 5 test whether the reduction in self-protective responding (i.e., denial) in response to the revised label introduced in Experiment 3 is greatest when concern over response ambiguity is heightened. Supporting the idea that the RRT backfires because people are concerned that an affirmative response will be interpreted as an admission of having done so, we find that the relative advantage of the revised label over the traditional label is greater when the stakes of responding affirmatively are high (Experiment 4), and among people who have *not* engaged in the target behavior (Experiments

5A&B).

All studies include a DQ condition as a benchmark to compare the prevalence estimates generated by the RRT. Experiments 1 and 5 are validation studies, enabling prevalence estimates to also be compared to the true prevalence. These benchmarks allow us to test whether the RRT is generally effective at eliciting more valid prevalence estimates of sensitive behaviors. Thus, in addition to providing evidence for when and why the paradoxical effect of the RRT is exacerbated or attenuated, we can also draw conclusions about the RRT's practical utility to marketers or scholars who are interested in obtaining valid sensitive information from questionnaires and surveys. We report how we determined our sample size, all manipulations, all measures, and that no data were excluded from analysis.

EXPERIMENT 1

Experiment 1 was a two condition between-subjects validation study in which we contrasted prevalence estimates obtained using RRT versus DQ.

Method

Follow-up emails were sent to people who had participated in a previous set of studies in which we tested psychological factors that cause cheating. We chose email as the method of contact as we thought it would produce the highest response rates. In these studies, we followed a procedure similar to that introduced by Mazar, Amir, and Ariely (2008): participants answered trivia questions, were given an answer key and asked to report the number of questions they had answered correctly, and were paid based on these self-reported scores. Unbeknownst to the participants, the workbooks into which they had written their answers were collected and linked to their self-reported scores. Therefore, we were able to tell whether each participant had cheated (by overstating his or her score), and hence, decided to use this prior study as a source of validation data for an RRT experiment.

Overstatement scores (OS). To determine actual scores, a research assistant graded the workbooks. To assess score overstatement (cheating), we subtracted each participant's actual score from their self-reported score. Since participants answered between forty to fifty questions, an OS of 1 could reflect innocent error – for example, making an arithmetic mistake tabulating one's score. However, people were much more likely to overstate their score than to understate it, so we suspect that even low OSs are likely to indicate cheating. For example, the proportion of participants who overstated their score by exactly one (34.9%) was much higher than the proportion understating it by the same amount (5.7%; $\chi 2(1)=23.62$, $p<.0001$), most participants (58.1%) had an OS of one or higher.[3]

The OS is a conservative measure because not all forms of cheating could be detected by comparing workbooks to self-reported scores. For example, some people may have scribbled out incorrect answers in their workbook and replaced them with correct answers. It is unclear whether such participants wrote the correct answer before or after receiving the answer key (only the latter is cheating). In cases where responses seemed to be erased or changed, participants were given the benefit of the doubt, and were credited with having given the correct answer.

Approximately one month after having participated in one of the cheating studies, the 352 participants were sent an email in which they were asked to visit a link to a follow-up survey in exchange for a chance at a $100 amazon.com gift card. Participants were sent up to two reminder emails to participate. We stopped collecting data approximately two weeks after the final reminder

[3] We do not have OSs for twenty-three participants: six participants took their workbooks away at the end of the cheating study (instead of throwing them into the lab's garbage bin as had been requested of them); seventeen participants either illegibly recorded their names in the cheating study or did not leave their name in the follow-up survey, so we could not link their responses to their OSs. However, the proportion of participants for whom we do not have OS data was no different between the inquiry conditions.

email had been sent, at which point it had been about seven days since the last response.

There were 198 participants (51.5% male, 42.9% Caucasian, 35.4% Asian, 7.1% African American; 83.8% were students; Mean_{age}=23.8 years, SD=6.4 years, $\text{Median}_{\text{age}}$=22.0; all *NS* between-conditions), a response rate of 56.3%.

Upon clicking the link in the email, participants were randomly assigned to one of two inquiry conditions (DQ vs. RRT). Since there were differences in cheating between the conditions of the cheating studies, we stratified participants based on cheating condition. In addition, due to the greater error in prevalence estimates generated by RRTs, to maximize statistical power given the sample size, we oversampled RRT 2:1 relative to DQ.

All participants were instructed:

> In the ‘Reading Other People’s Minds Study’ you were asked to answer a series of questions. You then graded your own answers and reported your score. You therefore had the opportunity to overstate your actual score. We would like to know whether you overstated your score in this study. Please note that there will be no repercussions to responding ‘yes’ to this question.

The question, “Did you overstate your score in this study?” was accompanied with a yes/no response scale and was the same for all participants. Prior to answering the question, participants in the RRT condition were told:

> We have developed a procedure designed to better protect people’s privacy, and hence, to make you feel more comfortable answering the question. Using this procedure, from your answers, we will not be able to determine whether you personally engaged in the behavior, but from looking at a large number of people's answers, we will be able to determine the overall fraction of respondents who have engaged in the behavior.
>
> Instructions:
>
> 1. Please flip one coin one time. You may flip one of your own, or visit the following link to be directed to a virtual coin flip page <link to http://www.random.org/coins/>.
> 2. If you flipped:
>   - Heads, respond “Yes” to the question below, *REGARDLESS* of whether or not you’ve

done the behavior.

- Tails, answer the question honestly.

These, and the RRT instructions for all of our experiments, are similar to those used in previous RRT studies documenting positive effects of RRTs (see Appendix 1).

In all studies, to determine the aggregate admission rate in RRT (denoted by *t-hat* in the equation below), we adjusted the number of ‘yes’ responses (denoted Y) based on the expected likelihood of flipping heads (which in this case was 0.5, denoted by p in the equation below):

$$\hat{t} = \frac{Y - p}{1 - p}$$

Because of the additional variation introduced by the randomizing procedure, we widened the confidence intervals surrounding the prevalence estimates produced by the RRT, making our statistical tests conservative. This adjustment is based on the procedure outlined by Warner (1965); additional details are provided in Appendix 2. We used an intention-to-treat approach to data analysis: participants who dropped out of the survey prior to answering the focal question were assumed to have denied the behavior. However, the results across studies are similar, if not stronger, when we treat these participants as missing data (i.e., we assume that blank responses denote neither affirmations nor denials).

On the subsequent screen, participants were asked to provide their first name, followed by the first initial of their last name. In the cheating studies, participants had provided this information alongside their self-reported scores. Obtaining this information in the follow-up study enabled participants’ admissions or denials of cheating to be linked to their individual OSs. The study (as did all studies in this paper) concluded with standard demographic questions.

Results

Participants who completed the follow-up survey were significantly more likely to have cheated (i.e. to have an OS of 1 or greater) relative to those who did not complete the follow-up survey (56.1% of those who took the follow-up survey vs. 42.8% who did not take the follow-up survey; $\chi^2(1)=6.24$, p=.012). More importantly, however, among those who completed the follow-up, the percent of participants who cheated was not significantly different between the inquiry conditions (61.3% of DQ participants had cheated vs. 54.9% of RRT participants; $\chi^2(1)=.674$, p=.41) – i.e., random assignment worked.

The cheating prevalence estimate was 24.3% in the DQ condition, and only 4.8% in the RRT condition (t(196)=1.97, p=.05). In both inquiry conditions, the proportion of participants who admitted to having cheated was significantly lower than the true cheating prevalence within the given inquiry condition (RRT: admission rate=2.6%[4] vs. true prevalence=54.9%; p<.0001; DQ: admission rate=24.3% vs. true prevalence=61.3%; p<.005).

EXPERIMENT 2

Experiment 1 provides evidence that RRTs can generate lower and less valid prevalence estimates relative to DQ, a result that is consistent with several comparative studies. As noted in the introduction, previous researchers have posited several possible explanations for the effect; in this paper, we empirically test one of them: respondents who flip heads may fail to check “yes” due to concern that their response will be misinterpreted as an affirmative admission (self-protective behavior).

If lower RRT estimates are entirely driven by self-protective behavior, we would expect the paradox to disappear with a manipulation of social desirability. When it is socially desirable to

[4] This prevalence estimate (2.6%) is different than that reported above (4.8%) because the former is restricted to participants for whom we had OSs. Given that here we are comparing the prevalence estimate to the true prevalence, it seemed appropriate to only include those for whom we had OSs (and therefore who had been included in the calculation of the true prevalence).

respond affirmatively, respondents who flip ‘heads’ should not be concerned about having their ‘yes’ responses interpreted as affirmative admissions (and they might even *like* the ambiguity introduced by the RRT). Experiment 2 tests this idea. The hypothesis is that when a behavior is framed as socially undesirable, admission rates using RRT will be lower relative to DQ, but that, when it is framed as desirable, the difference will disappear. The experiment was a 2x2 between-subjects design in which we manipulated the social desirability of the behavior (desirable vs. undesirable) and the inquiry method (DQ vs. RRT).

Method

The study was conducted during the inauguration of a private Northeastern university’s satellite lab in a large office building. The lab was shared by a pool of researchers; we collected as much data as we could in the two days we had been allotted to run the study. Office workers (N=158) were recruited as they walked by the lab (62.6% female; 83.0% Caucasian, 12.1% African American, 2.8% Asian, 1.4% Hispanic, .7% Indian, Mage= 43.62, SD=12.35, Median=45; all *NS* between-conditions). They were offered a chance at a $100 gift card in exchange for completing a short, online survey. The survey consisted of an introduction (which formed the social desirability manipulation), followed by the focal question: “Have you ever texted while driving?” We decided to ask about this behavior primarily because it is neither highly sensitive nor highly innocuous and thus could be credibly framed as either socially desirable or socially undesirable (as described below). We also had a practical constraint: the office building administration would not allow us to ask a highly sensitive question.

Social desirability manipulation. At the beginning of the survey, participants read a short paragraph about text messaging. In the socially undesirable condition, the paragraph read:

> It is overwhelmingly clear that texting while driving is a deadly, selfish, activity. As highlighted in recent media coverage, texting while driving has caused numerous traffic accidents, many of them fatal. Texting is not only dangerous for the driver him or herself, but imposes risks on men, women and children in other cars who are not even enjoying the minor benefits of 'staying connected' at every moment.

In the socially desirable condition, the paragraph read:

> In our busy world, texting has become almost as essential as breathing to people who are socially connected or in professional positions. Although texting while driving is dangerous, it is increasingly common among people who are highly educated, overworked and socially connected. Penalties for texting while driving therefore threaten to strain the criminal justice system with a different group from those who usually get caught up in it: the professionally active and socially popular.

Inquiry method manipulation. Participants were asked: "Have you ever texted while driving?" using either DQ or RRT. The RRT instructions were the same as Experiments 1 and 2, except that participants were not provided with a link to a simulated coin flip page; instead, they were asked to flip a real coin – either one of their own, or the one provided in front of their computer terminal. We chose this hybrid mode of data collection to address any suspicion that may arise from an exclusively online coin flip.

Results

A logistic regression revealed a significant main effect of inquiry method (βRRT=-1.26, p = .017), and, of more relevance to our hypothesis, an interaction between inquiry method and social desirability (βRRT*Desirability=1.55, p=.028) (Figure 1). Follow-up testing revealed that when texting while driving was framed as socially undesirable, its estimated prevalence was marginally significantly lower using the RRT method relative to DQ (17.0% vs. 42.1%, t(77)=1.70, p=.09). When the behavior was framed as socially desirable, however, the RRT estimated prevalence was not significantly different from the DQ estimate (RRT=40.0%, DQ=33.3%, t(77)<1, p=.42).

Experiment 2 is consistent with the notion that RRTs backfire in part because they create concerns over how affirmative responses will be interpreted. When texting while driving was framed as socially undesirable, participants wanted to unambiguously show that they had not texted while driving, even if that meant disobeying the RRT instructions. Experiment 2 also helps to rule out the possibility that the results of Experiment 1 are simply the result of some kind of trivial methodological mistake. If this were the case, we should not expect social desirability to have made a difference.

Although the paradox was removed when the behavior was framed as socially desirable, we do not advocate such framing as a method of eliciting confessions -- doing so could introduce harmful unintended consequences. For example, although framing drug use as socially desirable is likely to increase RRT effectiveness, doing so could also increase drug use.

EXPERIMENT 3

Experiments 1 and 2 provide evidence that RRTs, although intended to facilitate disclosure, can instead generate prevalence estimates that are lower and less valid compared to DQ. In contrast to Experiments 1 and 2, in Experiment 3, we ask participants a highly sensitive question.[5] The RRT is believed to display its greatest advantage over DQ for highly sensitive questions (Lensvelt-Mulders, et al., 2005; Warner, 1965); thus, Experiment 3 is a conservative test of the basic hypothesis that the RRT can backfire.

[5] For Experiment 3 and 4, we pretested questions and chose those that were rated to be highly sensitive.

More importantly however, and similar to Experiment 2, Experiment 3 provides a test of the reason for the paradox, namely, apprehension over response ambiguity. Whereas Experiment 2 tests a framing manipulation, Experiment 3 tests a subtle revision to the RRT response labels that communicates the surveyors' understanding that individual 'yes' responses do not necessarily connote admissions. Edgell et al. (1982) noted that some participants who had been forced by the randomizer to say 'yes' would "giggle, smile, or in some other manner try to communicate that the answer they were giving was not true." Experiment 3's revised response label is designed to satisfy this apparent urge, thereby mitigating the paradox.

Method

In Experiment 3, as in Experiments 4-5, participants (N=162) were recruited through Mturk in exchange for a small payment and a chance to win $30. We collected as much data as we could in one day. We administered the survey through this platform because it enabled us to: a) collect data quickly and inexpensively; b) pose sensitive questions; and c) collect validation data (in Experiment 5).

In DQ, participants were asked "Have you ever cheated on a relationship partner?" followed by a yes / no response scale. There were two RRT conditions; in both, participants were given the standard RRT explanation and instructions. In the standard label (RRT-SL) condition, participants were presented with the same yes/no response scale as DQ. In the revised label (RRT-RL) condition, the response options were labeled: "yes/flipped heads" and "no." The RRT-RL condition was therefore designed to communicate to respondents that the surveyors understood that a "yes" response is not necessarily indicative of an affirmative admission to the behavior in question. Screen shots of the response labels, by condition, are shown in Figure 2.

Therefore, in terms of prevalence estimates, we predicted that RRT-SL < DQ (i.e. a replication of the paradoxical RRT effect), but that RRT-RL >= DQ.

Results

As predicted, admission rates were significantly lower in RRT-SL relative to DQ (RRT-SL=-21.0%, DQ=25.4%; t(100)=2.47, p=.02), but not in RRT-RL relative to DQ (RRT-RL=30.0%, DQ=25.4%; t(117)=0.50, p=.72) (Figure 3). In other words, when we signaled our understanding that “yes” responses may arise simply because a respondent flipped “heads,” the paradoxical effect of the RRT disappeared.

EXPERIMENT 4

Experiments 2 and 3 are consistent with the explanation that RRTs can backfire because they introduce apprehension over response ambiguity: the results suggest that respondents are uncomfortable giving an affirmative response simply because they flipped ‘heads.’ Notably however, even when this concern is addressed, the RRT generates estimates comparable to – but not better than – DQ. This is not particularly surprising: though the RRT-RL makes it clear that the researcher will not misinterpret a “yes” response to mean that the respondent definitely engaged in the behavior, the meaning of a “yes” response is still ambiguous, whereas the meaning of a “no” response is not. Although the RRT-RL does not, therefore, eliminate the ambiguous response problem, we can still make predictions about when the problem should be more or less serious. Experiments 4 and 5 test the notion that the advantage of the revised response label introduced in Experiment 3 should be greatest in situations in which concern over response ambiguity is highest.

Experiment 4 was a 3x2 between-subjects design in which we manipulated the inquiry method (DQ / RRT-SL / RRT-RL) and the stakes of responding affirmatively (high vs. low). The latter was manipulated by varying the extent to which participants were identifiable. We predicted that among participants who were relatively identifiable, the revised label would produce higher prevalence estimates relative to the standard label, and possibly also relative to DQ.

Method

Participants (N=691) were recruited through Mturk in exchange for a small payment. We planned to keep data collection open for about two weeks, and sought to obtain a sample size of at least 75 per RRT condition (the same procedure was used in experiments 5A and 5B).

For half of participants, we raised the stakes of responding affirmatively by making them identifiable: these participants were asked to provide their full name and email address at the outset of the study. The other half of participants were not asked to provide this information.

Participants were asked: "have you ever provided misleading or incorrect information on your tax return?" and were randomized to one of three inquiry methods: DQ, RRT-SL, or RRT-RL.

Results

Among participants asked to provide identifying information at the start of the study, 82.8% complied. The propensity to comply with the request did not differ by inquiry condition (as expected, since the identifiability manipulation preceded the inquiry manipulation).

Replicating Experiment 3, admission rates in RRT-SL (-17.3%) were significantly lower relative to both RRT-RL (7.6%; t(553)=2.82, p<.005) and DQ (9.6%; t(412)=3.66, p<.0005), but not in RRT-RL relative to DQ (t(411)=0.33, *NS*).

More interestingly, however, the benefit of the revised label over the standard label was driven by participants in the identified condition (Figure 4).[6] Among participants in the identified condition, admission rates in RRT-SL (-35.3%) were dramatically lower relative to both RRT-RL (RRT-RL=6.0%; t(275)=3.34, p<.005;) and DQ (7.4%; t(205)=3.83, p<.0005). However, the RRT-RL was not significantly different from the DQ t(204)=.18, *NS*). Prevalence estimates in the anonymous conditions were similar across conditions (Figure 4; DQ=11.8%; RRT-SL=0.7%; RRT-RT=9.4%; all comparisons *NS*).

The pattern of results is even more pronounced when the identified condition is restricted to the 82.8% participants who provided identifying information (admission rates among Ss who provided identifying information: DQ=8.6%; RRT-SL=-19.3%; RRT-RL=19.7%).

EXPERIMENTS 5A & 5B

In Experiment 4, the benefit of RRT-RL over RRT-SL was highest when the stakes of responding affirmatively were relatively high. However, in both Experiments 3 and 4, the revised label did not fare better than DQ.

We have proposed that the revised label works because it addresses concerns of response ambiguity – respondents who flip heads no longer feel as though they are inadvertently incriminating themselves by checking 'yes.' Therefore, one would expect the advantage of RRT-RL over RRT-SL to be pronounced among people likely to be particularly concerned over inadvertently incriminating themselves: those who have *not* engaged in the target behavior. However, the revised label may also make it easier for those who *have* engaged in the behavior to truthfully say 'yes' when required by the random device, since a 'yes' response is no longer unambiguously associated with an affirmative response – a kind of "facilitated justification."

[6] Note that in Experiments 4 and 5 due to negative admission rates, we were unable to conduct a logistic regression to explicitly test for an interaction as we had in Experiment 2.

According to the response ambiguity account, the improvement of the RRT-RL over RRT-SL should be greatest for those who have *not* engaged in the target behavior. On the other hand, according to facilitated justification, the improvement should be greatest for those who *have* engaged in the target behavior. Experiments 5A and 5B are validation studies that test these two accounts.

Experiment 5A

Experiment 5A was a 3x2 between-subjects design in which we manipulated inquiry method (DQ / RRT-SL / RRT-RL) and the proportion of participants who had engaged in the target behavior (high vs. low). Although the latter manipulation failed, the results are nonetheless interesting because the validation data enable the results to be separated based on whether the respondents had engaged in the target behavior (in this case, lying).

Method

Participants (N = 1,386) were recruited through Mturk in exchange for a small fixed payment. At the start of the study, participants were told that they would receive a bonus payment if they were completing the survey from a location within the United States. Between-subjects, we attempted to manipulate the lying rates by varying the incentive to lie: the bonus payment was either 5 cents (low) or 30 cents (high). Participants were next asked whether they were completing the survey from the United States. Finally, they were asked the following question, as a function of either DQ, RRT-SL, or RRT-RL: “Earlier in this survey, you were asked whether you are completing this survey from outside of the United States. Did you lie on this question? (please note that your response to the question below will not affect your payment for this study).” We asked about lying about physical location because a) it lent itself to validation data (ip addresses) and b) pretests indicated that questions about lying and cheating are very sensitive

Critically, and unbeknownst to participants, we collected their IP addresses to validate their location claims. A condition-blinded research assistant coded the country denoted by each IP address.

Results

Overall, only 20 percent of participants lied about their physical location, and there were no differences in lying rates as a function of the lying manipulation. Therefore, we collapse across this manipulation in subsequent analyses.

Consistent with Experiments 1-4, the prevalence estimate of RRT-SL (-30.0%) was significantly lower and less valid than that of DQ (16.3%; t(897)=31.04, p<.0001) and RRT-RL (-20.2%; t(1192)=2.48, p=.013). Although the RRT-RL yielded prevalence estimates that were significantly higher and more valid than RRT-SL, in contrast to Experiments 3 and 4, they were significantly lower and less valid than those obtained in DQ (t(891)=26.3, p<.0001).

More interestingly, when the results are broken down by whether participants lied (Figure 5), the difference in prevalence estimates between RRT-SL and RRT-RL is driven by those who did *not* lie (RRT-SL vs. RRT-RL among participants who did not lie: t(959)=3.23, p=.001; among participants who lied: *NS*). Participants who did not lie are likely to be particularly concerned over response ambiguity, because false incrimination tends to be worse than true incrimination - hence, the benefit of a manipulation designed to alleviate that apprehension should be pronounced among this sub-sample.

The results of Experiment 5A are consistent with the response ambiguity explanation; however, since the lying manipulation failed, this differential effectiveness could be attributable to selection bias (perhaps people who do not lie are more anxious in general about being misconstrued). Experiment 5B addresses this shortcoming through a stronger lying manipulation.

Experiment 5B

Similar to Experiment 5A, in Experiment 5B participants were first asked whether they were completing the survey from the United States; later in the survey, they were asked (as a function of DQ, RRT-SL or RRT-RL), whether they had lied about their physical location. While in Experiment 5A, all participants had an incentive to lie about their physical location, in Experiment 5B, we strengthened this manipulation by providing half of participants with an incentive to lie, and the other half of participants with *no* such incentive. The latter condition is expected to produce very low lying rates (why lie when there is no incentive to do so?), and hence, a relatively high proportion of participants with a heightened concern over response ambiguity when they are subsequently asked (as a function of RRT), whether they had lied. In addition, in Experiment 5B we used a non-financial lying incentive, which we thought would induce higher lying rates relative to a financial incentive. Thus, we predicted that the relative advantage of the RRT-RL over the RRT-SL will be driven by the no lying incentive conditions – i.e., the conditions consisting of a relatively large proportion of participants who have not engaged in the target behavior (i.e., lying).

Method

Experiment 5B was a 3x3 between-subjects design manipulating inquiry mode (DQ / RRT-SL / RRT-RL) and lying incentive (*lying incentive* vs. two *no lying incentive* conditions, described below). Participants (N=609) were recruited on Mturk; the procedure, described below, was very similar to Experiment 5A.

For efficiency of data collection, we advertised the study only to prospective participants who were *not* from the United States (however, as in Experiment 5A, we nonetheless covertly collected participants' IP addresses to check that they did not originate from the United States).

On the first page of the survey, participants were asked: "Are you completing this survey from a location within the United States?"

In the *lying incentive* conditions, participants were told on the first page of the survey:

> On the next page, you will be asked to answer a series of math questions. All participants who are completing this survey from a location within the United States may choose to opt out of having to answer the math questions, and proceed directly to the final portion of the survey. (If your answer to the question below is "yes" then on the next page of the survey, you will be given the opportunity to opt out of answering the math questions if you would like).

To reinforce this incentive, the response options to the location question were labeled: "Yes (I can skip the math questions and proceed directly to the final portion of the survey)" and "No (I won't have the option of skipping the math questions)".

In the *no lying incentive* conditions, respondents' physical location had no impact on having to complete the math questions. We included two such conditions, in the (unlikely) event that having a choice of completing the math questions interacted with the subsequent inquiry manipulation (it did not). In the *choice* condition, participants were first asked whether they were completing the survey from the United States; regardless of their answer, they could then choose to skip the math questions. In the *no choice* condition, participants were first asked whether they were completing the survey from the United States and were then directed to the math questions. The propensity to lie about one's physical location was similar across the two *no lying incentive* conditions, and did not interact with the subsequent inquiry mode manipulation; we therefore combined these two conditions in the rest of the reported results.

Next, participants completed the math questions, if applicable – participants who had specified that they did not want to complete the math questions skipped this portion of the survey.

Finally, participants were asked whether they had lied about their physical location in the question at the beginning of the survey (same wording as that used in Experiment 5A), as a function of either DQ, RRT-SL, or RRT-RL.

Results

The incentive to lie produced modestly higher lying rates relative to the *no lying incentive* conditions (lying rates = 28% in *lying incentive* condition; 13% in *no lying incentive c*onditions; p<.0001)[7], making tests of our hypotheses conservative.

Consistent with Experiments 1-5A, the prevalence estimate of RRT-SL (-33.6%) was significantly lower and less valid than DQ (18.0%; t(364)=5.92, p<.0001) and RRT-RL (-12.8%; t(485)=-1.94, p=.053). Although RRT-RL yielded prevalence estimates that were significantly higher and more valid than RRT-SL, in contrast to Experiments 3 and 4, but similar to Experiment 5A, they were significantly lower and less valid than those obtained the in DQ (t(363)= 3.88, p<.0005).

Consistent with the response ambiguity explanation, the relative advantage of the RRT-RL over the RRT-SL was driven by the *no lying incentive* condition (Figure 6A). When there was a relatively low proportion of participants who had engaged in the behavior, presumably apprehension over inadvertently incriminating oneself was high, and therefore RRT-RL fared significantly better than RRT-SL (t(330)=2.01, p=.05). By contrast, in the *lying incentivized* condition, there was no advantage of the RRT-RL over RRT-SL (t(153)=.055, *NS*).

Breaking the results down by participants who did versus did not lie (as in Experiment 5A), reveals

[7] One might wonder why any participants from the no lying incentive condition lied about their physical location. It is possible that participants had the intuitive belief that indicating that they were from the United States was the “desired” answer. At any rate, regardless of the reason, the important point is that the incentive manipulation worked, in the sense that it produced different lying rates.

a similar pattern (Figure 6B)[8]. Notably, the difference in prevalence estimates between RRT-RL and RRT-SL is driven by those who did *not* lie (RRT-SL vs. RRT-RL among participants who did not lie: t(358)=2.51, p=.012; among participants who lied: *NS*).

GENERAL DISCUSSION

RRTs are intended to make individuals more comfortable admitting to having engaged in sensitive behaviors. And yet, as shown in the present research, RRTs can produce prevalence estimates that are lower than DQ estimates (Experiments 1-5), less valid than DQ estimates (Experiments 1 and 5), or even impossible (i.e., negative, Experiments 3-5). Experiments 2 and 3 show that the paradox is alleviated by manipulations that reduce apprehension over response ambiguity – in Experiment 2, by framing the target behavior as socially desirable; and in Experiment 3, by revising the RRT response option labels to communicate that affirmative answers will not be construed as personal admissions. Experiments 4 and 5 show that the reduction in self-protective responding (i.e., denial) in response to the revised label introduced in Experiment 3 is greatest when concern over response ambiguity is heightened. Supporting the idea that the RRT backfires because people are concerned that an affirmative response will be interpreted as an admission of having done so, we find that the relative advantage of the revised label over the traditional label is greater when the stakes of responding affirmatively are high (Experiment 4), and among people who have *not* engaged in the target behavior (Experiments 5A&B).

In the present studies, RRT prevalence estimates were always lower than DQ estimates and actual prevalence rates. The results of our studies are thus in stark contrast to the conclusion of the

[8] And interestingly, the pattern even flips for those who did lie, although it may be a false positive since it is the only time we have observed such an effect.

meta-analysis by (Lensvelt-Mulders, et al., 2005) that “…using randomized response questions in surveys on sensitive topics significantly improves the data quality of the surveys (p. 25, ibid).” As we outlined in the introduction, however, we believe that the effectiveness of the RRT has been overestimated in this meta-analysis, because a) only a small subset of RRT studies was included, b) lower RRT than DQ estimates were assumed to be more valid when boasting was believed to occur, even though no research to-date has shown that boasting is more likely in DQ than the RRT, and c) the file-drawer problem was not accounted for. Given that our, and previous studies (Brewer, 1981; Holbrook & Krosnick, 2010) yielded impossible prevalence estimates (i.e., negative or in excess of 100%), RRTs may perform much worse than is concluded in the meta-analysis.

However, the problem of impossible prevalence rates can effectively be overcome by accounting post-hoc for non-adherence to RRT instructions. Studies that used the RRT-modification proposed by Clark and Desharnais (1998), and studies that estimated latent groups of non-adherents to RRT instructions, all yielded positive prevalence estimates (e.g., Ostapczuk et al., 2009; Ostapczuk, Musch, and Moshagen, 2011; Böckenholt and van der Heijden, 2007; DeJong, Pieters, and Fox, 2010; DeJong, Pieters, and Stremersch, 2012). Furthermore, the study by DeJong, Pieters, and Fox (2010) demonstrated that such post-hoc-corrected RRT prevalence estimates were higher than DQ estimates. While these are encouraging developments, it is still not clear how good these post-hoc approaches are in accurately estimating non-adherence (i.e., there are no validations studies that compare post-hoc estimated non-adherence rates to known non-adherence rates). More importantly, though, we are still lacking studies in which post-hoc corrected prevalence estimates are compared to known prevalence estimates—so-called individual validation studies—the gold standard of validation.

The results of our studies demonstrate that revising the response labels in the RRT almost entirely eliminates the problem of self-protective behavior. By labeling ‘yes’ responses as ‘yes/flipped heads,’ participants are no longer afraid of self-incrimination and follow the RRT instructions to provide an affirmative response when required to do so by the random device. Hence, combining both approaches—revised response labels and latent class models (which correct prevalence estimates by accounting post-hoc for the number of non-adherence to RRT instructions due to misunderstanding of the instructions)—may yield the most accurate prevalence estimates for sensitive consumer behaviors such as responsible behavior regarding public health or environmental issues, tobacco, alcohol, and other drug consumption, gambling, and financial behavior. This appears to be a promising avenue for future consumer behavior research.

We have demonstrated our effects across different survey administration procedures, questions, and participant populations, which attests to the robustness of our findings. Future research however, might systematically manipulate some of these variables to gain further understanding of the situations under which the technique is more likely to work. In addition, our experiments focus on the under-reporting of undesirable behavior; future research might study when and why RRTs might lead to over-reporting of desirable behaviors.

There is a psychological perspective, rarely if ever questioned in the literature, underlying the expected success of the RRT: The RRT assumes that people have a desire to tell the truth, but are deterred from doing so by qualms about self-incrimination. By diminishing these qualms, this implicit perspective assumes, the desire to tell the truth will have a greater impact, leading to more truthful responses. The assumption that, all other things held equal, people are motivated to tell the truth is, however, no more than that – an assumption. If violated, as one might expect it to be in some circumstances, then there is no reason to think that the RRT will have the intended effect. Respondents who do not like or trust the researcher, for example, might choose to willfully lie; and the RRT will do nothing to increase their willingness to tell the truth.

The fact that many people are admitting to self-incriminating or embarrassing behaviors under DQ might appear to suggest that people are, to some degree, motivated to tell the truth; however, other motives are possible. For example, if people told the truth under DQ because they suspected, whether rightfully, or due to some kind of suspicion or superstition, that a lie would be discovered, then RRT would very likely backfire, because it would make it easier for people to avoid telling the truth. We have no way of knowing whether our studies found more consistent paradoxical effects of the RRT for reasons that had to do with participants' motivations, but that certainly is one possibility; the college sophomores of earlier times, who make up the bulk of participants in earlier studies, may have had very different motivations from the college students and internet recruits in our studies.

The present research demonstrates that RRTs can backfire despite (or perhaps in spite of) the fact that they provide greater privacy protection. Broadly, this finding shows how factors that should make people more forthcoming with information can suppress disclosure – in much the same way as confidentiality assurances have been found to make people *less* willing to respond to surveys on sensitive subjects (Singer, Hippler, & Schwarz, 1992). Researchers have also begun to demonstrate the opposite effect – factors that should make people less forthcoming with information can, paradoxically, increase divulgence (L. John, Acquisti, & Loewenstein, 2011). For example, people seem naturally more comfortable disclosing personal information on unprofessional-looking web sites, even though such sites are particularly prone to abusing the information that is disclosed on them. Taken together, these findings highlight how people's willingness to divulge can be at odds with the objective consequences of information revelation.

References

Akers, R. J., Massey, J., Clarke, W., & Lauer, R. M. (1983). Are Self-Reports of Adolescent Deviance Valid? Biochemical Measures, Randomized Response, and the Bogus Pipeline in Smoking Behavior. *Social Forces, 62*, 234-251.

Begin, G., & Boivin, M. (1980). Comparison of Data Gathered on Sensitive Questions via Direct Questioning, Randomized Response Technique, and a Projective Method. *Psychological Reports, 47*, 743-750.

Beldt, S. F., Daniel, W. W., & Garcha, B., S. . (1982). The Takahasi-Sakasegawa Randomized Response Technique. *Sociological Methods and Research, 11*, 101-111.

Böckenholt, U., & Van der Heijden, P. G. M. (2004). *Measuring noncompliance in insurance benefit regulations with randomized response methods for multiple items.* Paper presented at the Proceedings of the 19th International Workshop on Statistical Modelling, Florence, Italy.

Böckenholt, U., & Van der Heijden, P. G. M. (2007). Item Randomized-Response Models for Measuring Noncompliance: Risk-Return Perceptions, Social Influences, and Self-Protective Responses. *Psychometrika, 72*(2), 245-262.

Brewer, K. R. W. (1981). Estimating Marihuana Usage Using Randomized Response: Some Paradoxical Findings.

Campbell, C., & Joiner, B. L. (1973). How to get the answer without being sure you've asked the question. *The American Statistician, 27*(5), 229-231.

Cruyff, M. J., 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.

Cruyff, M. J., van den Hout, A., Van der Heijden, P. G. M., & Böckenholt, U. (2007). Log-linear randomized-response models taking self-protective response behavior into account. *Sociological Methods and Research, 36*(2), 266-282.

Dawes, R., & Moore, M. (1978). Guttman scaling orthodox and randomized responses. In F. Peterman (Ed.), *Attitude Measurement*.

de Jong, M. G., Pieters, R., & Fox, J.-P. (2010). Reducing Social Desirability Bias Through Item Randomized Response: An Application to Measure Underreported Desires. *Journal of Marketing Research, 47*(1), 14-27.

de Jong, M. G., Pieters, R., & Stremersch, S. (2012). Analysis of Sensitive Questions Across Cultures: An Application of Multigroup Item Randomized Response Theory to Sexual Attitudes and Behavior. *Journal of Personality and Social Psychology, 3*(543-564).

Duffy, J. C., & Waterton, J. L. (1988). Randomised Response vs. Direct Questioning: Estimating the Prevalence of Alcohol-Related Problems in a Field Survey. *Australian Journal of Statistics, 30*(1), 1-14.

Edgell, S. E., Himmelfarb, S., & Duchen, K. L. (1982). Validity of Forced Response in a Randomized Response Model. *Sociological Methods and Research, 11*, 89-110.

Goode, T., & Heine, W. (1978). Surveying the extent of drug use. *Statistical Society of Australia Newsletter, 5*, 1-3.

Himmelfarb, S., & Lickteig, C. (1982). Social desirability and the randomized response technique. *Journal of Personality and Social Psychology, 43*(4), 710-717.

Holbrook, A. L., & Krosnick, J. A. (2010). Measuring Voter Turnout By Using the Randomized Response Technique: Evidence Calling Into Question the Method's Validity. *Public Opinion Quarterly, 74*(2), 328-343.

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology, 19*(5), 640-646.

John, L., Acquisti, A., & Loewenstein, G. (2011). Strangers on a plane: context-dependent willingness to divulge personal information. *Journal of Consumer Research*.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices with Incentives for Truth-telling. *Psychological Science, 23*(517-523).

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: 35 years of validation. *Sociological Methods and Research, 33*(319), 319-348.

Locander, W., Sudman, S., & Bradburn, N. (1976). An investigation of interview method, threat, and response distortion. *Journal of American Statistics, 71*, 269-275.

Mazar, N., Amir, O., & Ariely, D. (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research, 45*(6), 633-644.

Ostapchuk, M., Moshagen, M., Zhao, Z., & Musch, J. (2009). Assessing Sensitive Attributes Using the Randomized Response Technique: Evidence for the Importance of Response Symmetry. *Journal of Educational and Behavioral Statistics, 34*(2), 267-287.

Ostapchuk, M., Musch, J., & Moshagen, M. (2011). Improving self-report measures of medication non-adherence using a cheating detection extension of the randomised-response-technique. *Statistical Methods in Medical Research, 2011*(20), 489-503.

Park, J. W., & Park, H. N. (1987). A new randomized response model for continuous quantitative data. *Proceedings of the College of Natural Science, 12*, 33-44.

Pashler, H., & Harris, C. R. (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science, 7*(6), 531-536.

Pollock, K. H., & Bek, Y. (1976). A Comparison of three Randomized Response Models for Quantitative Data. *Journal of the American Statistical Association, 71*, 884-886.

Reinmuth, J. E., & Geurts, M. D. (1975). The Collection of Sensitive Information Using a Two-Stage, Randomized Response Model. *Journal of Marketing Research, 12*, 402-407.

Rosenthal, R. (1979). The File Drawer Problem and Tolerance for Null Results. *Psychological Bulletin, 86*(3), 638.

Scheers, N. J. (1992). Methods, plainly Speaking: A Review of Randomized Response Techniques. *Measurement and Evaluation in Counseling and Development, 25*, 27-41.

Singer, E., Hippler, H.-J., & Schwarz, N. (1992). Confidentiality assurances in surveys: Reassurance or threat? *International Journal of Public Opinion Research, 4*, 256-268.

Sterling, T. D. (1959). Publication Decisions and and Their Possible Effects on Inferences Drawn from Tests of Significance - or Vice Versa. *Journal of the American Statistical Association*, 30-34.

Tamhane, A. C. (1981). Randomized Response Techniques for Multiple Sensitive Attributes. *Journal of the American Statistical Association, 76*, 916-923.

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

Tracy, P. E., & Fox, J. a. (1980). The validity of randomized response for sensitive measurements. *American Sociological Review, 46*, 187-200.

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

Weissman, A. N., Steer, R. A., & Lipton, D. S. (1986). Estimating illicit drug use through telephone interviews and the randomized response technique. *Drug and alcohol dependency, 18,* 225-233.

Williams, B. L., & Suen, H. (1994). A Methodological Comparison of Survey Techniques in Obtaining Self-Reports of Condom-Related Behaviors. *Psychological Reports, 7*, 1531-1537.

Wiseman, F., Moriarty, M., & Schafer, M. (1975). Estimating Public Opinion with the Randomized Response Model. *Public Opinion Quarterly*, *39*, 507-513.

Figure 1. Prevalence estimates in Experiment 2.

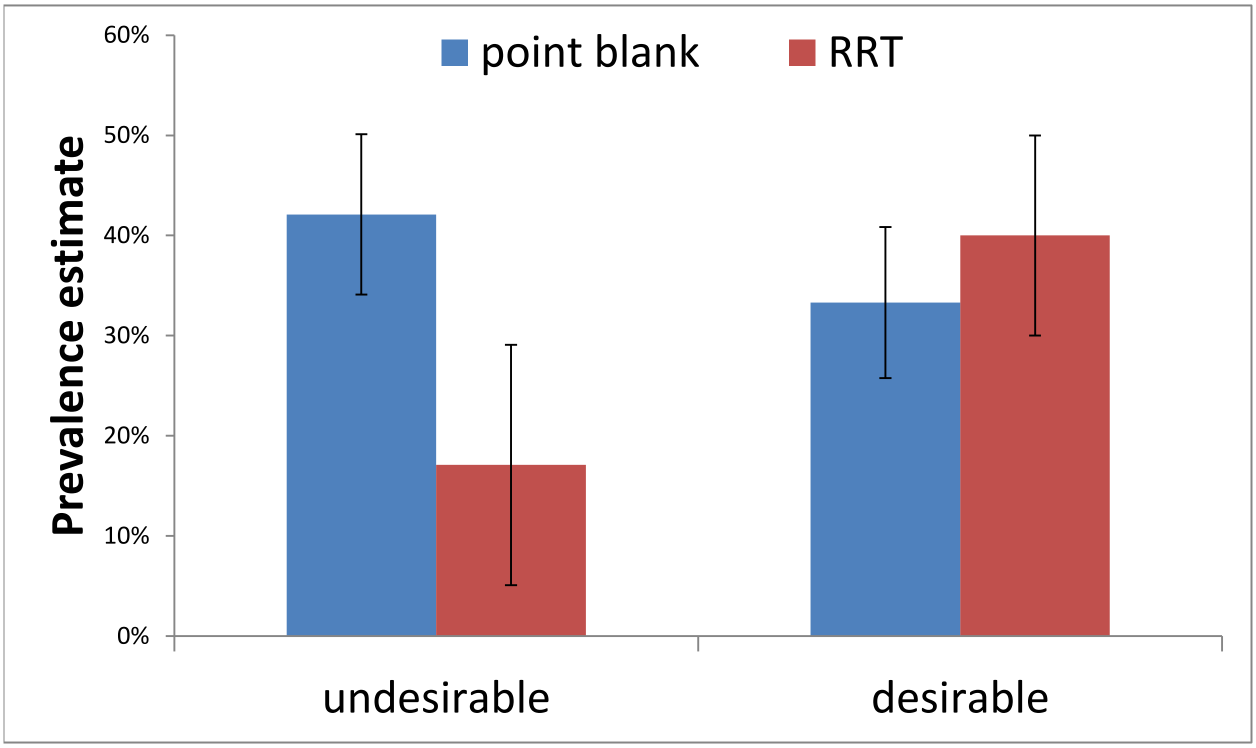


Figure 2. Screen shots of response labels used in Experiment 3 (top panel depicts labels used in DQ and RRT-Standard Label; bottom panel depicts labels used in RRT-Revised Label).

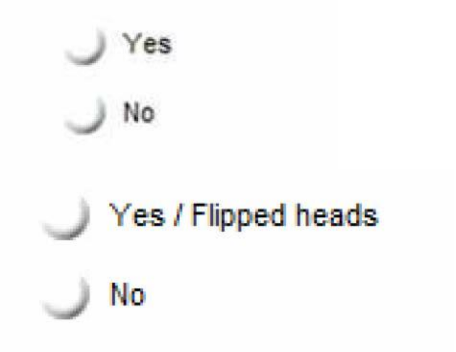


Figure 3. Prevalence estimates in Experiment 3.

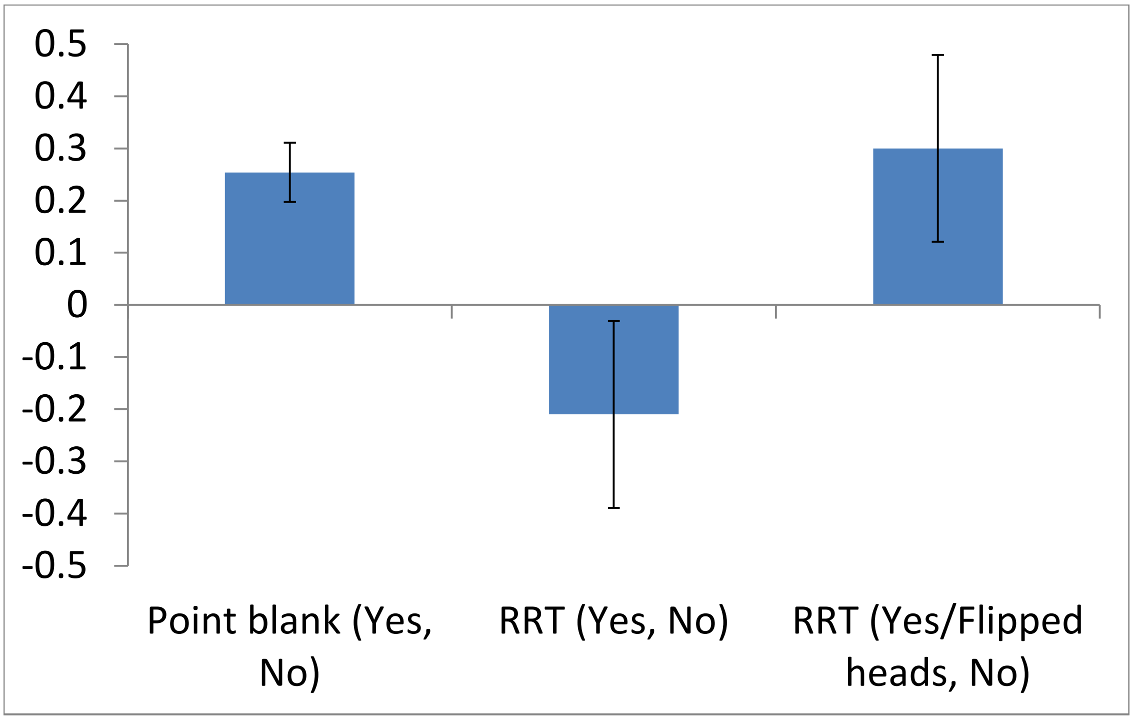


Figure 4. Prevalence estimates in Experiment 4.

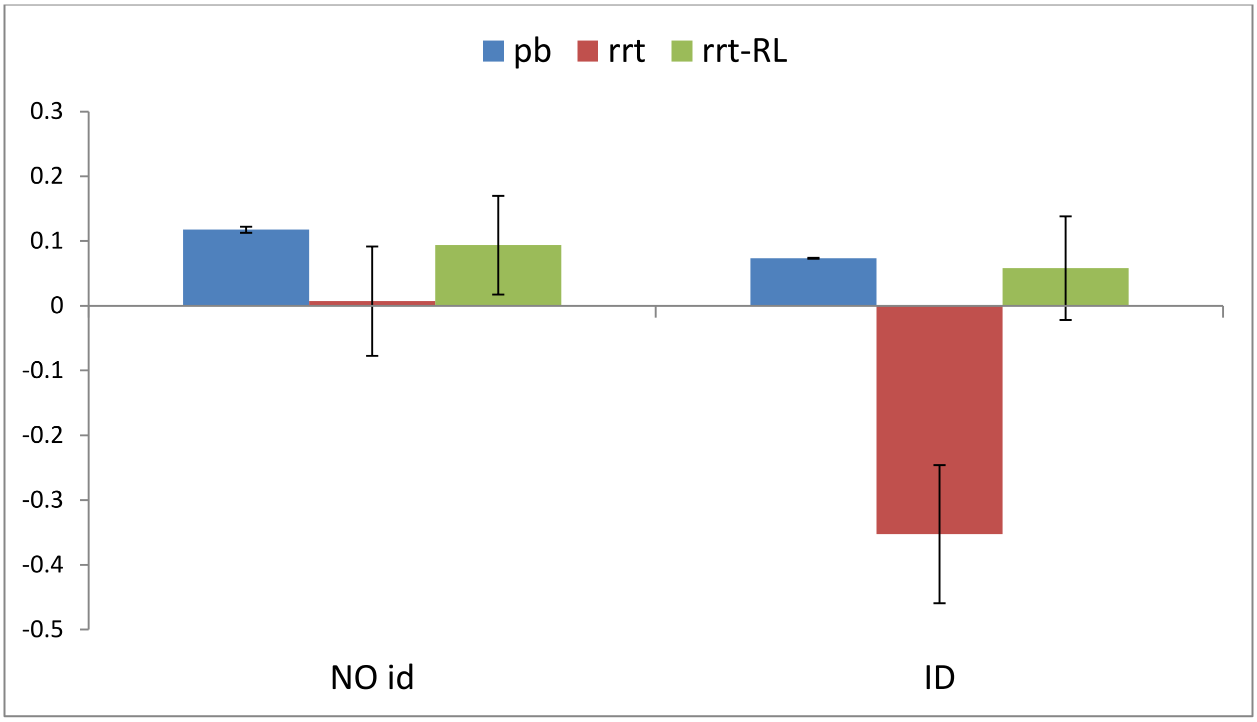


Figure 5. Prevalence estimates in Experiment 5A

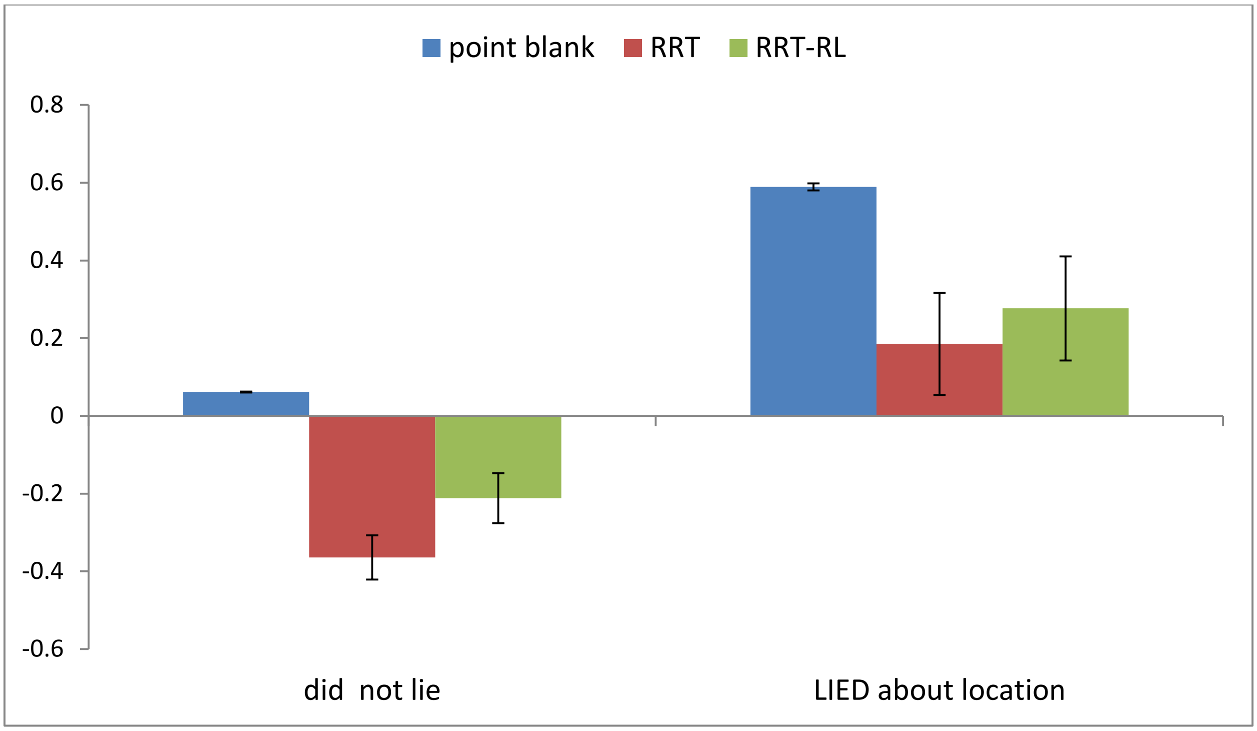


Figure 6A. Prevalence estimates in Experiment 5B. Solid blue lines denote true prevalence, by incentive to lie manipulation.

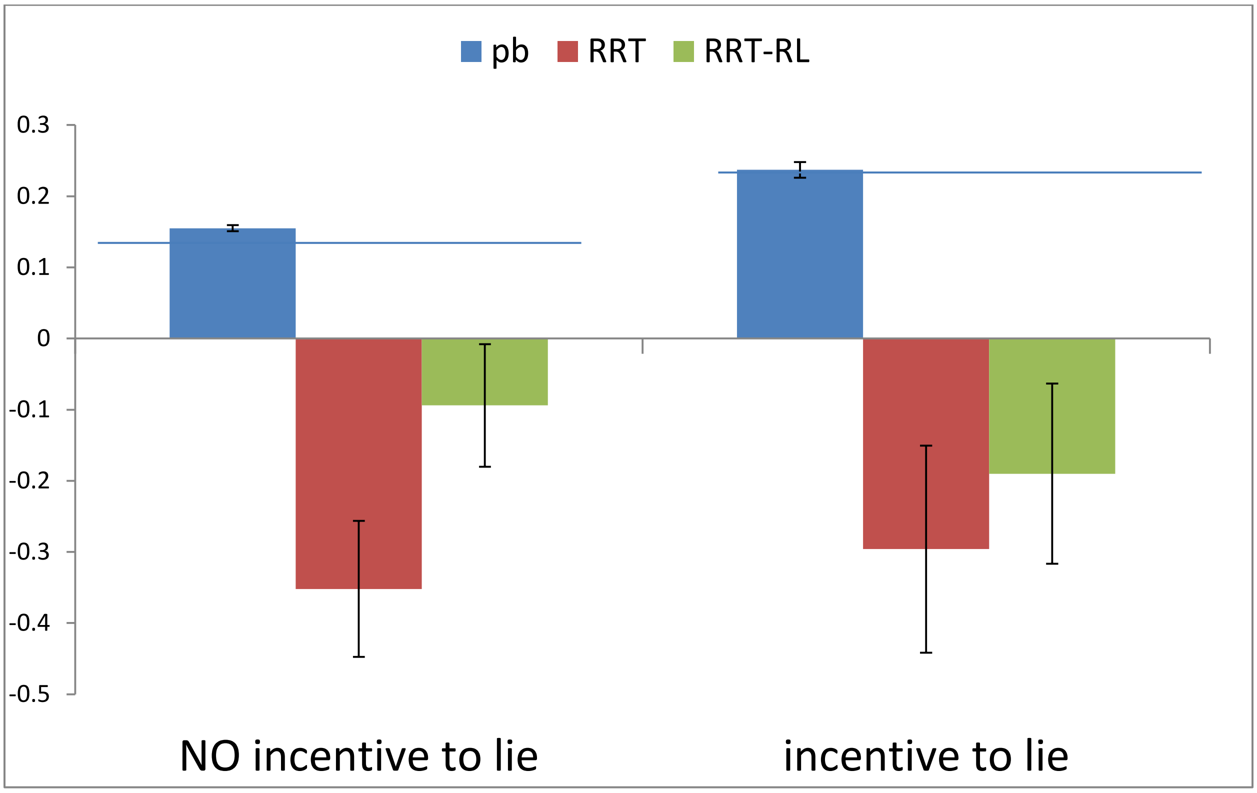
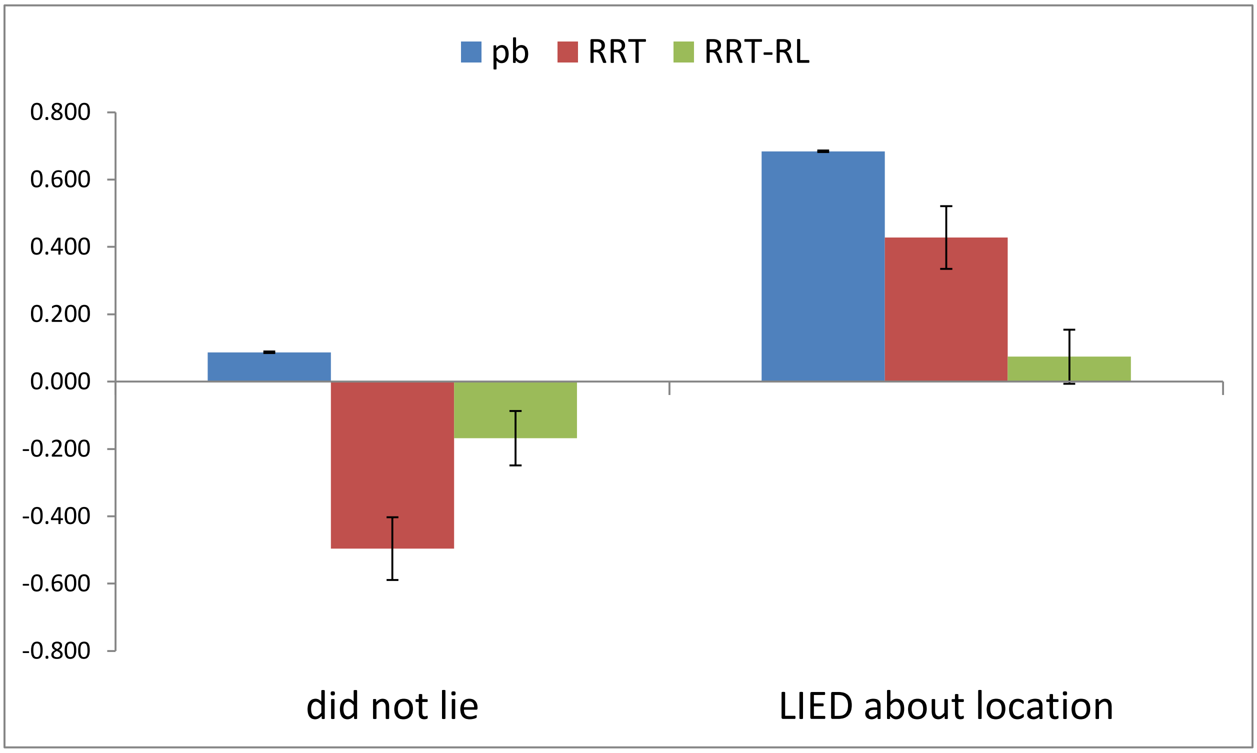


Figure 6B. Prevalence estimates in Experiment 5B, by lying status.



Appendix 1: Instructions from RRT studies showing positive effects:
Zdep Rhodes 1976

The questioning that was finally developed is presented below:

The next question is one which some people find hard to answer. It deals with the use of physical force on children. We also have a question dealing with attendance at PTA meetings (church or synagogue attendance).

I'm going to give you a nickel and a card with these two questions on it. I want you to take this coin and shake it in your hands. [DEMONSTRATE]. Let it rest on the palm of your hand. Don't let me see which side is up. If the heads side turns up, answer the question on the card next to the heads-up coin. If the tails side turns up, answer the question printed next to the tails-up coin. You are to answer "Yes" or "No" without telling me which question you are answering. [HAND RESPONDENT COIN AND EXHIBIT.].

The first question reads, "Have you or your spouse ever intentionally used physical force on any of your children in an effort specifically meant to hurt or cause injury to that child?"

The second question reads, "Have you attended a PTA meeting at school within the past 12 months (attended church or synagogue within the past week)?"

If the respondent hesitated or refused, the interviewer was instructed to offer this further reassurance:

There is absolutely no way we can tell which question you are answering if you don't ell us. On the average, half of the people we interview will answer the "heads" question, and half will answer the 'tails' question. By putting all the answers in our computer we can determine how many people answered "Yes" to each question, but we won't know which ones answered the "heads" question nor will we know which answered the "tails "question. Therefore, it is extremely important that you answer the question indicated by the coin.

Himmelfarb Lickteig 1982

Randomized response technique. Each subject was given an insulated foam cup containing three pennies. They were told that because people are sometimes reluctant to answer questions of a personal nature truthfully even under anonymous conditions, a way had been worked out to obtain the research information yet make certain that the answers could in no way be directly connected with any one individual.

Subjects were then told that before answering each question they were to shake the cup containing the three coins and let the coins fall to the bottom of the cup. If all three coins came up heads, they were not to answer the question but to check the yes position on the answer sheet. If all the coins came up tails, they were not to answer the question but to check the no position on the answer sheet. However, if the coins landed in any combination of heads or tails other than all heads or all tails, they were to answer the question truthfully.

Subjects were then given a practice questionnaire containing three innocuous questions (e.g., Do you own a dog?) and told to shake the cup and answer the questions according to the outcome of the coin toss. They were paced by the experimenter through each practice question. After each question, the experimenter went around the room, asked each subject what the outcome of the coin toss was and how the subject had responded. Each subject was interrogated aloud so that the other subjects could hear the correct procedure and so that they could learn that outcomes other than the one they obtained were possible.

After the practice questions were answered, the experimenter explained to the subjects how the technique maintained their confidentiality and pointed out that their answers could not be directly connected with any one of them if the experimenter did not know the outcome of the coin toss. The experimenter also assured them that valid results still could be obtained through the technique and that the data they provided were worthwhile if they all followed the procedure conscientiously.

From Barth Sandler 1976

“To ensure this anonymity, I have devised the following system: Earlier I passed out 2 dimes to each person in the classroom [subjects were allowed to keep dimes]. I will now ask you to flip each coin separately and remember whether they come up both heads, both tails, or one heads and one tails. If both coins come up heads, please answer question 1: Does your telephone number end in an odd digit? If the coins come up in any other combination (i.e., both tails or one heads and one tails), please answer question 2: Over the past year have you consumed 50 or more glasses (or drinks) of any alcoholic beverages? In marking your answer, please darken the box at the bottom of the page indicating either a ‘yes’ or ‘no’ answer to whichever question you have chosen from the coin flip. Please do not indicate on the questionnaire which question you have answered.

“The reason for the coin flip method is to ensure that I will have no idea which question anyone has answered. Do not write your name on the questionnaire. Please darken in the box at the top of the page which indicates either male or female. Please answer the question you have chosen as accurately and honestly as possible. Are there any questions?”

From van der Heijden at al. 2000

B1. FACE-TO-FACE DIRECT QUESTIONING

B1.1. “We now would like to ask a couple of questions about topics that we already touched upon, for example, your income and possessions, extra high expenses, looking for work, and providing information to the local welfare department. This can have to do with, for example, declaring part of your income from a side job, family reunion, or living together. In short, about information that for all sorts reasons often is not, only partly, or not in time provided to the local welfare department."

B1.2. "We ask you to answer the questions with 'yes' or 'no'".

B1.3. "We understand that this can sometimes be difficult because you will not always have a ready-made answer. That is why we ask you to answer 'yes' when the answer is 'mostly yes' and no when the answer is 'mostly no.'"

B1.4. "We will now ask you a few questions about your expenses and income and about providing information to the local welfare department."

B1.5. [Important. The questions have to be read word by word, including the explanation of the terms, so that the respondent does not need to ask for clarification.]

B1.6. Questions follow about (1) saving for a large expenditure; (2) providing address information to the local welfare department; (3) officially having a car worth more than approximately $15,000; (4) having a motor home; (5) going abroad for holiday longer than four weeks; (6) gambling a large amount (more than $25) at the horses, in casinos, in playing halls, or on bets; (7) having hobbies about which you or household members think cost too much, given the income you have; (8) having refused jobs, or taken care that employers did not want you for a job while you had a good chance to get the job; (9) working more than 20 hours as a volunteer without the local welfare department's knowledge; (10) not declaring part of your income to the local welfare, whereas this is obligatory by law; (11) living now with a partner without the local welfare department's knowledge; (12) having lived with a partner without the local welfare department's knowledge; and (13) giving the local welfare department insufficient or incorrect information about having a fortune. Note that (10) is the dependent variable that is the key variable in this article (see section 2.2.1 for the exact formulation). Also note that questions (1) to (7) are not referring to fraud in any way. They are meant simply to pave the way for more sensitive questions.

B3. RANDOMIZED RESPONSE: FORCED-RESPONSE PROCEDURE

The sensitive block starts with B1.1. Then,

B3.1 "Many people find it difficult to answer these types of questions straightaway because they find the topics too private. Yet, we do not want to embarrass anyone.

Therefore, we ask you these questions, experimentally, in a roundabout way. We let you answer in such a way that your privacy is guaranteed so that nobody can ever find out what you have done personally, including me."

B3.2. "You may answer in a few moments using two dice. With those, you can throw 2 or 12 or something in between. You(r) answer is dependent on what you throw with the dice." [Give the box to the respondent and look at it together.] "In the box you will find a card showing what you have to say when you have thrown the dice." [Let interviewee look and give directions with the next explanation.] "If you throw 5, 6, 7, 8,9, or 10, you always answer 'yes' or 'no' honestly. If you throw 2, 3, or 4, you always answer 'yes.' If you throw 11 or 12, you always answer 'no.' So, if you throw 2, 3, or 4, or 11 or 12, then your answer is based on the outcome of the throw. Because I cannot see what you have thrown, your personal privacy is guaranteed; thus your answer always remains a secret.

"This technique is a bit strange. But it is useful, since it allows people working for Utrecht University to estimate how many people of the group that we interviewed answered 'yes' because they threw 2,3., or 4 and how many people answered 'yes' because they had to give an honest answer.

"Let us take an example. I ask you the question: 'Do you live in Utrecht?' and you throw a 3. You answer with 'yes.'

"We can imagine that you find this a bit awkward, but it does not mean that you are lying or that someone can think that the honest answer to the question is also 'yes'. It means only that you stick to the rules of the game by which your privacy and that of everybody else taking part in this investigation is fully guaranteed. I propose that we now try out a few questions to practice."

B3.3. [Turn around] and B1.5.

"I ask you the first six questions to practice."

Questions follow about whether the respondent (1) read a newspaper today, (2) ignored a red traffic light, (3) received a fine for driving under the influence of alcohol, (4) used public transportation last year without paying at least once, (5) paid the obligatory fee for television and radio, (6) ever bought a bicycle suspecting it was stolen.

The instruction goes on with the following.

"Is it clear now? Then we will now ask the questions we are really interested in. Please take your time to answer them."

[Do not start with the real questions before you are certain the next points are understood. Do not read the following points aloud. Read one of the points aloud only when that point is unclear to the respondent.]

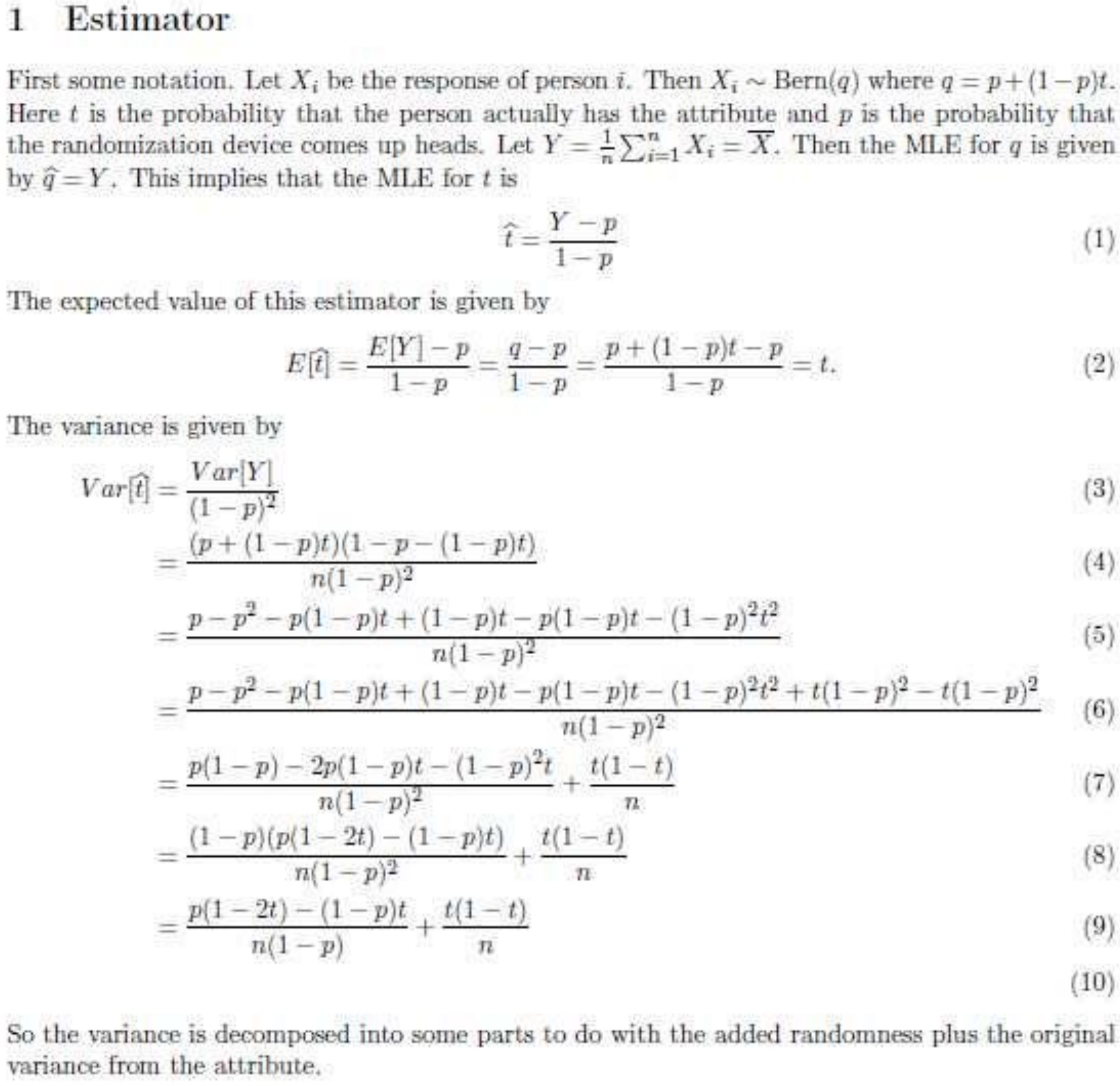
B3.4. [We do this to guarantee your privacy. Nobody sees what you throw and nobody will know what your personal answer is. According to the rules of the game, answers are possible that are in conflict with your feelings: "yes" when it is "no" and "no" when it is "yes". It is not lying: it simply guarantees your privacy. Based on all answers of the people that we interviewed, we can estimate afterward how many people have read a newspaper today or ignored a red traffic light, and so on.] Followed by B.1.4, B1.5, and B1.6.

From the web site:

http://www.randomisedresponse.nl/watisrrENG.htm

“We are about to ask you a few questions about attitudes towards your work, boss and collegues. From previous research we know that many people find it hard to answer this kind of questions, because they are considered too private. Some people fear that an honest answer might have negative consequences. But we do not want to embarrass anyone. That is why we asked Utrecht University to asked these question using a detour that completely guarantees your privacy. You are about to answer the questions with the aid of two dice. With the dice you can throw 2 to 12 and anything between. Your answer depends on the number you threw. This detour completely guarantees your privacy! Nobody, not the company, not the boss and not your collegues can ever know what exactly was your answer.

Appendix 2: RRT Prevalence Estimator



1 Estimator

First some notation. Let X_i be the response of person i. Then $X_i \sim \text{Bern}(q)$ where $q = p + (1-p)t$. Here t is the probability that the person actually has the attribute and p is the probability that the randomization device comes up heads. Let $Y = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$. Then the MLE for q is given by $\widehat{q} = Y$. This implies that the MLE for t is

$$\widehat{t} = \frac{Y-p}{1-p} \tag{1}$$

The expected value of this estimator is given by

$$E[\widehat{t}] = \frac{E[Y]-p}{1-p} = \frac{q-p}{1-p} = \frac{p+(1-p)t-p}{1-p} = t. \tag{2}$$

The variance is given by

$$\begin{aligned}
Var[\widehat{t}] &= \frac{Var[Y]}{(1-p)^2} && (3)\\
&= \frac{(p+(1-p)t)(1-p-(1-p)t)}{n(1-p)^2} && (4)\\
&= \frac{p-p^2-p(1-p)t+(1-p)t-p(1-p)t-(1-p)^2t^2}{n(1-p)^2} && (5)\\
&= \frac{p-p^2-p(1-p)t+(1-p)t-p(1-p)t-(1-p)^2t^2+t(1-p)^2-t(1-p)^2}{n(1-p)^2} && (6)\\
&= \frac{p(1-p)-2p(1-p)t-(1-p)^2t}{n(1-p)^2}+\frac{t(1-t)}{n} && (7)\\
&= \frac{(1-p)(p(1-2t)-(1-p)t)}{n(1-p)^2}+\frac{t(1-t)}{n} && (8)\\
&= \frac{p(1-2t)-(1-p)t}{n(1-p)}+\frac{t(1-t)}{n} && (9)\\
& && (10)
\end{aligned}$$

So the variance is decomposed into some parts to do with the added randomness plus the original variance from the attribute.