When and Why Randomized Response Techniques (Fail to) Elicit the Truth

Leslie K. John
Harvard Business School

George Loewenstein

Alessandro Acquisti

Joachim Vosgerau

Working Paper 16-125

Publish Version Can Be Found
Organizational Behavior and Human Decision Processes
148, 101-123 (2018)
https://doi.org/10.1016/j.obhdp.2018.07.004

When and Why Randomized Response Techniques (Fail to) Elicit the Truth

Leslie K. John

George Loewenstein

Alessandro Acquisti

Joachim Vosgerau

Abstract

By adding random noise to individual responses, randomized response techniques (RRTs) are intended to enhance privacy protection and encourage honest disclosure of sensitive information. Empirical findings on their success in doing so are, however, mixed. Supporting the idea that the noise introduced by RRTs can make respondents concerned that innocuous responses will be interpreted as admissions, seven experiments involving over 3,000 respondents document that RRTs can, paradoxically, yield prevalence estimates that are lower than direct questioning (Studies 1-5), less accurate than direct questioning (Studies 1, 3, & 4B-C), and even impossible (negative prevalence estimates, Studies 3, 4A-C, & 5). The paradox is reduced when the target behavior is framed as socially desirable (Study 2) and is mediated by respondents' concerns over response misinterpretation (Study 3). A simple modification designed to reduce apprehension over response ambiguity reduces the problem (Studies 4A-C), particularly when concerns over response ambiguity are heightened (Study 5).

Keywords: randomized response technique; response bias; survey research; privacy; disclosure

When and Why Randomized Response Techniques (Fail to) Elicit the Truth

> "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)

1. Introduction

Economists, psychologists, sociologists, managers, and policy makers have many reasons for asking personal and even intrusive questions. Sensitive questions of interest concern doping and tobacco, alcohol, and other drug consumption (Brewer 1981; Rohan 2013; Striegel et al. 2010), tax evasion (Houston and Tran 2001), employee theft (Wimbush and Dalton 1997), poaching (St John et al. 2011), regulatory compliance (Ellfers et al. 2003), the integrity of certified public accountants (Buchman and Tracy 1982), or participation or interest in deviant or illegal sexual practices (e.g., Akers et al. 1983, Weissman et al. 1986), to name just a few. 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 more elaborate techniques that guarantee a respondent's privacy?

Previous research on behavioral dimensions of privacy points to the prediction that attempts to reassure individuals about the security of their information can backfire, leading people to clam up rather than feeling freer to honestly share information (e.g., Singer et al. 1992). The overall conclusion of that research is that privacy is not a consideration that people consistently think of; indeed, most people have a deep need to share information, including personal information, with others. Privacy assurances can, in effect, 'ring alarm bells', leading individuals to become protective of their personal information in a way they would not have been had they not been alerted to the idea that their privacy could be violated (John et al. 2011). In one study for example (Joinson et al. 2008), participants were asked to answer an Internet-based survey that included sensitive personal questions. Half of the participants were first asked to complete

a separate questionnaire measuring their Internet privacy concerns. The questionnaire increased the salience of privacy issues, and decreased participants' self-disclosure: relative to the group of participants who had not been asked to discuss their privacy concerns, participants who had been primed with the questionnaire ended up answering fewer personal questions. These studies and others show that people's concern about privacy can be activated by environmental cues that may bear little, or sometimes even a negative, relationship to objective dangers associated with information sharing (for a review, see Acquisti et al. 2015).

Instead of assuring privacy, it has been argued that truthful information sharing can be motivated by introducing stochastic noise to the communication channel (Warner 1965; Forges 1986; Myerson 1986). For example, if messages are lost with probability p, the non-arrival of a message cannot be entirely attributed to an unwillingness to send a message, which increases senders' willingness to share information. Likewise, if respondents are instructed to answer a sensitive question truthfully only with probability p – a procedure called the randomized response technique (RRT) – affirmative responses cannot be interpreted on the individual level, thus increasing respondents' willingness to divulge sensitive information.

Although RRTs take many different forms (Böckenholt and Van der Heijden 2007, Campbell and Joiner 1973, de Jong et al. 2010, Park and Park 1987, Pollock and Bek 1976, Rohan 2013, Scheers, 1992, Tracy and Fox 1981, Warner 1965), they all add noise to individual responses, in theory making it easier for individuals to admit to sensitive behaviors, thoughts, and feelings. For example, in the coin flip technique (Boruch 1971, Dawes and 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 flips a coin and answers 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. The interviewer, who cannot see the outcome of the coin flip, 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 interviewer 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.

Ljungqvist (1993) provided a formalization of the RRT. Utility-maximizing respondents face a tradeoff between lying aversion – they prefer to tell the truth – and stigmatization aversion – they prefer not to be associated with the behavior/information in question. Whether a respondent answers truthfully or not thus depends on both conditional probabilities of being perceived as belonging to the sensitive group (A), p (A|no) – a measure of lying aversion – and p (A|yes) – a measure of stigmatization aversion. Based on Ljungqvist's model, Blume et al. (2013) developed a game-theoretic formulation of the RRT in which respondents face a tradeoff between lying and stigmatization aversion, and respondents' payoffs dependent on the interviewer's beliefs. The model allows for specifying the parameter space for lying and stigmatization aversion for which RRT will induce more truth-telling than direct questioning (DQ), and vice versa, when DQ will induce more truth-telling than RRT.

In line with the latter, in this paper we demonstrate that the RRT can yield lower and less valid prevalence estimates than those obtained by DQ, and in some cases, even impossible (negative) prevalence estimates. These paradoxical effects, however, occur for reasons beyond the specific parameter values for lying and stigmatization aversion. We show that the RRT can perform worse than DQ because the RRT makes respondents concerned that innocuous responses will be interpreted as admissions; because only one response (denial) has an unambiguous interpretation, it leads them to give that response. We present seven studies that demonstrate these paradoxical effects, provide evidence for why they occur, and finally, propose and test a simple modification that reduces the magnitude of the problem.

2. Review of Empirical RRT Findings

The RRT has been, and continues to be, used in a wide variety of applications, all of which hinge on uncovering the prevalence of sensitive attitudes and behaviors. Apart from the examples listed in the introduction, the RRT has been employed in censuses on sensitive topics including drug use, abortion, and anti-semitic attitudes (Brewer 1981, Adler et al. 1992, Krumpal 2012), in turn informing policy decisions

such as anti-drug doping policies in elite sports (Rohan 2013). It has also been used to assess demand for sensitive products and in turn, to prioritize the allocation of (scarce) marketing resources (de Jong et al. 2010) – as one practitioner notes, "the most effective ways to do this are through Randomized Response Techniques" (Insight Central 2010). What is the evidence, however, that RRTs achieve their intended purpose? Two types of studies have been used to assess the effectiveness of RRTs (Umesh and Peterson 1991, Tourangeau and Yan 2007): comparative studies and validation studies.

2.1 Comparative Studies

Comparative studies contrast prevalence estimates obtained using RRTs with those obtained via direct questioning. 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 (the "more-is-better" assumption, see Tourangeau and Yan 2007). In some comparative studies, RRTs have generated higher and hence presumably more valid prevalence estimates relative to DQ (e.g., de Jong et al. 2010, Wimbush and Dalton 1997). However, RRTs have also been found to produce the same or lower and hence presumably less valid prevalence estimates relative to DQ (Begin and Boivin 1980, Beldt et al. 1982, Brewer 1981, Coutts and Jann 2011, Höglinger et al. 2014, Lamb and Stern 1978, Duffy and Waterton 1988, Goode and Heine 1978, Locander et al. 1976, Tamhane 1981, Tracy and Fox 1981, Kulka et al. 1981, Williams and Suen 1994, Wiseman et al. 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 and Heine 1978). Similarly, Weissman et al. (1986) asked respondents whether they had used each of four illicit drugs (cocaine, heroin, PCP, and LSD) and found that drug usage estimates were equivalent across inquiry methods (RRT vs. DQ). Zdep et al. (1979) and Brewer (1981) found RRT estimates of marijuana usage among young adults to be *lower* than DQ estimates. In fact, Brewer (1981) and others (Coutts and Jann 2011, Hoeglinger et al. 2014) have found the RRT to generate prevalence estimates that are *negative* – a 'dead giveaway' of their inaccuracy.

2.2 Validation Studies

In validation studies, estimates are made of the prevalence of behaviors in situations in which the researcher can verify the validity of responses through some external source of data. In some studies, the researcher only knows the true prevalence of the behavior in aggregate (e.g., the percent of registered voters who voted in a given election). In stronger studies, the researcher has data on individual behavior, so responses can be compared at the individual level (e.g., whether a given registered voter voted). The inclusion of a comparison to the prevailing standard method of asking sensitive questions (i.e., DQ) is also informative.

Despite widespread agreement that "individual validation studies are doubtlessly the gold standard" (Lensvelt-Mulders et al. 2005), to the best of our knowledge, only nine validation studies have been published to date: two aggregate-level studies (Horvitz et al. 1967, Rosenfeld et al. 2015), and seven individual-level studies (all of which also include a DQ comparison; Kirchner 2015, Kulka et al. 1981, Lamb and Stern 1978, Locander et al. 1976, van der Heijden et al. 2000, Tracy and Fox 1981, Wolter and Preisendörfer 2013).[1] In the earliest validation study, the RRT generated a prevalence estimate that was 20 times larger than the true prevalence (Horvitz et al. 1967). The RRT also inaccurately measured prevalence in a second validation study, this time underestimating the corresponding population parameter by 35% (Locander et al. 1976). In the third study, Lamb and Stern (1978) compared students' self-reported course failure to actual failure retrieved through academic records. The students were asked one of two questions: *whether* they had ever failed a course or the *number of times* they had failed a course; and they were asked the question as a function of RRT or DQ. Results indicated that DQ performed just as well as the RRT for the first question; however, it performed significantly worse than the RRT for the second question. Similarly, another study, conducted on a criminal population, compared self-reported numbers of arrests

[1] Umesh and Peterson (1991) consider two additional papers to be validation studies: Edgell et al. (1982) and Akers et al. (1983). However, we consider neither to provide validation data. Edgell et al. (1982) covertly recorded whether the randomizer had instructed the given participant to respond 'Yes' versus to answer the truth. Since these researchers did not have information on whether the given participant had actually engaged in the target behavior, we do not consider it to be a validation study. Akers et al. (1983) attempted to assess the validity of smokers' abstinence claims using salivary tests designed to detect thiocynate, a cigarette byproduct, but the test was inconclusive: "given the inexact relationship between smoking and thiocynate levels, it is difficult to get exact numerical estimates of smoking" (Umesh and Peterson 1991, p.124).

with actual arrests records (Tracy and Fox 1981). The RRT produced higher prevalence estimates that were closer to the true rates than DQ for those who had been arrested more than once. However, the opposite – substantially lower and less valid prevalence estimates – was observed for individuals who had been arrested only once. In evaluating the six earliest validation studies, Wolter and Preisendörfer (2013) concluded that "the results in favor of RRT are not convincingly strong."

The two newest individual-level validation studies provide similarly inconclusive results. Wolter and Preisendörfer (2013) asked respondents whether they had been convicted of a crime as a function of either DQ or RRT. Unbeknownst to respondents, only those with a criminal record were included in the sample; hence the true prevalence was 100%. The prevalence estimate gleaned by the RRT (59.6%) was no more valid than that of DQ (57.5%), and both methods vastly underestimated true prevalence. Similarly, Kirchner (2015) found that the RRT was no better than DQ in estimating the propensity to collect welfare, and that both methods underestimated true prevalence.

The latest aggregate-level validation study examined Mississippians' propensity to reveal how they voted on a proposed constitutional amendment to define life as beginning at conception as opposed to birth (Rosenfeld et al. 2015). Respondents were asked whether they had voted "yes" as a function of different questioning techniques, including RRT and DQ. The RRT was hence implemented based on the premise that voting yes – a "pro-life" stance – is taboo. The prevalence estimates gleaned by the RRT were closer to the true prevalence and also higher than that obtained from DQ. These results should be interpreted with caution, however, because there is ambiguity as to which vote – "yes" or "no" – is taboo. Given the state's conservative reputation, and the fact that the amendment passed, it seems plausible that voting "no" rather than voting "yes" constituted a taboo. The RRT may have thus provided protection for the "wrong" answer option. As a consequence, those who were instructed by the random device to vote "no" but were afraid of being stigmatized may have been pushed to ignore the RRT-instructions and vote "yes" instead. In our studies we will provide empirical evidence for such intentional non-adherence of RRT-instructions, which, in Rosenfeld et al. (2015) may have artificially increased prevalence estimates in the RRT. Without *individual-level* validation data, it is unclear to which extent this may have happened, and hence it is

difficult to interpret the results of this study.

2.3 Meta-Analyses

Two comprehensive meta-analyses of RRT studies have been conducted. Umesh and Peterson (1991) – based on the five (or, by their count, seven, see footnote 1) validation studies that had been conducted to date (since that paper, three additional validation studies have emerged, as noted in section 2.2) – concluded that "contrary to common beliefs (and claims), the validity of the randomized response method does not appear to be very good." In contrast, a more recent meta-analysis that included 32 comparative studies (in addition to the same validation studies as Umesh and Peterson (1991)) reported that prevalence estimates obtained using RRTs were on average higher than those obtained using DQ (Lensvelt-Mulders et al. 2005). The authors concluded that "using randomized response questions in surveys on sensitive topics significantly improves the data quality of the surveys" (p. 25). Finally, a meta-analysis of DQ on sensitive topics such as income, drug use, and health – based on validation studies – found direct questioning to yield surprisingly accurate results (Marquis et al. 1986): no systematic underreporting of population values was found.

Together, the results of these three meta-analyses are far from conclusive. Whether the RRT is an effective tool for eliciting honest answers to sensitive questions is still an open question. As one group of researchers concluded: "despite the aforementioned meta-analyses [the two reported in Lensvelt-Mulders et al. 2005] and a huge amount of literature on the subject, it is still controversial whether RRT provides any benefit to response validity at all" (Wolter and Preisendörfer 2013, p. 323). It is disconcerting that the RRT continues to be used in the face of its many unexplained failures; in fact, the *Journal of the American Statistical Association* recently published a step-by-step guide on how to analyze RRT data (Blair et al. 2015).

3. Non-adherence

Clearly, respondents often fail to adhere with the RRT instructions, which can produce paradoxical

results – prevalence estimates lower than DQ estimates or even impossible (negative or in excess of 100%). Edgell et al. (1982), in an attempt to quantify the seriousness of the non-adherence problem, surreptitiously recorded the outcome of the randomizer, and found that 25% of respondents answered 'no' when the randomizer had instructed them to answer 'yes.'

3.1 Correcting for non-adherence

Sophisticated statistical methods have been introduced to measure non-adherence in order to correct for it post-hoc. Clark and Desharnais (1998) measure non-adherence by randomizing respondents to one of two probabilities of forced 'yes' responses. For example, the surveyors set the random device to instruct one group of respondents to answer truthfully in 75% of cases and to answer 'yes' in 25% of cases; and the other half of respondents to answer truthfully in 25% of cases and to answer 'yes' in 75% of cases. By comparing the prevalence of 'yes' responses of both groups, the prevalence of the behavior in question can be estimated, as well as the rate of non-adherence to RRT instructions. Ostapczuk et al. (2009) applied this methodology, estimating a non-adherence rate of 20.2% among Chinese students asked about cheating in exams. And, using the same method in a study with a German patient sample asked about compliance with doctor-prescribed medicine intake, Ostapczuk et al. (2011) estimated a 38.9–55.2% non-adherence rate.

The second approach of estimating non-adherence rates for post-hoc correction of RRT estimates uses latent class models (Böckenholt and van der Heijden 2007, Cruyff et al. 2008, Cruyff et al. 2007). Each construct is measured with multiple items and the RRT applied to each item. This allows for estimating each respondent's probability of belonging to a latent class of non-adherents, as well as his probability of belonging to a latent class of having exhibited the behavior in question. Using the multi-item latent class approach, Böckenholt and van der Heijden (2007) estimated rates of non-adherence of 13-17% in a study on fraudulent health insurance benefit claims. In two studies on the prevalence of sexual attitudes and behaviors, de Jong et al. (2010) estimated a non-adherence rate of 10.3% in a Dutch sample, and de Jong et al. (2012) estimated rates ranging from 5% (Brazil, Japan) to 20% (Netherlands) to 29% (India).

Although ingenious, these post-hoc correction methods have some important limitations. First, from a practical standpoint, as illustrated by the above applications, non-adherence rates vary greatly between populations and topics of inquiry. Thus these post-hoc correction methods require non-adherence to be measured anew each time the RRT procedure is administered. Doing so requires randomizing respondents between two different probabilities of answering the sensitive question, effectively doubling the required size of an already necessarily large sample size relative to simple direct questioning. These methods can also be cumbersome to administer; for example, using the latter method, the RRT has to be applied to each item – and each construct (i.e., sensitive topic) has to be assessed with multiple items. Perhaps more importantly, it seems that neither correction procedure has been empirically validated: to the best of our knowledge, non-adherence estimates gleaned by these correction methods have not been compared to true non-adherence rates. Such information would be important in ascertaining whether these correction methods work.

This lack of evidence is problematic for Clark and Desharnais' procedure, because it assumes that non-adherence is independent from the probability of being forced to answer 'yes.' It is reasonable to question the validity of this fundamental assumption because the probability that respondents are forced to answer 'yes' versus permitted to answer truthfully is typically transparent and known to respondents. A respondent might therefore reasonably feel more conspicuous answering 'yes' when p(forced yes) is 25% relative to when it is 75%. Hence, respondents may be less likely to adhere to the randomizer (i.e., to answer "yes" when instructed to do so by the randomizer) in the former situation than in the latter. In other words, non-adherence could conceivably be negatively correlated with p(forced yes), violating an assumption of the correction procedure.

Similarly, the lack of validation data is noteworthy for studies that use latent-class post-hoc corrections because the method – perhaps implausibly – assumes that adherence is the same across items and across the outcome of the randomizer. In other words, this correction method assumes that non-adherence is just as likely for benign questions as it is for sensitive questions (an assumption refuted by Wolter and Preisendörfer 2013), and similarly, that it is just as likely when instructed by the randomizer to

answer ‘yes’ versus to answer based on engagement in the behavior (an assumption refuted in our Study 3). Violations of this so-called “local independence” assumption (errors of individual items are uncorrelated) can lead to biased prevalence estimates (Kreuter et al. 2008, Yan et al. 2012). Finally, non-adherence is estimated by counting the number of times participants give non self-incriminating responses. The method thus cannot distinguish between true non-adherence and true non-engagement in the behaviors, which could make it prone to overestimating non-adherence.

In sum, although previous researchers have proposed means of addressing non-adherence that are ingenious from a statistical perspective, there is reason to question their practicality as well as their ability to produce valid prevalence estimates. But even despite these concerns, these methods are designed to correct for the problem after-the-fact; it would be preferable, if possible, to use an approach that avoided the problem altogether. In order to do so, it is important to understand the reasons for non-adherence.

3.2 Explaining Non-adherence

Why do respondents fail to comply with the RRT instructions? In this paper, we provide empirical evidence to answer this question, drawing on the psychology of non-adherence. We then use this understanding to devise and test a method of preventing intentional non-adherence, in turn generating more valid prevalence estimates.

Previous researchers have proposed three different explanations for non-response. We describe them below, reviewing their merit and the solutions proposed to address them. First, non-adherence could be unintentional: 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 fraudulent health insurance benefit claims. Clarity of instructions and participants’ education level were negatively related to the number of ‘no’ answers when a ‘yes’ answer was required. Although such unintentional non-adherence contributes to prevalence estimate distortion, our studies suggest that it is not the full explanation (indeed, as study 4B will show, paradoxical effects of RRTs

are observed even when participants must pass a quiz about the procedure prior to answering the focal question).

The second and third proposed reasons entail *intentional* non-adherence. Some have argued that intentional non-adherence, in the form of boasting, may distort prevalence estimates. In Zdep et al.'s (1979) study which found higher marijuana usage when adults between the ages of 18 and 25 were asked directly rather than through the RRT, the authors argued that marijuana usage was fashionable among this age group, leading to greater boasting under DQ than the RRT. Similarly, Lensvelt-Mulders et al. (2005) found RRT prevalence estimates to be lower than DQ estimates for items on which boasting was plausible (e.g., "How often do you donate blood?"). In these cases, the lower estimate (RRT) was taken to be the more valid estimate, contrary to the criterion applied to the other studies in this meta-analysis, where the method that produced the higher estimate was presumed to be more valid.[2] 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.

The third and perhaps most interesting explanation for non-adherence was proposed, though not tested, by Campbell (1987): Although participants may understand the instructions correctly, they are uncomfortable responding 'yes' when the technique calls for them to do so because it introduces the possible interpretation that they engaged in a behavior in which they truly did not engage, or do not want to be perceived as having engaged in. 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 (2007) called this form of non-adherence to RRT instructions "self-protective behavior." Broadly consistent with this explanation, Wolter and Preisendörfer (2013) found that respondents were particularly likely to disobey the randomizer for socially undesirable behaviors; however, they went on to conclude that "although many

[2] Since the authors did not identify the 32 comparative studies used in their meta-analysis it is not clear to which datasets such recoding was applied.

claims can be found in the literature (…), the exact reasons behind misreporting are still controversial." (Wolter and Preisendörfer 2013, p. 329). To the best of our knowledge, the current paper is the first to empirically test the explanation that RRTs backfire because participants are concerned about response misconstrual. In turn, we use this understanding to devise and test a method of preventing non-adherence.

4. Overview of Studies

In seven studies, we investigate whether and why non-adherence occurs in the RRT, and test a simple modification of the RRT to prevent this behavior. We consistently document paradoxical effects of RRTs – estimates that are lower than DQ (Studies 1-5), less valid than DQ (Studies 1, 3 & 4B-C), and even impossible (negative prevalence estimates, Studies 3 & 4A-C). We begin with a simple demonstration of the paradox using a validation study (Study 1). Then, in four subsequent experiments, we provide evidence that the paradox occurs in part because the noise introduced by RRTs makes respondents concerned that innocuous responses will be interpreted as admissions. Consistent with this explanation, the paradox is reduced by a manipulation that alleviates apprehension over response ambiguity – framing the target behavior as socially desirable (Study 2). Study 3 goes further, by showing that the propensity to respond affirmatively is mediated by respondents' concerns that their answers will be misinterpreted. We show that a simple modification to the RRT designed to reduce apprehension over response ambiguity reduces the problem (Studies 4A-C), particularly in situations in which concerns over response ambiguity are heightened (Study 5).

All experiments employ the coin flip method because it is one of the most commonly used RRT and is simple to administer. Critically however, our findings apply to RRTs in general, because as our seven experiments document, it is the ambiguity introduced to responses – a universal feature of, and in fact, the *defining* feature of the method – that is responsible for the paradox.

All studies include a DQ condition as a benchmark to compare the prevalence estimates generated by the RRT. Studies 1, 3, 4B, and 4C are individual-level validation studies, enabling prevalence estimates

to also be compared to 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 practitioners and scholars who are interested in obtaining valid sensitive information from questionnaires and surveys. For all studies, we explain how we predetermined sample sizes and report all manipulations and measures. No data were excluded from the analyses unless explicitly indicated.

4.1 Study 1

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

4.1.1 Method

Emails were sent to all individuals 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 cheating studies we had followed a procedure similar to that introduced by Mazar et al. (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 to use the information 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%; $\chi^2(1)=23.62$, $p<.0001$), and 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 email had been sent, at which point it had been about seven days since the last response.

Out of all participants from the first study, 198 responded to the survey (51.5% male, M_{age}=23.8 years, *SD*=6.4 years; all *NS* between-conditions), a 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, in all experiments, we oversampled RRT relative to DQ.[4]

All participants were instructed:

[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.

[4] In the studies conducted first (Experiments 1, 2, and 4A), the ratio was 2:1. Given that these studies produced RRT standard errors that were still much larger than DQ, in subsequent studies, we increased this ratio to 3:1 (in Experiment 4B) and finally, to 4:1 in Experiment 3 – the final experiment we conducted.

> 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 prevalence estimate 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 also analyzed the prevalence estimates in the direct questioning versus RRT with likelihood ratio tests. The results are virtually the same, with results slightly stronger for likelihood ratio tests.

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., when 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.

4.1.2 Results and Discussion

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: prevalence estimate=2.6%[5] vs. true prevalence=54.9%; $p<.0001$; DQ: prevalence estimate=24.3% vs. true prevalence=61.3%; $p<.005$). Before we analyze whether self-reported cheating (in DQ and RRT) is a function of whether participants actually cheated or not (Study 4 C), in Studies 2 and 3 we test whether lower RRT prevalence estimates are caused by self-protective behavior.

4.2 Study 2

Study 1 provides evidence that RRTs can generate lower and less valid prevalence estimates relative to DQ. In Study 2, we test whether this effect arises from self-protective behavior: respondents who flip heads may fail to check 'yes' due to concern that their response will be misinterpreted as an affirmative admission. If lower RRT estimates are driven by this concern over response ambiguity, we would expect the paradox to disappear when social desirability is manipulated. When it is socially desirable to 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). Study 2 tests this idea. The hypothesis is that when a behavior is framed as socially undesirable, prevalence estimates 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). In other words, if concerns over response ambiguity are a cause of the paradox, the RRT should be more responsive to manipulations of concerns over response ambiguity relative to DQ.

4.2.1 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; 62.6% female;

[5] 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).

M_{age}= 43.62, SD=12.35; all *NS* between-conditions) were recruited as they walked by and 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 Studies 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.

4.2.2 Results and Discussion

A logistic regression revealed a significant main effect of inquiry method (β_{RRT}=-1.26, p=.017) and, of more relevance to our hypothesis, an interaction between inquiry method and social desirability (β_{RRT}*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).

Study 2 supports the idea 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. Study 2 also helps to rule out the possibility that the results of Study 1 are simply the result of some kind of trivial methodological mistake, and/or participants' misunderstanding of the procedure. 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.

4.3 Study 3

Studies 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. Study 3 is a validation study that provides additional evidence that the RRT underperforms relative to DQ. More importantly, Study 3 provides direct evidence of the process underlying the propensity to respond affirmatively, by measuring respondents' concern over response ambiguity. Study 3 was a three condition between-subjects design in which we knew the outcome of the randomizer and hence, could look separately at RRT participants forced to answer 'yes' versus those instructed to respond truthfully. We predicted that participants in the RRT-

forced-yes condition would report greater concern over having their response to the sensitive question misinterpreted relative to those in the RRT-answer-truthfully and DQ conditions, and that this concern would mediate the propensity to respond affirmatively.

4.3.1 Method

Participants (*N*=650) were recruited through Mturk in exchange for a small fixed payment. In Studies 3 and 4B, which were conducted considerably later than the other experiments, we sought to have at least 250 participants in each RRT condition. In Study 3, this implied a total target sample size of approximately 625 participants (since RRT was oversampled 4:1 relative to DQ). We administered the survey through Mturk because it enabled us to: a) collect lots of data quickly and inexpensively; b) pose sensitive questions; and c) collect validation data (IP addresses, as described below). The target sample size was much higher in Studies 3 and 4B relative to the studies we had conducted earlier because of exposure to thought-leaders who now suggest that even n's of 50 are not large (Simmons et al. 2013).

At the beginning of the survey, participants were asked: "Are you completing this survey from a location within the United States?" Later, they would be asked (as a function of DQ or RRT) whether they had lied on this question. For efficiency of data collection, we advertised the study only to prospective participants who were *not* from the United States. Participants were not aware of this filter. To provide participants with a motive to lie, 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)." Unbeknownst to participants, we collected their IP addresses to

validate that they were in fact not in the United States.[6] 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. [7]

Manipulation. Next, participants were asked whether they had lied about their physical location. In the DQ condition, they were simply asked: "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)." In the RRT condition, this question was preceded by a page describing the RRT procedure and providing instructions on how to answer the question. Although the probability of being forced to answer 'yes' (50%) was the same as in Studies 1 & 2, we used a different randomizer that, when combined with the responses to the demographic questions from the end of the survey, enabled us to look separately at RRT participants who had been randomized to answer 'yes' versus those randomized to respond truthfully. Specifically, half of RRT participants were instructed to:

> "answer the question depending on whether you were born in the first half of the year (i.e. January - June) or in the second half of the year (i.e. July-December). Specifically:
>
> If you were born in the FIRST HALF of the year (i.e. January - June)....
>
> - we will ask you to respond "Yes" to the question, REGARDLESS of whether you've done the behavior.
>
> If you were born in the SECOND HALF of the year (i.e. July - December)...
>
> - we will ask you to answer the question honestly (i.e. to indicate whether you've actually done the behavior).
>
> Once you have read and understood the above instructions, click the >> button to proceed."

For the other half of RRT participants, the mapping of birth month to response type was reversed

[6] We validated the IP addresses using this coding tool: http://software77.net/geo-ip/multi-lookup/
[7] In a previous study, we found that answering the math questions did not interact with the subsequent inquiry method condition.

– participants born in early months were instructed to answer truthfully; those born in later months were instructed to answer 'Yes.' This control manipulation did not affect outcomes and therefore we collapse across it in describing the results.

Process measure. On the next page, participants indicated the extent to which they agreed with the statement: "When answering the question on the previous page, I was concerned that my answer would be misinterpreted" on a scale from 1 (not at all concerned) to 7 (very concerned).

The survey ended with basic demographic questions which, critical to breaking the RRT condition into its components (i.e. forced-yes vs. answer-truthfully), included an assessment of birth month. Although the propensity to answer this question was statistically significantly different between conditions (percent of respondents providing their birth month: DQ=98.5%; RRT-answer-truthfully=92.1%; RRT-forced-yes=100%; *Fisher's Exact*=27.08, $p<.0005$), the difference is small in magnitude – across *all* conditions, the vast majority (i.e. 96.5%) of participants provided this information. There were no differences in reported birth months by condition and birth months were approximately uniformly distributed (as they should be: http://www.panix.com/~murphy/bday.html). Note that lying about one's birth month would have only made it more difficult to detect differences between conditions.

4.3.2 Results

Twenty-three percent of participants lied about their physical location (i.e., said that they were completing the survey from the United States when in fact they were not; *NS* between conditions). Consistent with Studies 1 and 2, the RRT produced a lower and less valid prevalence estimate relative to DQ (RRT=-23.1%; DQ=19.7%; $t(648)=4.69$, $p<.0005$). DQ accurately measured prevalence – it produced a prevalence estimate (19.7%) that was not significantly different from the true prevalence (23%, $p=.37$) – while the RRT did not ($p<.0005$).

Consistent with the response ambiguity explanation, there were significant differences in concern

over being misinterpreted as a function of inquiry condition ($F(2, 636)=11.23$, $p<.0005$).[8] Specifically, participants in the RRT-forced-yes condition were significantly more concerned over being misinterpreted relative to those in the RRT-answer-truthfully condition ($M_{\text{RRT-forced-yes}}=4.7$, SD=2.17; $M_{\text{RRT-answer-truthfully}}=3.9$, $SD=2.23$, $t(505)=4.37$, $p<.0005$) and the DQ condition ($M_{\text{DQ}}=3.8$, $SD=2.39$; $t(245.3)=3.51$, $p=.001$).

Mediation. A mediation analysis revealed that the relationship between inquiry method and the propensity to respond affirmatively ($\beta_{\text{inquiry}}=1.32$, $SE=.14$, $p<.0005$) was significantly reduced when concern of misinterpretation was included in the model ($\beta_{\text{inquiry}}=1.26$, $SE=.14$, $p<.0005$; $\beta_{\text{concern}}=.18$, $SE=.04$, $p<.0005$), providing support for partial mediation (*Sobel test* = 2.95, $p=.003$). This pattern holds when controlling for the propensity to lie about one's location (*Sobel test* = 2.57, $SE=.03$, $p=.001$).

4.4 Studies 4A, B & C

Taken together, Studies 1-3 suggest that RRTs can backfire because they make respondents concerned that innocuous responses will be interpreted as incriminating. Studies 4A-C test a possible antidote to the problem: a subtle revision to the RRT response labels that communicates the surveyors' understanding that individual 'yes' responses do not necessarily connote admissions. The revision was inspired by Edgell et al.'s (1982) anecdotal observation 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." The revised response label tested in Studies 4A-C is designed to satisfy this apparent urge, thereby mitigating the paradox.

In addition, in Study 4A, we asked participants an extremely sensitive question (as ascertained by a pilot study). The RRT is believed to display its greatest advantage over DQ for highly sensitive questions (Lensvelt-Mulders et al. 2005, Warner 1965); thus Study 4A is a conservative test of the basic hypothesis that the RRT can backfire.

[8] Note that the denominator degrees of freedom are 636, implying a sample size of N=639 instead of 650 as reported in the methods section of Experiment 3. This inconsistency exists because 11 participants did not answer the process measure.

4.4.1 Study 4A

4.4.1.1 Method

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. 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.

4.4.1.2 Results and Discussion

As predicted, prevalence estimates were significantly lower in RRT-SL relative to DQ (RRT-SL=-21.0%, DQ=25.4%; $t(100)=2.47$, $p=.015$), 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. Note however that the revised RRT (RRT-RL) did not yield appreciably higher prevalence estimates: it simply did not produce a paradoxical effect.

4.4.2 Study 4B

In Study 4A, a subtle revision to the RRT response label designed to address concerns over response ambiguity eliminated the paradox: prevalence estimates in the RRT-RL condition were indistinguishable from those in the DQ condition. Although we propose that this effect occurs because the revised label reduces concerns over response ambiguity, one could argue that the unusual-looking revised label prompts respondents to read, and hence comply with, the RRT instructions. Study 4B rules out this alternative

explanation by showing that the revised label outperforms the standard label even when we can be sure that participants understand the RRT procedure, and hence, how they are supposed to answer (all participants must pass a quiz about the procedure before answering the target question).

4.4.2.1 Method

Participants (*N*=1,135) were recruited through Mturk in exchange for a small fixed payment. We planned to close data collection after approximately two weeks, provided that a sample size of at least 250 per RRT condition was reached.

Participants were first asked to indicate whether they were completing the survey from a location within the United States. Participants then completed a brief filler task. Next, participants were told that they would be asked to answer a follow-up question. In the RRT conditions, the RRT instructions were then provided (as in Study 4A).

Participants in the comprehension conditions were further told that they would answer two quiz questions "to make sure you understand the questioning procedure that was described on the previous page." In the first quiz question, participants were presented with a hypothetical scenario in which a person named Lucy was asked: "Have you cheated on your tax return?" using the RRT. Participants were told that Lucy had never cheated on her taxes and had flipped heads; they were then asked "Given this information, what is the correct response that Lucy should give?" (Lucy should respond… "Yes" / "No"). In the second quiz question, participants were told that Lucy had not cheated on her taxes, and that she had flipped tails. Participants could not proceed with the survey until they had correctly answered both quiz questions (upon entering an incorrect response, participants were looped back to the RRT instruction page).

Finally, participants were asked whether they had lied about their location (same item as Study 3), as a function of either: DQ, RRT-SL, or RRT-RL. Again, as in Study 3, we collected their IP addresses to validate their location claims. We did so by running the IP addresses through an automated IP address coding tool.

4.4.2.2 Results

Instruction Quiz. Most (72.9% of) participants in the instruction quiz conditions answered the first

quiz question correctly on the first attempt (*NS* between conditions). Almost all (96.9% of) participants proceeded to answer the second question correctly on the first attempt (*NS* between conditions). The primary results below are intent-to-treat; all participants are included in the analysis regardless of their performance on the instruction quiz.

Prevalence estimates. Ten percent of participants lied (*NS* between conditions). Figure 4 presents mean estimated prevalence rates in all five conditions. Although all prevalence estimates were significantly different from the true prevalence of lying (10.0%), collapsing across the quiz manipulation, prevalence estimates were significantly higher in RRT-RL compared to RRT-SL, suggesting that the revised label was partially effective in reducing response ambiguity apprehension (RRT-SL=-27.4%, RRT-RL=-8.8%; $t(1045)=2.89$, $p<.01$).

More importantly, there was no difference in prevalence estimates as a function of the quiz in either the RRT-SL conditions (SL-NoQuiz=-23.4%, SL-Quiz=-31.1%; *NS*) or the RRT-RL conditions (RL-NoQuiz=-9.6%, RL-Quiz=-8.0%; *NS*), suggesting that the revised label facilitates disclosure not because it simply cues participants to read the RRT instructions, but because it reduces response ambiguity apprehension.

4.4.3 Study 4C

An interesting question is whether the magnitude of the paradox, and the benefit of the revised label, depends on whether the respondent has engaged in the target behavior. One might expect the paradox to emerge among respondents who have not engaged in the behavior. In contrast to direct questioning, with the RRT, innocent people who flip heads are deprived of the opportunity to unambiguously express their innocence. Wanting to do so, these individuals may be tempted to respond 'No' despite the randomizer's instruction to respond affirmatively.

By similar logic, one might also expect the paradox to emerge among those who *have* engaged in the behavior. To see why, consider that at least some guilty individuals will want to lead others to think that they are innocent. In direct questioning, such individuals can unambiguously feign innocence (by lying, explicitly indicating that they have not done the behavior). With the RRT however, some guilty people will

be deprived of the opportunity to unambiguously feign innocence. Wanting to do so, these individuals may be tempted to deny having engaged in the behavior despite the randomizer's instruction to respond affirmatively.

In sum, both innocent and guilty individuals who are posed sensitive questions using RRTs have reason to be concerned over response ambiguity and thus, the RRT may backfire for both groups. In Study 4C, we explore this possibility by pooling data from all nine validation studies that we have conducted to date – three of which are included in this paper (Studies 1, 3, and 4B). The validation data enabled us to examine prevalence estimates by engagement status (i.e. whether the respondent had engaged in the target behavior).

4.4.3.1 Method

In all nine validation studies that we have conducted, participants (N=4,144) were asked whether they had engaged in a sensitive behavior (either lying about one's physical location, 7 studies; or cheating on a previous task, 2 studies). Method of inquiry (DQ vs. RRT) was randomized between-subjects (as in the previous studies, we used the coin flip method of the RRT). In addition, seven of the studies also included a revised label version of the RRT. Therefore, we were able to see whether the effectiveness of these inquiry techniques (DQ, RRT-SL, RRT-RL) differed by whether participants had engaged in the given behavior.

4.4.3.2 Results

We replicated the basic paradox: regardless of engagement status, prevalence estimates were highest in DQ (18.5%) and lowest in RRT-SL (-21.4%); prevalence estimates in the RRT-RL fell in the between these two estimates (-.03%). All pairwise comparisons were strongly statistically significantly different from each other (all *p*s <.0001). Both the estimates generated by the standard and revised labels were significantly lower than the true prevalence (18.8%, *p*s <.0001). DQ however, provided an accurate prevalence estimate – the estimate it provided (18.5%) was not statistically significantly different from the true prevalence (18.8%; z=0.193, *p*=.85; 95% CI of the DQ prevalence estimate = 15.9% - 21.5%).

More interestingly, as can be readily seen in Figure 5, the RRT-SL backfired for both people who

had and who had not engaged in the behavior. Similarly, for both groups, the RRT-RL facilitated affirmative responding, and hence, reduced the paradox.[9]

4.5 Study 5

Studies 2-4 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 when instructed to do so by the randomizer. Studies 4A, B & C show that addressing this concern increases the validity of the prevalence estimates; although the modified version of the RRT did not outperform DQ. The latter point 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, and in turn, when the advantage of the revised label (relative to the standard label) is expected to be particularly great.

Study 5 tests the following idea: the paradox should be heightened by a manipulation that increases apprehension over response ambiguity. In turn, the relative effectiveness of the revised label should be particularly great in this context. The study 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 the RRT-SL would be particularly likely to backfire (relative to DQ) among participants who were relatively identifiable. We also predicted that relative to the RRT-SL, the revised label would be particularly

[9]We do not have validation data for 60 participants (4% of the sample of 4,144 participants). Seven of these participants were part of a study in which IP addresses were the source of validation data. These participants had blocked their IP addresses and hence their physical location could not be ascertained. The remaining 53 of these 60 participants were part of a study in which cheating was the source of validation data, as in Study 1. We were unable to link these participants' data from the initial cheating study to their responses on the follow-up survey in which we asked them (as a function of DQ or RRT) whether they had cheated. Data linkage was done on the basis of participants' first name and first initial of their last name; for these participants, we could not find an exact match between the name information they provided in the follow-up survey and the cheating study data set (which was the source of the validation data).

effective for these participants.

4.5.1 Method

Participants (*N*=691) were recruited through Mturk in exchange for a small payment. 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.

4.5.2 Results and Discussion

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 Study 4A, prevalence estimates 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*).

Moreover, the benefit of the revised label over the standard label was driven by participants in the identified condition (Figure 6).[10] Among participants in the identified condition, the prevalence estimate in RRT-SL (-35.3%) was dramatically lower relative to both 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, although directionally consistent with our other studies (DQ=11.8%; RRT-SL=0.7%; RRT-RL=9.4%; all comparisons *NS*).

The pattern of results is even more pronounced when the identified condition is restricted to the 82.8% of participants who provided identifying information (prevalence estimate among Ss who provided

[10] In Experiments 4 and 5 due to negative prevalence estimates, we were unable to conduct a logistic regression to explicitly test for an interaction as we had in Experiment 2.

identifying information: DQ=8.6%; RRT-SL=-19.3%; RRT-RL=19.7%).

5. General Discussion

RRTs are intended to make individuals more comfortable admitting to having engaged in sensitive behaviors. And yet RRTs can produce prevalence estimates that are lower than DQ estimates (Studies 1-4), less valid than DQ estimates (Studies 1, 3, and 4B), or even impossible (i.e., negative, Studies 3 & 4). Study 2 shows that the paradox is alleviated by a manipulation that reduces apprehension over response ambiguity – framing the target behavior as socially desirable. Study 3 provides direct evidence for this explanation, by showing that concern of misinterpretation mediates the relationship between inquiry method and the propensity to respond affirmatively. Studies 4A-C show that a simple modification to the RRT designed to reduce apprehension over response ambiguity reduces the problem. Finally, Study 5 shows that the relative advantage of the revised label in alleviating the paradox is greatest in situations in which concerns of response ambiguity are heightened (Study 5).

In the studies in this manuscript, RRT prevalence estimates were always lower than actual prevalence estimates, and were only, at best, equal to DQ estimates. The results of our studies are thus consistent with Umesh and Peterson's (1991) conclusion that "contrary to common beliefs (and claims), the validity of the RRM [RRT] does not appear to be very good." Our results, however, stand in stark contrast to the conclusion of the 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)." Given that our, and previous studies (Brewer 1981, Holbrook and Krosnick 2010) yielded impossible prevalence estimates (i.e., negative or in excess of 100%), RRTs may perform much worse than is concluded by Lensvelt-Mulders et al. (2005).

Our results demonstrate that revising the response labels in the RRT reduces self-protective behavior. By labeling 'yes' responses as 'yes/flipped heads,' participants seem to be less afraid of self-incrimination, and are more likely to follow the RRT instructions to provide an affirmative response when required to do so by the random device. Hence, to glean the most accurate prevalence estimates of sensitive

behaviors such as those relevant to public health or environmental issues, tobacco, alcohol, and other drug consumption, gambling, and financial behavior, our approach, which *prevents* self-protective responding, could be combined with latent class approaches which *correct for* any remaining response bias after-the-fact. (That is, assuming that the latent class approach is empirically tested against proper control conditions and found to be effective). This appears to be a promising avenue for future 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. While people in general seem to be motivated to tell the truth (Gneezy 2005, Sànchez-Pagés and Vorsatz 2007), they may not be so in some circumstances, and then there is no reason to think that the RRT will have the intended effect (see for example, Blume et al. 2013). 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 consistently paradoxical effects of the RRT than did prior research 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, office workers, and internet recruits in our studies.

The present research demonstrates that providing people with protections that, logically, should make them more forthcoming with information, can in fact backfire, leading people, instead, to 'clam up'. Prior research has, analogously, shown that confidentiality assurances that, logically, should ease people's fears of sharing information in fact make people *less* willing to respond to surveys on sensitive subjects (Singer et al. 1992). Researchers have also demonstrated the obverse effect – factors that should make people less forthcoming with information can, paradoxically, increase divulgence. 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 (John et al. 2011). Taken together, these findings highlight how people's willingness to divulge can be at odds with the objective consequences of information revelation.

The tactic of Robert Penn Warren's protagonist in *All the King's Men* is to ask sensitive questions directly, "for if you ask something quick and sharp out of a clear sky you may get an answer you never would get otherwise" (Warren 1946); we concur.

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 tell 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 [*sic*]. From previous research we know that many people find it hard to answer this [*sic*] kind [*sic*] 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 [*sic*] 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 [*sic*] 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.

References

Acquisti A, Brandimarte L, Loewenstein G (2015) Privacy and human behavior in the age of information. *Science* 374(6221):509-514.

Adler NE, David HP, Major BN, Roth SH, Russo NF, Wyatt GE (1992) Psychological factors in abortion: A review. *American Psychologist* 47(10):1194-1204.

Akers RJ, Massey J, Clarke W, Lauer RM (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 SF, Daniel WW, Garcha BS (1982) The Takahasi-Sakasegawa randomized response technique. *Sociological Methods and Research* 11:101-111.

Blair G, Imai K, Zhou Y (2015) Design and analysis of the randomized response technique. *Journal of the American Statistical Association* 110(511):1304-1319.

Blume A, Lai EK, Lim W (2013) Eliciting private information with noise: The case of randomized response, Working Paper.

Böckenholt U, Van der Heijden PGM (2007) Item randomized-response models for measuring noncompliance: Risk-return perceptions, social influences, and self-protective responses. *Psychometrika* 72(2):245-262.

Boruch RF (1971). Assuring confidentiality of responses in social research: a note on strategies. *The American Sociologist* 6(4):308-311.

Brewer KRW (1981) Estimating marihuana usage using randomized response: Some paradoxical findings. *Australian Journal of Statistics* 23:139-48.

Buchman TA, Tracy JA (1982) Obtaining responses to sensitive questions: Conventional questionnaire versus randomized response technique. *Journal of Accounting Research* 20:263-271.

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

Campbell AA (1987) Randomized response technique. *Science* 236(4805):1049.

Clark SJ, Desharnais RA (1998) Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychological Methods* 3(2):160-168.

Coutts E, Jann B (2011) Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). *Sociological Methods & Research* 40:169-93.

Cruyff MJ, Böckenholt U, van den Hout A, van der Heijden PGM (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 MJ, van den Hout A, van der Heijden PGM, 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 MG, 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 MG, 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 JC, Waterton JL (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 SE, Himmelfarb S, Duchen KL (1982). Validity of forced response in a randomized response model. *Sociological Methods and Research* 11:89-110.

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

Forges F (1986) An approach to communication equilibria. *Econometrica* 54:1375-1385.

Gneezy U (2005) Deception: The role of consequences. *American Economic Review* 95:384-394.

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.

Höglinger M, Jann B, Diekmann A (2014) Sensitive questions in online surveys: An experimental evaluation of the randomized response technique and the crosswise model. *University of Bern Social Sciences Working Paper No 9*.

Holbrook AL, Krosnick JA (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.

Horvitz DG, Shah BV, Simmons WR (1967) The unrelated question randomized response model. *Proceedings of the Social Statistics Section* American Statistical Association 65-72.

Houston J, Tran A (2001) A survey of tax evasion using the randomized response technique. *Advances in Taxation* 13:69-94.

Insight Central (2010) Randomized responses: More indirect techniques to asking sensitive survey questions, https://analysights.wordpress.com/2010/06/23/randomized-responses-more-indirect-techniques-to-asking-sensitive-survey-questions/

John L, Acquisti A, Loewenstein G (2011) Strangers on a plane: context-dependent willingness to divulge personal information. *Journal of Consumer Research* 37(5):858-873.

Joinson AN, Paine C, Buchanan T, Reips U-D (2008) Measuring self-disclosure online: Blurring and non-response to sensitive items in web-based surveys. *Computers in Human Behavior* 24(5):2158-2171.

Kirchner A (2015) Validating sensitive questions: A comparison of survey and register data. *Journal of Official Statistics* 31(1):31-59.

Kreuter F, Yan T, Tourangeau R (2008) Good or bad item – can latent class analysis tell? The utility of latent class analysis for the evaluation of survey questions. *Journal of the Royal Statistical Society, Series A 171* (Part 3):723-738.

Krumpal I 2012 "Estimating the Prevalence of Xenophobia and Anti-Semitism in Germany: A Comparison of Randomized Response and Direct Questioning." *Social Science Research* 41: 1387–1403.

Kulka RA, Weeks MF, Folsom Jr. RE (1981) A comparison of the randomized response approach and the direct question approach to asking sensitive survey questions. Working Paper, Research Triangle Institute, NC.

Lamb C, Stern DE (1978) An empirical validation of the randomized response technique. *Journal of Marketing Research* 15:616-21.

Lensvelt-Mulders GJLM, Hox JJ, van der Heijden PGM, Maas CJM (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.

Ljungqvist L (1993) A united approach to measures of privacy in randomized response models: a utilitarian perspective. *Journal of the American Statistical Association* 88:97-103.

Marquis KH, Marquis MS, Polich JM (1986) Response bias and reliability in sensitive topic surveys. *Journal of the American Statistical Association* 81:381-89.

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.

Myerson R B (1986) Multistage games with communication. *Econometrica* 54:323-358.

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 JW, Park HN (1987) A new randomized response model for continuous quantitative data. *Proceedings of the College of Natural Science* 12:33-44.

Pollock KH, Bek Y (1976) A comparison of three randomized response models for quantitative data. *Journal of the American Statistical Association* 71:884-886.

Rohan T (2013) Antidoping agency delays publication of research. The New York Times, August 23, 2013.http://www.nytimes.com/2013/08/23/sports/research-finds-wide-doping-study-withheld.html

Rosenfeld B, Imai K, Shapiro, J (2015) An empirical validation study of popular survey methodologies for sensitive questions. *American Journal of Political Science*. doi: 10.1111/ajps.12205.

Sànchez-Pagés S, Vorsatz M (2007) An experimental study of truth-telling in a sender-receiver game. *Games and Economic Behavior* 61(1):86-112.

Scheers NJ (1992) Methods, plainly speaking: A review of randomized response techniques. *Measurement and Evaluation in Counseling and Development* 25:27-41.

Simmons JP, Nelson LD, Simonsohn U (2013) Life after p-hacking. Meeting of the Society for Personality and Social Psychology, New Orleans, LA, 17-19 January 2013. Available at SSRN: http://ssrn.com/abstract=2205186.

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

St John FAV, Keane AM, Edwards-Jones GE, Jones L, Yarnell RW, Jones JPG (2011) Identifying indicators of illegal behaviour: Carnivore killing in human-managed landscapes. *Proceedings of Royal Society Biological Science*.

Striegel H, Ulrich R, Simon P (2010) Randomized response estimates for doping and illicit drug use in elite athletes. *Drug and Alcohol Dependence* 106:230-232.

Tamhane AC (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 PE, Fox J (1981) The validity of randomized response for sensitive measurements. *American Sociological Review* 46:187-200.

Umesh UN, Peterson RA (1991) A critical evaluation of the randomized response technique. *Sociological Methods & Research* 20(1):104-38.

van der Heijden PGM, van Gils G, Bouts J, Hox JJ (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* 28:505-37.

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

Warren RP (1946) All the King's Men. New York: Time Incorporated.

Weissman AN, Steer RA, Lipton DS (1986) Estimating illicit drug use through telephone interviews and the randomized response technique. *Drug and Alcohol Dependence* 18:225-233.

Williams BL, Suen H (1994) A methodological comparison of survey techniques in obtaining self-reports of condom-related behaviors. *Psychological Reports* 7:1531-1537.

Wimbush JC, Dalton DR (1997) Base rate for employee theft: Convergence of multiple methods. *Journal of Applied Psychology* 82:756-63.

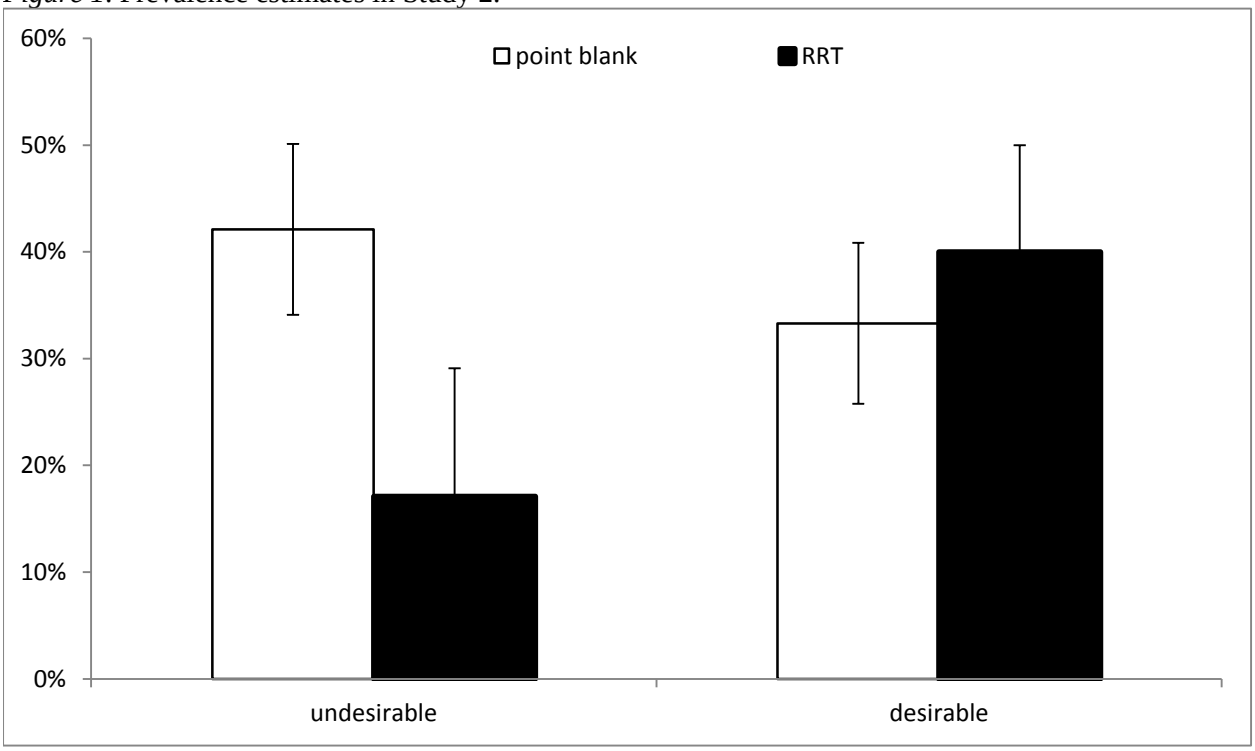
Wiseman F, Moriarty M, Schafer M (1975) Estimating public opinion with the randomized response model. *Public Opinion Quarterly* 39:507-513.

Wolter F, Preisendorfer P (2013) An evaluation of the randomized response technique versus direct questioning using individual validation data. *Sociological Methods & Research* 42(3):321-353.

Yan T, Kreuter F, Tourangeau R (2012) Latent class analysis of response inconsistencies across modes of data collection. *Social Science Research* 41:1017-1027.

Zdep SM, Rhodes IN, Schwarz RM, Kilkenny MJ (1979) The validity of the randomized response technique. *Public Opinion Quarterly* 43:544-49.

Figure 1. Prevalence estimates in Study 2.



Note: Error bars represent 1 standard error above/below the estimate.

Figure 2. Screen shots of response labels used in Study 4 (top panel depicts labels used in DQ and RRT-Standard Label; bottom panel depicts labels used in RRT-Revised Label).

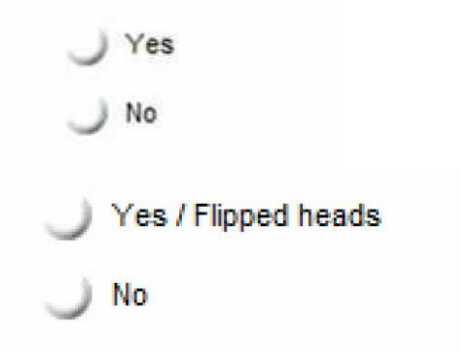
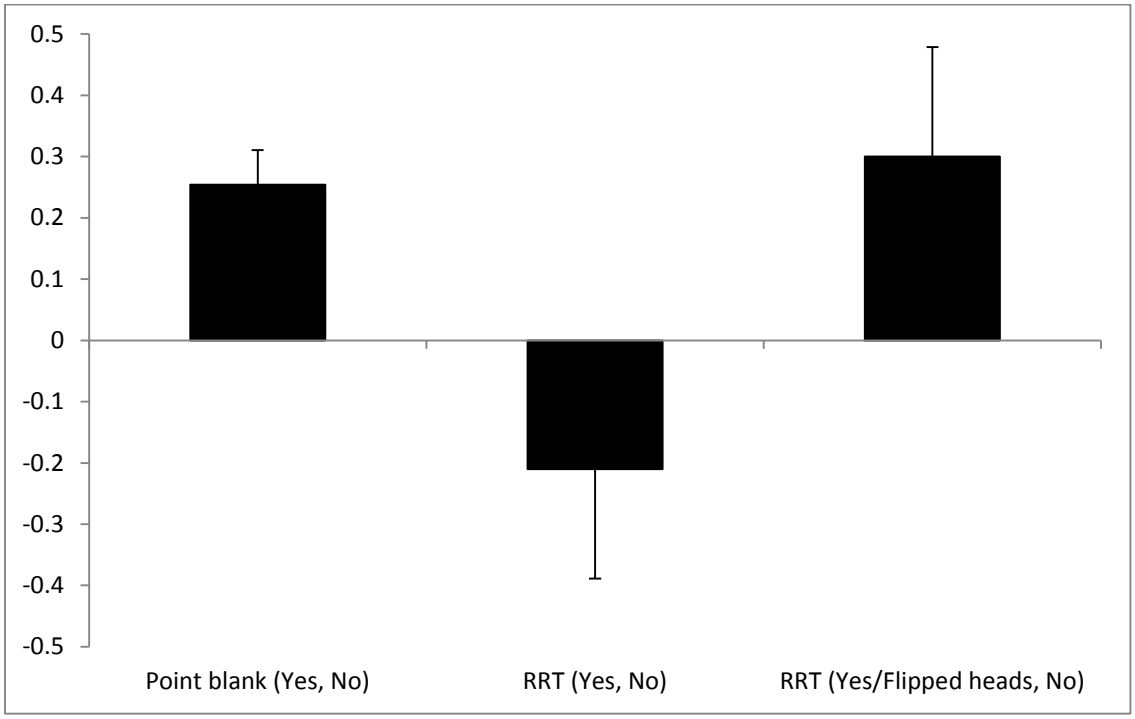
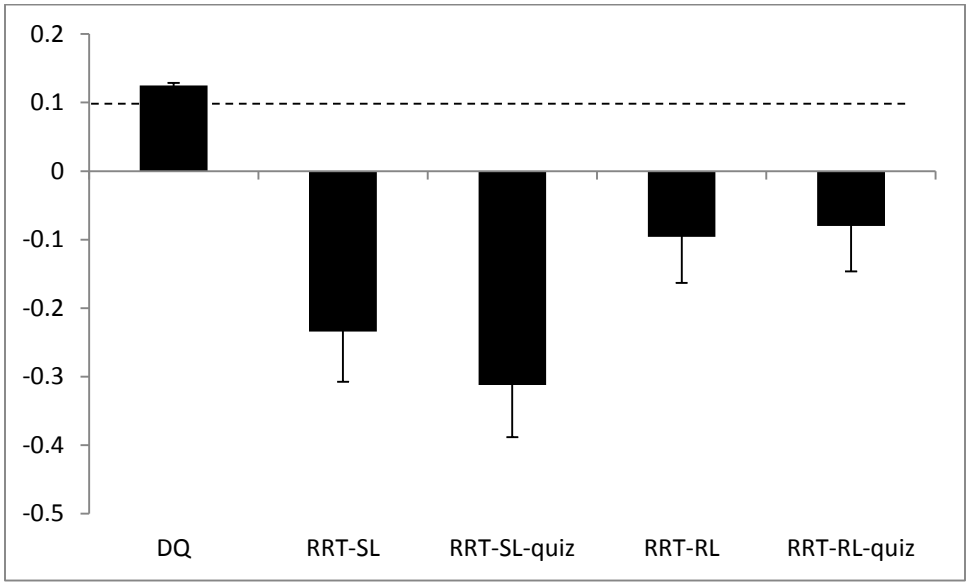


Figure 3. Prevalence estimates in Study 4A.



Note: Error bars represent 1 standard error above/below the estimate.

Figure 4. Prevalence estimates in Study 4B. Dashed line denotes true prevalence.



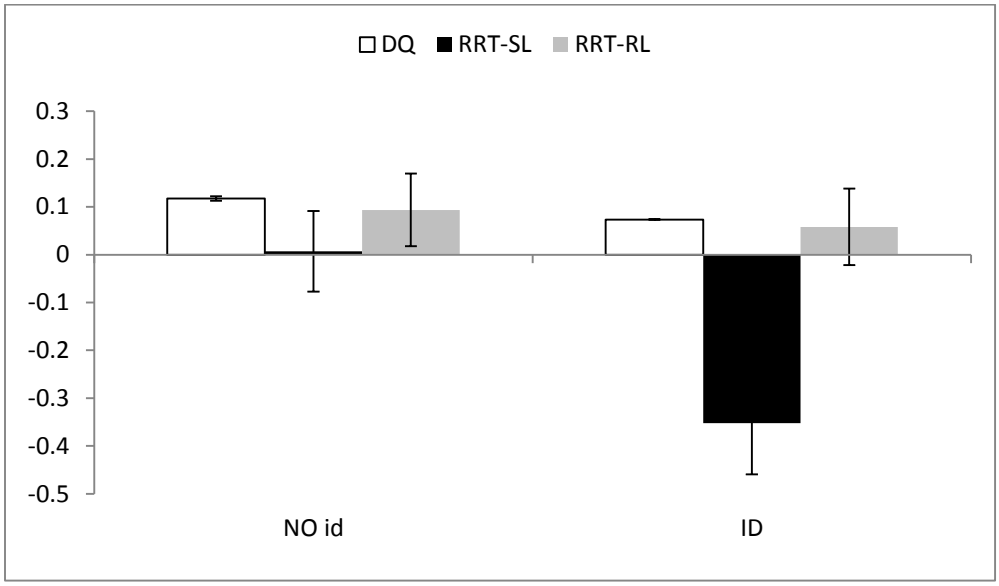
Notes:

- Error bars represent 1 standard error above/below the estimate.
- One may wonder why the prevalence estimate in DQ is higher than the true prevalence (dashed line). Although this could be indicative of boasting, we think it more likely to have arisen from the fact that in rare cases, IP addresses do not reflect a person's physical location. For example, if a respondent completed our survey from outside of the United States, but was connected to the internet through an American proxy server, his IP address would erroneously denote that he was completing the survey within the United States.

Figure 5. Results of pooled analysis (Study 4C).

Note: Error bars represent 1 standard error above/below the estimate.

Figure 6. Prevalence estimates in Study 5.



Note: Error bars represent 1 standard error above/below the estimate.