Organizational Behavior and Human Decision Processes 148 (2018) 101–123

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp

When and why randomized response techniques (fail to) elicit the truth

Leslie K. John[a,*], George Loewenstein[b], Alessandro Acquisti[c], Joachim Vosgerau[d]

[a] Baker Library 467, Negotiations, Organizations, and Markets Unit, Harvard Business School, Soldiers Field Drive, Boston, MA 02163, United States
[b] Department of Social and Decision Sciences, Carnegie Mellon University, United States
[c] Heinz College of Information Systems and Public Policy, Carnegie Mellon University, United States
[d] Marketing Department, Bocconi University, Italy

ARTICLE INFO

Keywords:
Truth-telling
Lying
Privacy
Information disclosure
Survey research

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. In nine experiments, we show that the noise introduced by RRTs can make respondents concerned that innocuous responses will be interpreted as admissions, and as a result, yield prevalence estimates that are lower than direct questioning (Studies 1–4, 5A, & 6), less accurate than direct questioning (Studies 1, 3, 4B, & 5A), and even nonsensical (i.e., negative; Studies 3–6). Studies 2A and 2B show that the paradox is eliminated when the target behavior is socially desirable, even when it is merely framed as such. Study 3 shows the paradox is driven by respondents' concerns over response misinterpretation. A simple modification designed to reduce concerns over response misinterpretation reduces the problem (Studies 4 & 5), particularly when such concerns are heightened (Studies 5 & 6).

"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

How prevalent is employee pilfering? How widespread is sexual harassment in the workplace? Is the misuse of prescription drugs more common than that of illegal drugs? The list of unethical and downright illegal behaviors that cause great damage to businesses and societies is almost endless: sexual harassment, blackmail, prostitution, online harassment, bullying, defamation, police abuse, favoritism, racism, sexism, fraud, and so on. Similarly long is the list of preferences and behaviors that are considered taboo, undesirable, or sensitive based on the norms of a given community or a group – from sexual preferences to political or religious affiliations. Determining the prevalence of such sensitive attitudes and behaviors is often critical for managerial decision making and public policy, but difficult because people are generally reluctant to disclose or to admit to holding and engaging in them.

Here, we examine the efficacy of an elicitation method called the randomized response technique (RRT) in obtaining truthful responses to questions about sensitive behaviors. The RRT is a type of indirect questioning technique designed to encourage truth-telling by introducing stochastic noise to the communication channel (Forges, 1986; Myerson, 1986; Warner, 1965). Respondents queried with the RRT are instructed to answer a sensitive question truthfully only with probability p, so affirmative responses cannot be interpreted at the individual level. In principle, this should increase disclosure by making individual responses less dispositive, while still making it possible to back out population-level prevalence rates. The RRT has been used to estimate the prevalence of a wide range of managerially relevant behaviors that are sensitive or illegal, including tax evasion (Houston & Tran, 2001), regulatory compliance (Elffers, Van der Heijden, & Hezemans, 2003), and dishonesty in certified public accountants (Buchman & Tracy, 1982).

There are, however, reasons to question whether the RRT is likely to improve truth-telling, or whether it may, instead, have paradoxical effects. Research on behavioral dimensions of privacy suggests that reassuring individuals about the security of their information can in some cases backfire, leading people to clam up rather than feeling freer to honestly share information (Singer, Hippler, & Schwarz, 1992). A conclusion of this 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 (Tamir & Mitchell, 2012). In contrast to conventional wisdom, therefore, privacy assurances can, in effect, ring "alarm bells," making people more, rather than less protective of their personal information. In one

* Corresponding author.
E-mail address: ljohn@hbs.edu (L.K. John).

https://doi.org/10.1016/j.obhdp.2018.07.004
Received 25 January 2017; Received in revised form 28 June 2018; Accepted 20 July 2018
Available online 17 August 2018
0749-5978/

study for example (Joinson, Paine, Buchanan, & Reips, 2008), participants were asked to answer an Internet-based survey that included sensitive personal questions. Participants who were first asked to complete a separate questionnaire measuring their Internet privacy concerns were subsequently less forthcoming with their personal information than participants who did not complete that questionnaire. These studies and others show that privacy concerns can be activated by environmental cues that bear little, or sometimes even a negative, relationship to objective dangers associated with information sharing. For example, casual-looking websites can keep privacy concerns at bay and induce people to share their personal information – a perverse effect given that such sites are typically less secure than their professional-looking counterparts (John, Acquisti, & Loewenstein, 2011). Applied to the RRT, these studies raise the question of whether obvious attempts to encourage truth-telling might backfire by activating privacy concerns that might be dormant if questions were asked more casually and directly.

Given these competing considerations, is the RRT an effective tool for obtaining truthful responses to sensitive queries, or, as the protagonist of Robert Penn Warren's classic proposes, is it better to ask questions directly — "quick and sharp?" Our research points to the surprising conclusion that the RRT commonly has the opposite of its intended effect, and provides insights into why this is the case.

2. Overview of the RRT

There are two basic varieties of RRTs: forced-response variants and forced-question variants. In forced-response variants (Boruch, 1971; Warner, 1965) – our focus as they are more common – respondents are randomized to either answer the question truthfully (with probability p), or to a forced-response condition (with probability 1-p) in which they are told which response option to endorse. For example, in the coin flip technique, a common instantiation which we focus on here, 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. Similarly, in forced-question variants (Greenberg, Abul-Ela, Simmons, & Horvitz, 1969), respondents are randomized to answer one of two questions: the sensitive, focal question (with probability p); or an innocuous one (with probability 1-p).

The RRT interviewer is blinded to the outcome of the randomizer, and so cannot tell whether any given stigmatizing response represents a truthful admission to the sensitive question. In forced-response variants, this means that the interviewer does not know whether endorsing a sensitive response denotes a forced stigmatizing response versus a real admission. Similarly, in forced-question variants, the interviewer does not know to which question – sensitive versus innocuous – a given admission corresponds. By correcting for the (known) probability that respondents were instructed to truthfully answer the focal question, the interviewer can deduce the population-wide prevalence of the behavior. In principle therefore, RRTs can be used to better estimate the prevalence of sensitive behaviors. In this paper, however, we show that RRTs often backfire, providing lower and less accurate estimates of the prevalence of sensitive attitudes and behaviors. We provide evidence supporting a particular account of why this occurs, based on the idea that people want to rule out potentially unfavorable interpretations of the RRT's noisy signal.

Psychologists and economists have long argued, and demonstrated with research, that people care about the image they project to others, and even themselves (Akerlof & Kranton, 2000; Bem, 1972; Dhar & Wertenbroch, 2012; Leary & Kowalski, 1990; Prelec & Bodner, 2003). For example, people engage in prosocial behavior in part to signal the commonly valued identity of being a good person (Ariely, Bracha, & Meier, 2009; Bénabou & Tirole, 2006; Gneezy, Gneezy, Riener & Nelson, 2012). Similarly, we propose that RRTs can backfire because they fail to eliminate the signaling process, while introducing greater leeway in responding. Respondents steered by the randomizer to endorse the stigmatizing answer, in forced-response variants; or to answer the benign question, in forced-question variants, may be worried that their innocuous responses will be misinterpreted as sensitive admissions. We therefore posit that RRTs can induce anxieties about being misunderstood – of sending the wrong signal to oneself and others – causing people to disobey the randomizer's instructions and to endorse the non-stigmatizing response pattern.

Consistent with this account, we show that the RRT can yield lower and less valid prevalence estimates than those obtained by direct questioning (DQ) because the RRT makes respondents concerned that innocuous responses will be misinterpreted as admissions. Based on this evidence, we propose a simple modification that reduces the magnitude of the problem. Before describing this empirical evidence, we first provide a review of previous RRT research.

3. Review of empirical RRT findings

The RRT has been, and continues to be, used in a variety of applications, all of which hinge on uncovering the prevalence of sensitive attitudes and behaviors. It is used in surveys designed to elicit the propensity of sensitive behaviors, such as drug use, sexual behavior and abortion, and sensitive attitudes such as antisemitism (Adler et al., 1992; Brewer, 1981; Chen et al., 2014; Krumpal, 2012), in turn informing policy decisions such as anti-drug doping policies in elite sports (Rohan, 2013).

Two types of studies have been used to assess the effectiveness of RRTs (Tourangeau & Yan, 2007; Umesh & Peterson, 1991): comparative studies and validation studies.

3.1. 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 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 (de Jong, Pieters, & Fox, 2010; Lara, Strickler, Olavarrieta, & Ellertson, 2004; Musch, Bröder, & Klauer, 2001; Rider, Harper, Chow, & Cheng, 1976; Shotland & Lynn, 1982). However, RRTs have also generated the same or lower, and hence presumably less valid, prevalence estimates relative to DQ (Bégin & Boivin, 1980; Beldt, Daniel, & Garcha, 1982; Brewer, 1981; Coutts & Jann, 2011; Duffy & Waterton, 1988; Goode & Heine, 1978; Höglinger, Jann, & Diekmann, 2014; Kulka, Weeks, & Folsom, 1981; Locander, Sudman, & Bradburn, 1976; Tamhane, 1981; Williams & Suen, 1994; Wiseman, Moriarty, & Schafer, 1975). For example, a national survey conducted by the Australian Bureau of Statistics to estimate the prevalence of drug use concluded that the RRT "did not significantly increase the number of affirmative responses to the controversial question, and was rather time-consuming" (Goode & Heine, 1978). Similarly, Weissman, Steer, and Lipton (1986) asked respondents whether they had used each of four illicit drugs (cocaine, heroin, PCP, and LSD) and found that drug usage estimates were equivalent across inquiry methods (RRT vs. DQ). Zdep, Rhodes, Schwarz, and Kilkenny (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 (e.g., Coutts & Jann, 2011; Hoglinger et al., 2014) have found the RRT to generate nonsensical *negative* prevalence estimates.

3.2. Validation studies

In validation studies, estimates are made of the prevalence of behaviors in situations in which the researcher, unbeknownst to participants, 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 latter type of study almost invariably involves deception, because the RRT only works if respondents do not believe that the questioner has access to the truth. 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, Hox, Van Der Heijden, & Maas, 2005), to the best of our knowledge only ten articles reporting validation studies have been published: two with aggregate-level validation data (Horvitz, Shah, & Simmons, 1967; Rosenfeld, Imai, & Shapiro, 2015), and eight with individual-level validation data (Fox, Avetisyan, & van der Palen, 2013; Kirchner, 2015; Kulka et al., 1981; Lamb & Stem, 1978; Locander et al., 1976; Tracy & Fox, 1981; van der Heijden, van Gils, Bouts, & Hox, 2000; Wolter & Preisendörfer, 2013).[1] In only one of these ten articles did the RRT consistently produce prevalence estimates equivalent to the true prevalence (Lamb & Stem, 1978). In another article (Horvitz et al., 1967), one study produced accurate prevalence estimates but a second study produced estimates significantly higher than true prevalence. And in the remaining eight articles, the RRT significantly underestimated true prevalence.

However, especially for sensitive topics, it is probably unrealistic to expect RRTs to eliminate response bias. It is perhaps for this reason that nine of the ten validation articles also included a DQ comparison group. Some of these articles obtained RRT prevalence estimates that were systematically higher than DQ (Fox et al., 2013; Lamb & Stem, 1978; van der Heijden et al., 2000; Rosenfeld et al., 2015). However, others documented mixed or null effects of the RRT relative to DQ (Locander et al., 1976; Kirchner, 2015; Tracy & Fox, 1981; Wolter & Preisendörfer, 2013), and one study found RRT prevalence estimates that were significantly lower than those obtained using DQ (Kulka et al., 1981). As with the comparative studies, especially given potential publication bias, these null and negative effect studies are noteworthy.

There may even be reason to question some of the evidence that RRTs generate higher prevalence estimates than DQ, to the extent that the RRT may have protected the "wrong" – i.e., non-stigmatizing, answer. Consider Rosenfeld et al. (2015), an aggregate-level validation study measuring Mississippians' propensity to reveal how they voted on a proposed constitutional amendment to define life as beginning at conception as opposed to birth. Respondents were asked whether they had voted "yes" as a function of different questioning techniques, including RRT (coin flip method) and DQ. The RRT was implemented based on the premise that voting "yes" — a "pro-life" stance — is taboo: respondents who flipped "heads" were instructed to answer "yes," regardless of whether they had, in fact, voted yes; those who flipped "tails" were instructed to answer truthfully. Relative to DQ, RRT prevalence estimates were closer to the true prevalence. However, given the state's conservative reputation, and the fact that the amendment passed, voting "no" could have been the taboo behavior. The RRT may have thus provided protection for the wrong answer option.

Protecting the wrong answer option is problematic because it can artificially increase prevalence estimates, giving illusory evidence of RRT effectiveness. In Rosenfeld et al. (2015), those possessing the stigmatized trait (i.e., who had voted "no") and instructed by the randomizer to respond truthfully may find it unfair that by virtue of a trivial coin flip, they have to admit to a stigmatized behavior. The coin flip could have just as easily come up heads, these respondents may readily rationalize, in which case they would have been not only permitted to, but required to, give the non-stigmatized response. This salient counterfactual may tempt participants to act as if the coin flip had, in fact, turned up heads, by giving the non-stigmatizing response. In our studies we 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 what extent this may have happened, and hence it is difficult to interpret the results of this study.

In sum, in examining previous validation studies, Wolter & Preisendörfer (2013) concluded that "the results in favor of the RRT are not convincingly strong." We concur.

3.3. Meta-analyses

Two comprehensive meta-analyses of RRT studies have been conducted. Umesh and Peterson (1991) — based on the five validation studies that had been conducted to date — 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 also included 32 comparative studies 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). However, contrary to the "more-is-better" assumption (Tourangeau and Yan, 2007), for items on which boasting was plausible (e.g., "How often do you donate blood?") the authors presumed the *lower* estimate to be the more valid estimate. Such recoding is potentially problematic; for example, it assumes that boasting is more likely in DQ than RRT but we are unaware of research indicating this to be the case.

Finally, a meta-analysis of DQ on sensitive topics such as income, drug use, and health — based on validation studies — found that direct questioning yielded surprisingly accurate results (Marquis, Marquis, & Polich, 1986): no systematic underreporting of population values was found. Together, the results of these three meta-analyses, plus those of the individual comparative and validation studies reviewed above, are inconclusive. 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 & Preisendörfer, 2013). Yet, the RRT continues to be used (e.g., Insight Central, 2010; Krumpal, 2012; Rohan, 2013) in the face of its unexplained failures.

3.4. Non-adherence

Clearly, respondents often fail to adhere to RRT instructions, which can produce undesired results — prevalence estimates lower than DQ estimates or even impossible (negative or in excess of 100%). Edgell, Himmelfarb, and Duchan (1982), in an attempt to quantify the seriousness of the non-adherence problem, surreptitiously recorded the

[1] Umesh and Peterson (1991) consider two additional papers to be validation studies: Edgell, Himmelfarb, and Duchan (1982) and Akers, Massey, Clarke, and Lauer (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 & Peterson, 1991).

outcome of the randomizer, and found that 25% of respondents answered "no" when the randomizer had instructed them to answer "yes."

3.5. Statistical corrections for non-adherence

Sophisticated statistical methods measure non-adherence to correct for it post-hoc. Clark and Desharnais (1998) measured 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 target behavior can be estimated, as well as the rate of non-adherence to RRT instructions. After accounting for expected variation from the randomizer, prevalence estimates should be invariant with respect to the probability of a forced "yes"; any observed differences can therefore be used to calculate non-adherence. Ostapczuk, Moshagen, Zhao, and Musch (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, Musch, and Moshagen (2011) estimated a 38.9–55.2% non-adherence rate.

The second approach to estimating non-adherence for post-hoc correction of RRT estimates uses latent class models (Böckenholt & van der Heijden, 2007; Cruyff, Bockenholt, van den Hout, & van der Heijden, 2008; Cruyff, van den Hout, van der Heijden, & Böckenholt, 2007). Each construct is measured with multiple items and the RRT is applied to each item. This allows for estimating each respondent's probability of belonging to a latent class of non-adherents, as well as their probability of belonging to a latent class of having exhibited the target behavior. Using the multi-item latent class approach, Böckenholt and van der Heijden (2007) estimated 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 non-adherence of 10% in a Dutch sample, and de Jong, Pieters, and Stremersch (2012) estimated rates ranging from 5% (Brazil, Japan) to 20% (Netherlands) to 29% (India).

Although ingenious and informative, these post-hoc correction methods have limitations. First, from a practical standpoint, non-adherence varies greatly between populations and topics. Thus these post-hoc correction methods require non-adherence to be measured anew for different populations and topics. Doing so requires randomizing respondents between two different probabilities of answering the sensitive question, effectively doubling the required size of the sample, which is already necessarily large relative to those sufficient for direct questioning. These methods can also be cumbersome to administer; for example, using the latent class model approach, 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, to the best of our knowledge, these post-hoc correction methods have not been empirically validated, in the sense that non-adherence estimates gleaned by these correction methods have not been compared to true non-adherence rates. For a further discussion of whether statistical post-hoc corrections successfully address the non-adherence issue see Appendix 1.

In sum, techniques correcting for non-adherence have merit but they can be cumbersome, and, given their assumptions, it is not entirely clear whether they produce more valid prevalence estimates. But even despite these concerns, these methods correct for non-adherence after-the-fact; it would be preferable, if possible, to avoid the problem altogether. To do so, the reasons for non-adherence must be understood.

3.6. Explaining non-adherence

Failure to comply with the RRT instructions can be either unintentional or intentional. Unintentional non-adherence arises when participants do not understand the RRT instructions. For example, participants misunderstanding the randomizer's instruction to give a forced "yes" response may instead respond with a truthful "no" response. Consistent with this explanation, a study on fraudulent health insurance benefit claims indicated that clarity of instructions and participants' education level were negatively related to the number of "no" answers when a "yes" answer was required (Böckenholt & Van der Heijden 2007). Although such unintentional non-adherence contributes to prevalence estimate distortion, our studies suggest that it is not the full explanation (the RRT backfires even when participants must pass a quiz about the procedure prior to answering the focal question, Study 4B).

Intentional non-adherence occurs when respondents disobey the randomizer despite understanding its instructions, which can occur for several reasons. First, people who have not engaged in a sensitive behavior may respond "no" even when the technique calls for them to respond "yes" regardless of whether they have, so as to avoid giving a response that provides any ambiguity about whether they have engaged in the behavior (this idea was proposed by Campbell, 1987). Second, people who have engaged in the behavior may also not wish to provide a response that has any possibility of being interpreted as an admission. As a result, they may deny having engaged in the behavior when they are supposed to respond honestly, and may also respond that they have not when the technique calls for them to respond that they have, regardless of whether they have done so. Böckenholt and van der Heijden (2007) called such non-adherence "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.

Intentional non-adherence could explain previous findings that, at first blush, may seem surprising. In one study respondents, who the researchers secretly knew to have criminal records, were asked how many times they had been arrested (Tracy & Fox, 1981), either via DQ or RRT. In the RRT condition, half of participants were asked to answer the question truthfully. The other half were instructed to report a specific number; their chance of being instructed to say: "zero" was 28%; "1" was 28%; "2" was 16%; "3" or "4" each with probability 8%; and "5," "6," "7," or "8," each with probability 4%. Overall, prevalence estimates were higher when elicited using RRT than with DQ (though still lower than true prevalence); however, subgroup analysis suggested that the RRT was only effective for those who had been arrested more than once. For those arrested "only" once, the RRT backfired relative to DQ. This pattern could have arisen because those arrested once were more likely to be forced to give a stigmatized response – to report a number higher than their true value – relative to those who had been arrested multiple times. Specifically, in the RRT condition, those who had been arrested once faced a 24% chance of being forced to give a number higher than one; for those arrested more than once, this likelihood was only 0 to 16%, depending on the number of arrests for the given participant.

Intentional non-adherence can also explain why some studies have found nonsensical *negative* prevalence estimates (Brewer 1981; Coutts & Jann, 2011; Hoglinger et al., 2014). In the coin flip technique for example, if a large number of participants responds with "no" when forced by the randomizer to answer "yes," the overall proportion of "yes" responses can become smaller than 50% (which is the proportion of "yes" responses resulting from coin flips), producing a nonsensical, negative prevalence estimate. In sum, despite the existence of different methods of addressing nonadherence, there is ongoing debate about the usefulness of the RRT.

4. Overview of the present research

In nine studies, we investigate whether and why RRT prevalence estimates are lower and less valid than those obtained through simple direct questioning. We consistently document paradoxical effects of RRTs — estimates that are lower than DQ (Studies 1–4, 5A, & 6), less valid than DQ (Studies 1, 3, 4B, & 5A), and even impossible (negative prevalence estimates, Studies 3–6). We begin with a simple demonstration of the paradox using a validation study (Study 1). In five subsequent experiments, we show that the paradox occurs in part because RRTs make respondents concerned that innocuous responses will be misinterpreted as admissions. Consistent with this explanation, paradoxical effects are pronounced when: the behaviors in question are socially undesirable (Studies 2A & 2B); and among subgroups particularly likely to be concerned over response misinterpretation: respondents who have not engaged in the behavior (Studies 3, 5A, & 5B) and respondents who are identifiable (i.e., not responding anonymously, Study 6). Study 3 also provides direct evidence of the process proposed to underlie the paradox, showing that the propensity to respond affirmatively is mediated by respondents' concerns over response misinterpretation. Finally, a simple modification to the RRT designed to reduce concern over response misinterpretation reduces the problem (Studies 4A, 4B, 5A, & 5B), particularly when this concern is acute, whether due to characteristics of the respondent (Studies 5A & 5B) or of the situation (Study 6).

Most studies include a DQ condition as a benchmark. Studies 1, 3, 4B, 5A, and 5B 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 effective at eliciting truthful answers to sensitive questions. 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 scholars, managers, and policy makers who are interested in obtaining valid sensitive information.

For all studies, the sample size was determined independently from the results: In Study 1 it was dictated by the number of people who had participated in a previous study; in Studies 2B and 4A it was based on a pre-set time window; and in all other studies it was based on a pre-set target sample size. The studies were run over a span of several years; therefore the variation in sample sizes across studies reflects evolving best practices in behavioral science (cf. Simmons, 2014; Simmons, Nelson, & Simonsohn, 2011). We report all manipulations and measures. No data were excluded from the analyses unless explicitly indicated.

5. Study 1

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

5.1. Materials and methods

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, Amir, & Ariely (2008): Participants answered trivia questions, were given an answer key and asked to report the number of questions they had answered correctly, and were paid based on these self-reported scores. Unbeknownst to participants, the workbooks into which they had written their answers were collected and linked to their self-reported scores. Therefore, we knew whether each participant had cheated (by overstating his or her score), and could 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. Because 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 more likely to overstate their score than to understate it, so we suspect that even low OSs indicate cheating. For example, the percentage of participants who overstated their score by exactly one point (34.9%) was much higher than the percentage understating it by one point (5.7%; $\chi^2(1) = 23.62$, $p < .001$), and most participants (58.1%) had an OS of one or higher.[2]

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 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, $SD = 6.4$), a response rate of 56.3%. Upon clicking the link in the email, participants were randomly assigned to one of two inquiry conditions (DQ vs. RRT). Because 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 this and all studies, we oversampled RRT relative to DQ.[3]

All participants were instructed:

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

The question, "Did you overstate your score in this study?" was accompanied by 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

[2] 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 condition.

[3] In earlier studies, we set the sampling ratio of RRT relative to DQ to 2:1 (Studies 1 & 6). 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 (Study 2A), to 4:1 (Study 3), and finally, to 6:1 (Study 4B).

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 2).

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.

In all studies, to determine the aggregate prevalence estimate in RRT (denoted by *t-hat* in the equation below), we adjusted the fraction 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. This adjustment is based on the procedure outlined by Warner (1965); additional details are provided in Appendix 3. 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.

In this and all studies, we used an intent-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. The results across studies are similar, and in some cases stronger, when we treat these participants as missing data (i.e., when we assume that blank responses to the focal question denote neither affirmations nor denials). In Studies 3, 4B, 5A, and 5B multiple participants attrited prior to randomization (11.5% in Study 3, 9.1% in Study 4B, 7.0% in Study 5A, and 3.5% in Study 5B); therefore these responses could not be included in the analysis. We suspect this attrition arose because in each of these studies, prior to randomization, participants faced a moderately invasive question (participants were asked to indicate their physical location), coupled with a somewhat annoying task (answering quiz questions).

5.2. Results and discussion

Participants who completed the follow-up survey were 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 (57.1% 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, 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; $\chi^2(1) = 0.67$, $p = .41$) — i.e., random assignment worked.

The cheating prevalence estimate was 24.3% in the DQ condition, and only 4.8% in the RRT condition ($t(197) = 1.97$, $p = .05$). In both inquiry conditions, the percentage of participants who admitted to having cheated was lower than the true cheating prevalence within the given inquiry condition (RRT: prevalence estimate = 2.6%[4] vs. true prevalence = 54.9%; $p < .001$; DQ: prevalence estimate = 27.4% vs. true prevalence = 61.3%; $p < .001$).

6. Studies 2A & 2B

Study 1 shows that RRTs can generate lower and less valid prevalence estimates relative to DQ. In Study 2 we test whether this effect arises because of concern over response misinterpretation: respondents who flip heads may fail to check "yes" due to concern that their response will be misinterpreted as an admission. If lower RRT estimates are driven by this concern, the paradox should disappear for questions that, despite being intrusive, inquire about socially desirable behaviors, relative to equally intrusive questions that inquire about socially undesirable behaviors (Study 2A). It should also disappear when the same target behavior is framed as socially desirable versus undesirable (Study 2B). The logic is that when it is socially desirable to respond affirmatively, respondents who flip 'heads' should not be concerned about having their "yes" responses misinterpreted as admissions (and they might even like the ambiguity introduced by the RRT).

7. Study 2A

In Study 2A each participant was asked whether they had engaged in two different behaviors, both of which were intrusive, but only one of which was socially undesirable. Between-subjects, we manipulated the inquiry method; half of participants were asked, point blank, whether they had engaged in the behaviors, while the other half were asked using the RRT. To ensure that results are not due to a particular set of questions, between-subjects we manipulated the question set: Each participant was randomized to one of five different question pairs, pretested to be equivalent with respect to intrusiveness but to differ in social desirability. The study was therefore a $2 \times 2 \times 5$ mixed design.

7.1. Materials and method

Pretest. Participants ($N = 80$ MTurk workers; 62.5% male; $M_{age} = 33.4$ years, $SD = 9.5$) were presented with 66 questions (Appendix 4) and rated each on two dimensions: intrusiveness and social desirability. For intrusiveness, participants were asked: "We are interested in how sensitive you think these questions are – that is, how intrusive would it feel to be asked each of these questions?" using the response scale "Not at all intrusive," "Somewhat intrusive," "Intrusive," and "Very intrusive." For social desirability, participants were asked: "We are interested in how taboo it is to engage in each of these behaviors. In other words, how socially approved – or not socially approved – is it to do each of these things?" using the response scale "Not at all taboo," "Somewhat taboo," "Taboo," and "Very taboo." Half of participants were randomized to first rate intrusiveness, then social desirability; the other half rated in the opposite order (there were no order effects). Based on this pilot, we identified five pairs of items that were equivalent in intrusiveness but significantly different in social desirability (Appendix 5). In the main study, each participant was randomized to one of these five question pairs.

[4] This prevalence estimate (2.6%) is different than that reported above (4.8%) because the former is restricted to participants for whom we had OSs. Likewise, the DQ prevalence estimate in this sentence (27.4%) is different from that above (24.3%) for the same reason. 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).

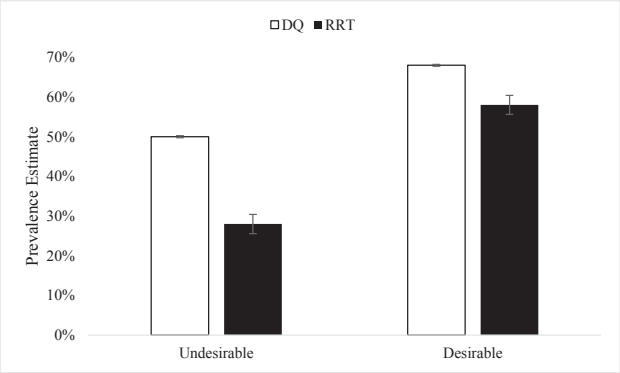

Fig. 1A. Prevalence estimates in Study 2A. Error bars represent 1 standard error above/below the estimate.

Participants (N = 827; 52.3% male; M_{age} = 37.4, SD = 11.4) were recruited through MTurk in exchange for a small fixed payment and randomly assigned to either a DQ or RRT condition (the RRT was described using the same text as Study 1). As a control variable, the order of presentation of the two questions was randomized between-subjects (this factor had no effect therefore the results collapse across it).

7.2. Results

A logistic regression revealed that overall, admission rates were higher for the socially desirable behaviors than the socially undesirable ones ($\beta_{desirable}$ = 0.75, $p < .001$); and when asked directly as opposed to using the RRT ($\beta_{RRT} = -0.94$, $p < .001$). Of more relevance to our hypothesis, these main effects were qualified by an interaction between inquiry method and social desirability ($\beta_{RRT*desirable}$ = 1.75, $p < .001$; Fig. 1A). Follow-up testing revealed that the estimated prevalence of the socially undesirable behaviors was lower using the RRT relative to DQ (RRT = 28%, DQ = 50%; $t(825) = 5.19$, $p < .001$). Although for the socially desirable behaviors the RRT prevalence estimates were also lower than from DQ (RRT = 58%, DQ = 68%; $t(825) = 2.32$, $p = .021$), the magnitude of this difference pales in comparison to that of the undesirable behaviors. Table 1 depicts the prevalence estimates by item and inquiry method. We designed the study to test for differences in prevalence estimates aggregated across items, therefore not surprisingly most of the individual RRT versus DQ comparisons are non-significant (we were not sufficiently powered to detect such differences).

8. Study 2B

Study 2A is consistent with the idea that, because of concerns over response misinterpretation, the RRT is prone to backfiring for socially undesirable behaviors. Although we pre-tested the questions used in Study 2A to differ in social desirability but to be matched in intrusiveness, it is nonetheless possible that the socially undesirable behaviors differed in some other respect relative to their desirable counterparts. Study 2B addresses this issue by manipulating the social desirability of the same target behavior. The study was a 2 × 2 between-subjects design manipulating inquiry method (DQ vs. RRT) and behavior framing (desirable vs. undesirable).

8.1. Materials and method

The study was conducted at a private US university's satellite lab in a large office building. We collected as much data as we could in the two days we had been allotted to run the study. Office workers (N = 158; 37.4% male; M_{age} = 43.6 years, SD = 12.4) were recruited as they walked by and were offered a chance at a $100 gift card for completing an 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 asked 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 2A, 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.

Table 1
Prevalence estimates by question and inquiry mode.

Pair #	Question	DQ prevalence	RRT prevalence	t test of RRT vs DQ
1	Have you ever picked up someone at a bar?	28%	33%	t(164) = 0.60, p = .55
	Have you ever flatulated audibly in public?	67%	57%	t(164) = 1.05, p = .30
2	Have you ever showered with a partner?	80%	72%	t(165) = 0.68, p = .05
	Have you ever been late in paying a bill or your rent?	77%	61%	t(165) = 1.79, p = .08
3	Are you currently in a relationship that involves physical intimacy?	98%	83%	t(164) = 1.32, p = .19
	Have you ever bounced a check?	36%	13%	t(164) = 2.11, p = .04
4	Do you maintain (e.g. shave, wax, trim) your pubic hair?	67%	62%	t(163) = 0.50, p = .62
	Have you ever had sexual thoughts about a member of your same sex?	44%	33%	t(163) = 1.19, p = .24
5	Did you reach orgasm the last time you had sex?	66%	44%	t(165) = 2.66, $p < .05$
	Have you ever had sexual desires for a minor?	20%	−21%	t(165) = 3.31, $p < .05$

Within each pair, the first question refers to a behavior that is relatively socially desirable; the second question to a behavior that is socially undesirable.

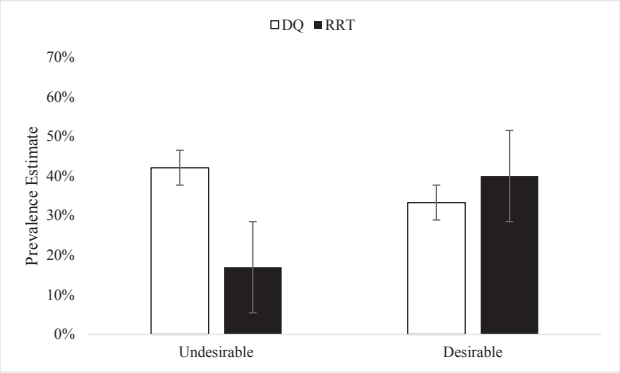

Fig. 1B. Prevalence estimates in Study 2B. Error bars represent 1 standard error above/below the estimate.

8.2. Results

A logistic regression revealed a main effect of inquiry method ($\beta_{RRT} = -1.26$, $p = .017$) and, of more relevance to our hypothesis, an interaction between inquiry method and social desirability ($\beta_{RRT*desirable} = 1.55$, $p = .028$; Fig. 1B). Follow-up testing revealed that when texting while driving was framed as socially undesirable, its estimated prevalence was marginally lower using RRT relative to DQ (DQ = 42.1%, RRT = 17.0%; $t(78) = 1.70$, $p = .09$). When the behavior was framed as socially desirable however, the RRT estimated prevalence was not statistically different from the DQ estimate (DQ = 33.3%; RRT = 40.0%; $t(78) = 0.82$, $p = .42$).

Studies 2A and 2B support the idea that the RRT backfires in part because it creates concern that innocuous responses will be misinterpreted as admissions to sensitive behaviors. Consistent with this idea, the RRT indicated lower prevalence rates than DQ when intrusive questions asked about socially undesirable behaviors, but the two methods yielded similar estimates when intrusive questions asked about socially desirable or neutral behaviors. Likewise, when texting while driving was framed as socially undesirable, RRT prevalence estimates were lower compared to DQ, but when texting while driving was framed as socially desirable, both inquiry methods yielded similar results. In addition to showing that the RRT is especially likely to backfire when used to estimate the prevalence of socially undesirable behaviors, these studies also help to rule out the possibility that the results of Study 1 are 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.

Despite the success of the framing manipulation, we do not advocate socially desirable 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. This could be particularly likely if such framing operates via perceptions of the popularity of the behaviors: Given that socially undesirable behaviors tend to be relatively uncommon, framing a behavior as socially desirable may cause a person to deem it common, increasing comfort in admitting to, and hence engaging in, the behavior (Cialdini, Reno, & Kallgren, 1990; Goldstein, Cialdini, & Griskevicius, 2008).

9. 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 than DQ. In Study 3 we provide direct evidence of our explanation for the paradox in two ways. First, we measure concern over response misinterpretation, enabling us to test whether this concern mediates the propensity to respond affirmatively. Second, we recorded the outcome of the randomizer; this information, coupled with the validation data, meant that we knew how each participant should have responded to the focal question – i.e., the correct answer for each participant. We test whether incorrect responding is particularly strong when correct responding is psychologically difficult: among innocent individuals who are concerned about their responses being misinterpreted (i.e., those who have not engaged in the behavior instructed by the randomizer to respond "yes"); and possibly among guilty individuals asked to tell the truth (i.e., to admit to having lied).

9.1. Materials and method

Participants ($N = 756$; 66.2% male; $M_{age} = 34.3$ years, $SD = 9.6$) were recruited through MTurk in exchange for a small fixed payment. At the beginning of the survey, participants were asked: "Are you completing this survey from a location within the United States?" Later, participants 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 induce 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)." Participants who endorsed the latter, "No" response option were then directed to the math questions. Participants who endorsed the former, "Yes" response option skipped the math question portion of the survey.[5] Unbeknownst to participants, we collected their IP addresses to validate that they were in fact not in the United States.[6]

Manipulation. Next, participants were asked if 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 the previous studies, 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:

"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.

[5] In a previous study, we found that answering the math questions did not interact with the subsequent inquiry method condition.

[6] We validated the IP addresses using this coding tool: http://software77.net/geo-ip/multi-lookup/

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 — 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. Effectively then, for the purpose of testing our hypotheses, there were three inquiry conditions of interest: DQ, RRT-forced-yes, and RRT-answer-truthfully.

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*).

At the end of the survey participants were asked to supply basic demographic information, including birth month, which we used to separate responses in the RRT condition based on whether participants had been instructed to respond "yes" versus answer truthfully. Although the propensity to answer the birth month item was different between conditions (percent of respondents providing their birth month: DQ = 98.5%; RRT-forced-yes = 100%; RRT-answer-truthfully = 92.1%; *Fisher's exact* = 27.08, $p < .001$), the difference is small in magnitude — across all conditions almost all (i.e., 96.5% of) participants supplied this information. There were no differences in reported birth month 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 added noise, making it more difficult to detect differences between conditions.

9.2. Results and discussion

Prevalence estimates. Twenty-three percent of participants lied about their physical location (i.e., said they were completing the survey from the United States when in fact they were not; *NS* between conditions). The RRT (collapsing across randomizer outcome) produced a lower and less valid prevalence estimate (−25.6% — a negative prevalence estimate) relative to DQ (19.4%; $t(668) = 7.23$, $p < .001$). DQ accurately measured prevalence — it produced a prevalence estimate (19.4%) that was not different from the true prevalence (23%, $p = .67$). By contrast, the RRT generated a nonsensical, negative estimate (−25.6%).

Concern over response misinterpretation. There were differences in concern over being misinterpreted as a function of inquiry condition ($F(2, 636) = 11.23$, $p < .001$).[7] Specifically, participants in the RRT-forced-yes condition reported greater concern over being misinterpreted relative to those in the RRT-answer-truthfully condition ($M_{\text{RRT-forced-yes}} = 4.71$, SD = 2.17; $M_{RRT\text{-answer-truthfully}} = 3.86$, $SD = 2.23$, $t(505) = 4.37$, $p < .001$) and the DQ condition ($M_{\text{DQ}} = 3.84$, $SD = 2.39$; $t(245) = 3.51$, $p = .001$).

A mediation analysis revealed that the relationship between inquiry method and the propensity to respond affirmatively ($\beta_{\text{RRT}} = 1.32$, $SE = 0.14$, $p < .001$) was reduced when concern over misinterpretation was included in the model ($\beta_{\text{RRT}} = 1.26$, $SE = 0.14$, $p < .001$; $\beta_{\text{concern}} = 0.18$, $SE = 0.04$, $p < .001$), 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, $SE = 0.03$, $p = .001$).

Correct responses. In this study we know each participant's guilt status (i.e., whether they lied), as well as the outcome of the randomizer. As a result, we know the correct response for each participant – i.e., how each participant should have answered the question. We used this information to calculate the percentage of correct responses in each condition (see Appendix 6 for calculations), which served as a measure of inquiry method performance.

A logistic regression revealed a main effect of both inquiry method ($\beta_{\text{inquiry}} = -1.25$, $p < .001$) and guilt status ($\beta_{\text{guilt}} = -1.85$, $p < .001$) on the propensity to answer correctly. However, these main effects were qualified by a significant interaction ($\beta_{\text{inquiry*guilt}} = 1.32$, $p < .001$). Consistent with our account, relative to DQ, correct responding in RRT was lowest when it was particularly psychologically difficult to do so (Fig. 2): among innocent individuals forced to respond "yes," and among guilty individuals asked to tell the truth (i.e., to admit to having lied). Specifically, for the innocent, the percentage of correct responses is lower among those forced to respond "yes" relative to those asked by the randomizer to tell the truth (RRT-forced-yes = 64.2%; RRT-tell-truth = 92.3%; $\chi^2(1) = 46.65$, $p < .001$), and also relative to DQ (DQ = 91.4%; $\chi^2(1) = 25.95$, $p < .001$). The latter two groups – innocents in the RRT condition instructed to tell the truth, and innocents in the DQ condition – had equivalent rates of incorrect responding ($\chi^2(1) = 0.067$, $p = .80$).

For the guilty, the percentage of correct responses is lower among those asked by the randomizer to tell the truth relative to those forced to respond "yes" (RRT-tell-truth = 41.7%; RRT-forced-yes = 59.4%; $\chi^2(1) = 3.89$, $p = .049$). It is also marginally significantly lower relative to DQ (DQ = 63.0%; $\chi^2(1) = 3.38$, $p = .066$) – noteworthy because these conditions make the same request of participants: answer the question truthfully. Yet, the RRT backfires for these individuals; it made guilty individuals slightly more uncomfortable admitting than they otherwise would, had they simply been asked directly.

10. Studies 4A & 4B

Taken together, Studies 1–3 suggest that RRTs can backfire because they make respondents concerned that innocuous responses may be misinterpreted as incriminating. Studies 4A and 4B test a possible antidote: 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. (1982)'s 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 and 4B 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 from this perspective Study 4A is a conservative test of the basic hypothesis that the RRT can backfire.

11. Study 4A

11.1. Materials and method

Participants ($N = 162$; 38.8% male; $M_{\text{age}} = 34.7$ years, $SD = 12.4$) 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

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

Fig. 2. Percentage of correct responses by condition, Study 3.

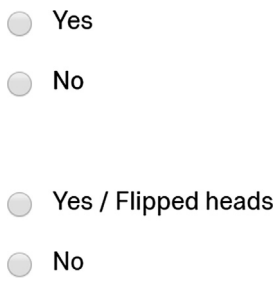

Fig. 3. Screen shots of response labels used in Study 4A (top panel depicts labels used in DQ and RRT-Standard Label; bottom panel depicts labels used in RRT-Revised Label).

RRT conditions; in both, participants were given the standard RRT explanation and instructions. In the standard label (RRT-standard) condition, participants were presented with the same response scale as DQ. In the revised label (RRT-revised) condition, the response options were labeled: "yes/flipped heads" and "no." The RRT-revised 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 target behavior. Screen shots of the response labels, by condition, are shown in Fig. 3. In the RRT-standard condition, we predicted prevalence estimates to be lower than DQ (i.e., a replication of the paradoxical RRT effect). However, we predicted the RRT-revised condition to reduce or possibly even alleviate the paradox, by generating a prevalence estimate greater than RRT-standard (and in the case of full alleviation of the paradox, it would generate an estimate equal to or higher than DQ).

11.2. Results and discussion

As predicted, prevalence estimates were lower in RRT-standard relative to DQ (RRT-standard = -21.0%, DQ = 25.4%; $t(100) = 2.47$, $p = .015$), but not in RRT-revised relative to DQ (RRT-revised = 30.0%, DQ = 25.4%; $t(117) = 0.50$, $p = .62$). 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.

12. Study 4B

In Study 4A, a subtle revision to the RRT response label designed to address concerns over response misinterpretation eliminated the paradox: prevalence estimates in the RRT-revised condition were indistinguishable from those in the DQ condition. Although we propose that this effect occurs because the revised label reduces concerns over response misinterpretation, one could argue that the unusual-looking revised label works by reducing unintentional non-adherence, prompting respondents to read, and hence comply with, the RRT instructions. Study 4B addresses 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 (all participants must pass a quiz about the procedure before answering the target question).

12.1. Materials and method

Participants ($N = 1{,}249$; 57.3% male; $M_{age} = 31.4$ years, $SD = 10.3$) were recruited through MTurk in exchange for a small fixed payment. Participants were first asked to indicate whether they were completing the survey from a location within the US. 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-standard, or RRT-revised. Again, as in Study 3, we collected respondents' IP addresses to validate their location claims.

12.2. Results and discussion

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). Fig. 4 presents mean estimated prevalence rates in all five conditions. Although prevalence estimates in all RRT conditions were different from the true prevalence of lying (10.0%), collapsing across the quiz manipulation, prevalence estimates were higher in RRT-revised compared to RRT-standard, suggesting that the revised label was partially effective in reducing concerns over response misinterpretation (RRT-standard = −27.4%, RRT-revised = −8.8%; $t(1046) = 2.62$, $p = .001$) – only partially effective because it was lower than the true prevalence (10%) and also in this case lower than the DQ prevalence estimate (DQ = 12.5%, $t(611) = 3.62$, $p < .001$).[8]

Importantly, there was no difference in prevalence estimates as a function of the quiz in either the RRT-standard conditions (NoQuiz = −23.4%, Quiz = −31.1%; *NS*) or the RRT-revised

[8] One may wonder why the prevalence estimate in DQ is directionally higher than the true prevalence. Although this could be indicative of boasting, we think it is more likely to have arisen from the fact that occasionally, 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.

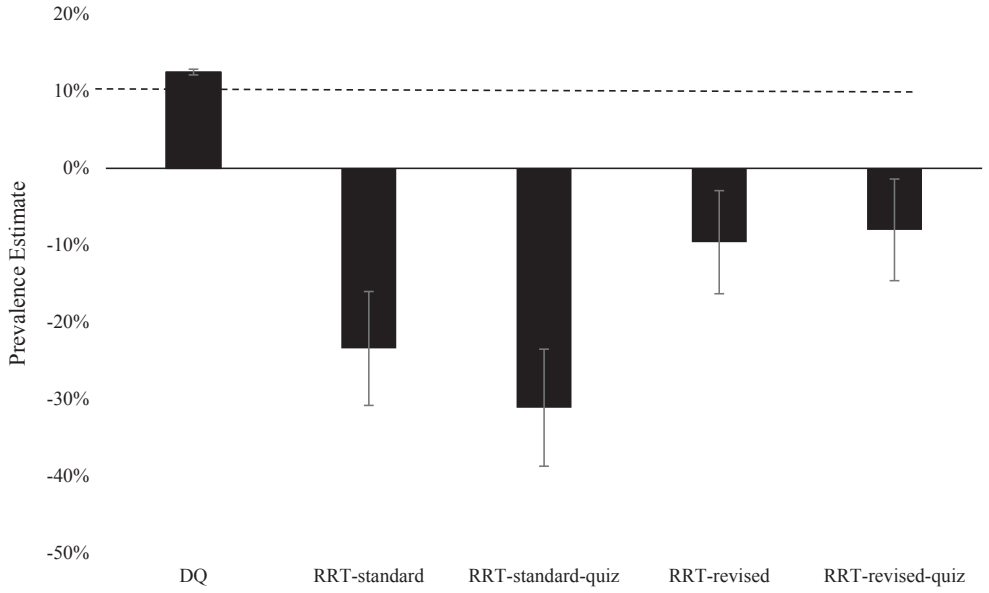

Fig. 4. Prevalence estimates in Study 4B. Dashed line denotes true prevalence. Error bars represent 1 standard error above/below the estimate.

conditions (NoQuiz = −9.6%, 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 concerns over response misinterpretation.

13. Studies 5A & 5B

Studies 5A and 5B test whether the magnitude of the paradox, and the benefit of the revised label, depends on whether the respondent has engaged in the target behavior. We have proposed that the revised label works because it addresses concerns of response ambiguity – respondents who flip heads no longer feel as though they are inadvertently incriminating themselves by checking 'yes.' Therefore, one would expect the advantage of the RRT-revised label over the RRT-standard label to be pronounced among innocent individuals, for they may be particularly concerned over inadvertently incriminating themselves.

Although the desire to avoid incriminating oneself is likely to be active, at least to some degree, for all respondents, we propose it to be heightened among the innocent – i.e., those who have not engaged in the target behavior. Consistent with this idea, research in criminology suggests that falsely accused individuals often insist on their innocence, sometimes even when doing so entails substantial risk; for example, defendants may reject attractive plea bargains that require an admission of guilt but that offer a dramatically reduced penalty relative to what would be sought in trial (Dervan & Edkins, 2013; Tor, Gazal-Ayal, & Garcia, 2010). Thus, given the particular anguish over false incrimination, we propose that the revised label is particularly likely to increase accuracy of responding among innocent individuals forced to respond "yes."

We had initially attempted to test this idea by exogenously varying the ratio of innocent to guilty individuals between experimental conditions. For example, using a similar validation set-up as Studies 3 and 4B, we manipulated the financial incentive to lie. Although we replicated the basic paradox in these studies – RRT prevalence estimates were lower than both actual prevalence and DQ estimates – our lying inductions failed to induce sufficiently different lying rates between conditions (in our strongest manipulation, 25 percent of participants lied when given a strong incentive, compared to 13 percent that lied when given a weak incentive). These experiments are nonetheless informative because, as we do next in Study 5A, their validation data enable the results to be separated based on whether the respondents engaged in the target behavior (in this case, lying).

In Study 5A, we pooled the data from nine of the ten validation studies we have conducted to date — three of which are included in this paper (Studies 1, 3, and 4B); the remaining six are integrated into the results of Study 5A (the tenth validation study is Study 5B, described next). We report the results in two ways: first using only the studies reported in this paper, then using all studies. Study 5B is a validation study in which, as in Study 3, we knew both whether each participant was guilty, as well as the randomizer outcome. Thus Study 5B enables us to pinpoint the benefit of the revised label: we expected it to be particularly effective at increasing adherence among innocent individuals instructed to answer "yes."

14. Study 5A

14.1. Materials and method

In all nine validation studies, participants (N = 4844) were asked whether they had engaged in a sensitive behavior (either lying about one's physical location, seven studies; or cheating on a previous task, two studies). Method of inquiry (DQ vs. RRT) was randomized between-subjects. 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-standard, RRT-revised) differed by whether participants had engaged in the given behavior.[9]

14.2. Results and discussion

Validation studies reported in this paper. Regardless of engagement status, prevalence estimates were highest in DQ (19.0%) and lowest in RRT-standard (−23.6%); prevalence estimates in RRT-revised fell in between (−8.8%). All pairwise comparisons were significantly

[9] We do not have validation data for 102 participants (2.1% of the sample of 4844 participants). In some cases this was because IP addresses were the source of validation data and the participants had blocked their IP addresses and hence their physical location could not be ascertained. And in other cases cheating was the source of validation data, as in Study 1, and 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.

different from each other (all *ps* < .001). Each inquiry method produced a prevalence estimate that was significantly lower than the true prevalence in the given condition (RRT-standard = −23.6% vs. true prevalence of 20.8%, $p < .001$; RRT-revised = −8.8% vs. true prevalence of 10.1%, $p < .001$; DQ = 19.0%, vs. true prevalence of 25.0%, $p = .02$).[10]

More interestingly, the RRT-standard backfired for both those who had (guilty) and who had not (innocent) engaged in the behavior (prevalence estimates… at guilty: DQ = 43.7%, RRT-standard = −5.8%; $t(314) = 5.53$, $p < .001$; at innocent: DQ = 10.8%,[11] RRT-standard = −28.4%; $t(1139) = 8.64$, $p < .001$; Fig. 5A). Moreover, among the innocent, the RRT-revised generated more valid prevalence estimates, bringing prevalence estimates significantly closer to zero relative to the standard label (RRT-standard = −28.4%; RRT-revised = −8.8%; $t(1397) = 3.07$, $p = .002$). Among the guilty, prevalence estimates were equivalent across the two different RRT label conditions (RRT-standard = −5.8%; RRT-revised = −9.4%; $t(296) = .22$, $p = .83$). We now turn to the full analysis using all nine validation studies.

All validation studies. Regardless of engagement status, prevalence estimates were highest in DQ (18.2%) and lowest in RRT-standard (−26.0%); prevalence estimates in RRT-revised fell in between (−12.2%). All pairwise comparisons were significantly different from each other (all *ps* < .001). Each inquiry method produced a prevalence estimate that was significantly lower than the true prevalence in the given condition (RRT-standard = −26.0% vs. true prevalence of 20.9%, $p < .001$; RRT-revised = −12.2% vs. true prevalence of 17.5%, $p < .001$; DQ = 18.2%, vs. true prevalence of 23.8%, $p < .001$).[12]

More interestingly, the RRT-standard backfired for the guilty and the innocent (prevalence estimates… at guilty: DQ = 48.9%, RRT-standard = 10.2%; $t(633) = 6.95$, $p < .001$; at innocent: DQ = 8.6%, RRT-standard = −35.6%; $t(2292) = 13.55$, $p < .001$; Fig. 5B). The RRT-revised generated more valid prevalence estimates for innocents, bringing prevalence estimates closer to zero (RRT-standard = −35.6%, RRT-revised = −18.0%; $t(2947) = 3.94$, $p < .001$). Among the guilty, prevalence estimates were equivalent across the two different RRT label conditions (RRT-standard = 10.2%, RRT-revised = 14.4%, $t(713) = 0.64$, $p = .52$).

15. Study 5B

15.1. Materials and method

Participants ($N = 1946$; 64.4% male; $M_{age} = 36.0$ years, $SD = 13.3$) were recruited through MTurk in exchange for a small fixed payment.

Fig. 5A. Results of pooled analysis using the validation studies in this manuscript (Study 5A). Error bars represent 1 standard error above/below the estimate.

Fig. 5B. Results of pooled analysis using all validation studies (Study 5A). Error bars represent 1 standard error above/below the estimate.

Study 5B used a similar validation set-up as Study 3. As in Study 3, participants were told that participants indicating that they are completing the survey from the US would be permitted to skip the math questions. Also as in Study 3, for efficiency of data collection and unbeknownst to participants, we advertised the study only to people who were *not* in the United States.

Participants were asked: "Are you completing this survey from a location within the United States? (note: if you say yes to this question, you won't have to answer the math questions. And your answer to this location question will have no impact on your payment… we will pay you the full payment for this HIT regardless of whether you say you are in the US)." Participants who did not lie (i.e., who indicated that they were outside of the United States) went on to complete the math questions.

Next, using the RRT, all participants were asked the focal question: "Earlier in this survey, you were asked whether you are completing this survey from outside of the United States. Did you lie on this question? (please note that your response to the question below will not affect your payment for this study)." We used the same randomizing procedure as Study 3 in which participants were instructed to answer based on their birth month: half of participants were instructed to answer "yes" regardless of whether they had lied if they were born in the first half of the year (as a control variable, this mapping was reversed for the other half of participants; the mapping had no effect therefore our results collapse across this factor). Between-subjects, we manipulated the response option labels: half of participants received the standard labels (i.e., "Yes," and "No,"); the other half received the revised labels ("Yes/I was born in first [second] half of year"; and "No").

[10] The true prevalence was significantly lower in the RRT-revised relative to both the RRT-standard and DQ. We suspect this was because the RRT-revised was not administered in all of the validation studies reported in this paper; specifically, it was only administered in Study 4B, where the true prevalence tended to be lower than that in Studies 1 and Study 3, which did not include a RRT-revised label condition. Consistent with this explanation, the true prevalence was statistically equivalent in RRT-standard and DQ.

[11] Although the true prevalence among innocent is 0%, on occasion DQ produced a prevalence estimate slightly higher than this rate. We suspect this is because we validated location by IP address, which is an excellent but imperfect proxy for location (for example when certain types of VPN software are in use, a computer may transmit an IP address different than the country in which it is physically located).

[12] As with the preceding analysis restricting the sample to the validation studies within this paper, when using data from all nine validation studies, the true prevalence was significantly lower in the RRT-revised relative to both the RRT-standard and DQ. Again we suspect this is because the RRT-revised was not administered in all of the validation studies. Consistent with this explanation, the true prevalence was statistically equivalent in RRT-standard and DQ.

Fig. 6A. Prevalence estimates by inquiry method, Study 5B. Error bars represent 1 standard error above/below the estimate.

15.2. Results and discussion

Prevalence estimates. Thirty-nine percent of participants lied about their physical location (i.e., said they were completing the survey from the United States when in fact they were not; *NS* between conditions). The RRT-standard produced a lower and less valid prevalence estimate (−2.2% — a negative prevalence estimate) relative to the RRT-revised (11.8%, $t(1876) = 3.20$, $p = .001$). More interestingly, the performance discrepancy between the RRT-standard and the RRT-revised appears to be driven by the innocent (Fig. 6A): among the innocent, the prevalence estimate in the RRT-standard (−29.2%) is significantly less valid relative to that of the RRT-revised (−8.0%, $t(1145) = 3.12$, $p = .002$). By contrast, among the guilty, the prevalence estimate in the RRT-standard (43.0%) is equivalent to that of the RRT-revised (41.0%, *NS*).

Correct responding. To pinpoint the locus of this advantage, we conducted follow-up tests using the percentage of correct responses as the outcome measure (facilitated by having recorded the randomizer outcome). We were particularly interested in whether the benefit of the revised label lies in its capacity to increase correct responses among innocent individuals instructed by the randomizer to answer "yes." A logistic regression testing the effect of guilt status (innocent vs. guilty), RRT label type (standard vs. revised), and instruction (tell the truth vs. say yes) revealed a significant 3-way interaction ($\beta_{guilt*label*instruction} = 1.53$, $p < .001$). We decomposed this interaction by examining the innocent versus guilty separately, conducting two parallel logistic regressions assessing the effect of label type and randomizer outcome on correct responding.

Among the innocent and consistent with Study 3, the percentage of correct responses was lower when instructed to say "yes" relative to when instructed to tell the truth ($\beta_{instruction} = -1.99$, $p < .001$). A significant interaction suggested that this decrement was buffered by the revised label ($\beta_{label*instruction} = 1.20$, $p = .002$, Fig. 6B). Specifically, among innocents instructed to respond "yes," the percentage of correct responses was higher in RRT-revised (86.0%) relative to RRT-standard (64.9%), $\chi^2(1) = 33.57$, $p < .001$. By contrast, among innocents instructed to respond truthfully, the percentage of correct responses was identical in the standard versus revised label conditions: 93.1%. This makes sense, because it is easy for innocent individuals to respond correctly when asked to tell the truth (i.e., to respond "no").

Among the guilty and consistent with Study 3, the percentage of correct responses was lower when instructed to respond truthfully relative to when instructed to say yes ($\beta_{instruction} = 0.889$, $p < .001$). Unlike for innocents, for the guilty, we did not have a strong prediction of where the locus of the revised label benefit might lie. There was a significant interaction ($\beta_{inquiry*instruction} = 0.861$, $p = .021$, Fig. 6B). Follow-up tests revealed that the revised label marginally significantly increased the percentage of correct responses among guilty individuals instructed to respond "yes" (RRT-revised = 88.0%; RRT-standard = 81.0%; $\chi^2(1) = 3.25$, $p = .071$). Among those instructed to tell the truth, correct responding was unaffected by the label (RRT-revised = 56.0%; RRT-standard = 63.6%; $\chi^2(1) = 2.19$, $p = .14$).

In sum, the consistent finding emerging from Studies 5A and 5B is that the RRT-revised is particularly effective at increasing correct responding among innocent individuals – a subgroup we posited to be particularly concerned about response misinterpretation. Study 5B delved deeper into the locus of this effect, demonstrating that the benefit of the RRT-revised is seen predominantly in the situation in which innocents would be most concerned over response misinterpretation: when they are instructed to respond "yes."

16. Study 6

Studies 2–5 are consistent with the explanation that RRTs can backfire because they introduce apprehension over response misinterpretation. Studies 4 and 5 show that addressing this concern improves prevalence estimates, although the modified version of the RRT did not outperform DQ. The latter point is not particularly surprising: though the RRT-revised 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. Plus, the revised label is a very subtle modification. Although the RRT-revised 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, along the lines of Study 5B, when the advantage of the revised label (relative to the standard label) is expected to be pronounced.

Studies 5A and 5B showed that the relative advantage of the revised label is moderated by a characteristic of the respondent (guilt status). Study 6 tests whether it is moderated by a characteristic of the situation that we posited to affect concerns over response misinterpretation: the stakes of responding affirmatively. We did so by varying the extent to which participants were identifiable (in the studies so far, following typical RRT administration, we did not ask respondents for identifiable information). The study was a 3 × 2 between-subjects design in which we manipulated the inquiry method (DQ/RRT-standard/RRT-revised) and the stakes of responding affirmatively (low vs. high). First, we predicted that the RRT-standard would backfire relative to DQ,

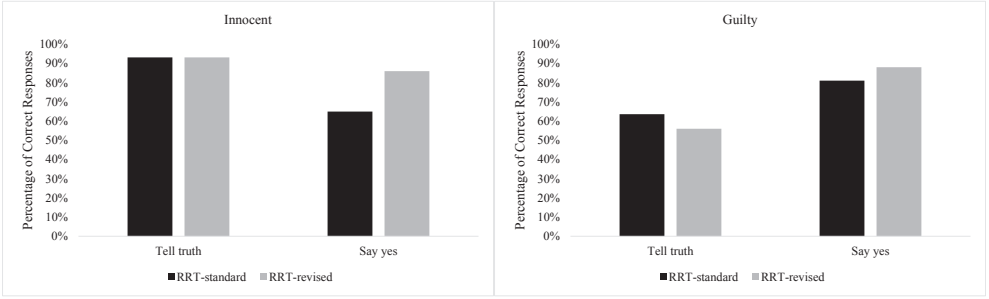

Fig. 6B. Percentage of correct responses among the innocent (left panel) and among the guilty (right panel), Study 5B.

especially when respondents are identifiable. Second, we predicted that for such participants, the relative advantage of the revised label (over the standard one) would be particularly great.

16.1. Materials and method

Participants (N = 695; 50.6% male; M_{age} = 33.9 years, SD = 12.7) 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 was not asked for 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-standard, or RRT-revised.

16.2. 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).

Prevalence estimates in RRT-standard (−17.2%) were lower relative to both RRT-revised (7.6%; $t(554) = 2.81$, $p = .005$) and DQ (9.6%; $t(414) = 3.66$, $p < .001$), but not in RRT-revised 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 (Fig. 7). Among participants in the identified condition, the prevalence estimate in RRT-standard (−35.2%) was lower relative to both RRT-revised (5.8%; $t(276) = 3.08$, $p < .005$) and DQ (7.4%; $t(206) = 3.82$, $p < .001$). However, the RRT-revised was no different from the DQ ($t(205) = 0.19$, *NS*). Prevalence estimates in the anonymous conditions were similar across conditions, although directionally consistent with our other studies (RRT-standard = 0.8%; RRT-revised = 9.4%; DQ = 11.8%; all comparisons *NS*).

The pattern of results is pronounced when the identified condition is restricted to the 82.8% of participants who actually provided identifying information (prevalence estimate among Ss who provided identifying information: RRT-standard = -19.3%; RRT-revised = 19.7%; DQ = 8.6%).

17. General discussion

RRTs are intended to make people 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, 5A, & 6), less valid than DQ estimates (Studies 1, 3, 4B, & 5A), or even impossible (i.e., negative, Studies 3–6). Beyond providing new evidence that RRTs can backfire, our studies also present new evidence about *why* they backfire. The data suggest that RRTs backfire due to respondent concerns over response misinterpretation – in particular, the concern among innocent people that a "yes" response will be interpreted as meaning they have engaged in a behavior when they have not. Supporting such an interpretation, Studies 2A & B show that the RRT underperformance is mitigated when apprehension over response misinterpretation is reduced; for undesirable behaviors, the RRT backfired, but for target behaviors that are socially desirable the RRT performed as well as DQ. Study 3 provides direct evidence for this explanation, by showing that concern over response misinterpretation mediates the relationship between inquiry method and the propensity to respond affirmatively. Studies 4A and 4B show that modifying the response labels to address this concern improves RRT performance, although not to the point of surpassing the performance of DQ. Finally, Studies 5A, 5B and 6 show that the advantage of the revised label relative to the standard RRT is greatest in situations in which concerns of response misinterpretation are heightened.

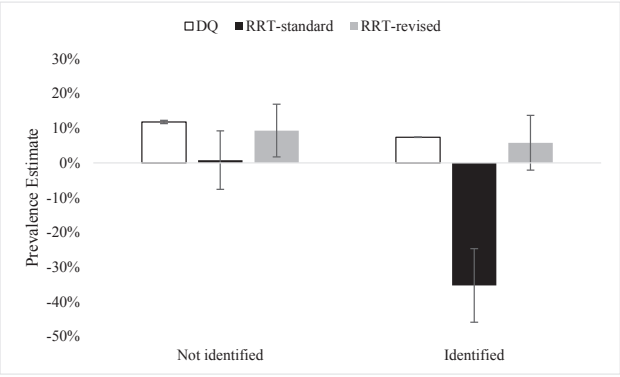

Fig. 7. Prevalence estimates in Study 6. Error bars represent 1 standard error above/below the estimate.

In the studies in this paper, RRT prevalence estimates, including those obtained using the revised RRT, were always lower than actual prevalence estimates, and were only, at best, equal to DQ estimates. These results are consistent with Umesh and Peterson (1991)'s conclusion that "contrary to common beliefs (and claims), the validity of the RRM [RRT] does not appear to be very good," but stand in 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 the discrepancy between our findings and those of Lensvelt-Mulders et al. (2005), a conservative interpretation of our results is the RRT may, in at least some cases, perform worse than DQ. Can we go further, however, and identify the conditions in which the RRT is especially prone to underperforming?

17.1. When does the RRT underperform?

Given that the RRT's intended benefit is in facilitating sensitive disclosures, it has been suggested that its relative advantage over DQ may be reduced in situations that respondents already perceive to be minimally threatening — for example, when respondents are anonymous. However, our research suggests the opposite: the RRT performs poorly in the very situations in which it is intended to perform well. Indeed as Study 6 demonstrated, the RRT performs particularly badly when respondents are identifiable. In the same vein, it has been argued that the RRT's performance should be heightened when inquiring about taboo behaviors, but Studies 2A and 2B provide direct empirical evidence that it performs worse relative to DQ in this situation. Future research might test additional moderators of RRT performance as well as proxies for nonadherence. For example, RRT performance may be moderated by aspects of the randomizer (Hoglinger et al., 2014), such as trust that it is truly random and cannot be monitored. Although this prediction is intuitive, Coutts and Jann (2011) found that, although participants trusted the randomizer more to the extent that they could control it (for example, by flipping a physical coin as opposed to a virtual one), a virtual coin flip produced greater compliance and higher prevalence estimates than the private flip of a real coin.

A peculiar difficulty in applying the RRT is that researchers need to know which behaviors are undesirable, sensitive, or taboo. If researchers mistakenly protect the wrong answer option – as may have happened in the previously discussed Rosenfeld et al. (2015) study measuring Mississippians' propensity to define life as beginning at conception as opposed to birth – prevalence estimates can be artificially increased, thus inflating (or giving illusory support for) evidence of the RRT's effectiveness. Furthermore, when choosing to protect one answer option or another as the socially undesirable behavior, a researcher may not only assume that respondents share her/his perception of which

behavior is socially undesirable, but also that there is little or no heterogeneity in respondents' perceptions of social undesirability. This may be true for many socially undesirable behaviors. However, there are surely also behaviors for which there is considerable heterogeneity in perceptions of social undesirability – for example drug use, uncommon sexual practices, in-group favoritism, extra-marital relationships, pro-life versus pro-choice beliefs, etc. The less respondents agree on which behavior is the socially undesirable one, the less can the RRT protect the stigmatizing response. The empirical consequences of such heterogeneity have yet to be explored.

Two alternative techniques to protect respondents' privacy and encourage truthful responding, which do not rely on randomization of answers, are the item count technique (ICT, also called the unmatched count technique, unmatched block design, or block total response; Miller, 1984; Raghavarao & Federer, 1979), and the crosswise-model (Yu, Tian, & Tang, 2007). Both techniques seem less susceptible to the problems of protecting the wrong answer option and of non-adherence to instructions, however, neither technique has been empirically validated. For a more detailed discussion of these techniques see appendix 7.

17.2. A broader perspective

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; Sánchez-Pagés & Vorsatz, 2007), they may not be so in some circumstances, and in such situations there is no reason to think that the RRT will have the intended effect (e.g., Blume, Lai, & Wooyoung, 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.

Moreover, to the extent that respondents do have a desire to tell the truth, the choice to use privacy-protecting questioning techniques could nonetheless reduce truth-telling. Respondents may infer that because the researchers deem an RRT necessary to inquire about a given behavior, the behavior being asked about must be sensitive and/or engaging in the behavior must be seen as undesirable. Consistent with this idea, in an exploratory study we asked respondents two questions about cheating: (1) whether they had cheated on their taxes, and (2) whether they had "cheated" in a relationship. We always asked one of these questions using DQ, and the other using the coin flip method of RRT, and randomized across subjects which question was asked using which technique. We then asked respondents which of the two forms of cheating they believed that we (the researchers) viewed as worse (more undesirable). People generally believed that we, the researchers, viewed the behavior inquired about with the RRT, as worse than that inquired about with DQ. Respondents may therefore infer that the use of the RRT (or any other privacy-protecting technique) is a signal that they are being asked about a behavior that is likely to be condemned, which could reduce their willingness to admit engaging in it, and to comply with the randomizer's instructions. The use of privacy-protecting techniques could thus lead to a self-fulfilling prophecy, creating the very apprehension that they seek to mitigate.

Indeed, a wealth of research in organizational behavior and allied fields points to how assumptions that precede actions can be self-fulfilling. For example, when a management system facilitates supervisor monitoring of subordinates, it causes the supervisors to believe that those subordinates cannot be trusted – in the sense that without the monitoring, the subordinates would not be intrinsically motivated to perform (Strickland, 1958; Kipnis, 1972). In turn, employers' beliefs about their employees can shape employees' beliefs and behavior toward the employer. For example, imposing systems to monitor and control employees tells those employees that their employer believes they cannot be trusted, which can cause those employees to both distrust the employer and to withhold effort when they cannot be observed or punished for doing so (Cialdini, 1966; Fehr & Rockenbach, 2003; Kruglanski, 1970; Lingle, Brock, & Cialdini, 1977; Tenbrunsel & Messick, 1999). The use of privacy-protecting techniques to elicit sensitive disclosures may represent a kind of self-fulfilling prophecy in which researchers' assumptions about human nature – regardless of their accuracy – become true (Jones, 1986; Ferraro, Pfeffer, & Sutton, 2005; MacGregor, 1960). The more we use these techniques, the more people may become concerned about their privacy and the less they may be motivated to trust companies, follow instructions, and tell the truth.

All techniques discussed in this paper, the RRT, DQ, ICT, and the cross-wise model are privacy-protective techniques for overt inquiries (i.e., asking data subjects to reveal sensitive information). However, firms are increasingly capable of covertly collecting sensitive information on their employees and consumers through a variety of digital surveillance tools, that may obviate or reduce the need to rely on techniques that attempt to elicit disclosures overtly. Such covert collection tools facilitate information harvesting in part because they operate outside of the data subject's awareness. However, when people become aware that a firm has been covertly collecting their personal information, it can cause them to respond negatively toward that firm (Kim, Barasz, & John, 2018). The firm may be left with the challenge of restoring eroded trust – a task that has proven difficult to accomplish, especially when people feel misled (Schweitzer, Hershey, & Bradlow, 2006), as is plausible when covert data collection is exposed. On the other hand however, research also suggests that people adapt to privacy invasions (Acquisti, John, & Loewenstein, 2012). Given this adaptation, and the challenges in overtly eliciting information from data subjects (of which this research is an example), for better or for worse, firms may increasingly rely on covert data collection techniques. Or perhaps more likely, firms may supplement information gleaned covertly with that elicited overtly. Indeed firms have been increasingly adopting tools from behavioral research in the design of online interfaces so as to increase users' propensity to disclose (John, 2015; Adjerid, Acquisti, & Loewenstein, in press).

17.3. Conclusion

The research presented here suggests ways to decrease respondent apprehension about response misinterpretation, and hence the likelihood that the RRT will backfire. However, in none of our experiments did the RRT, even with such modifications, produce more accurate propensity measures than DQ. We cannot rule out that such situations exist; perhaps the RRT produces more accurate propensity measures for behaviors that are extremely undesirable or illegal, or perhaps there are subpopulations for whom the RRT will outperform DQ. Yet, our results should, at a minimum, reinforce concerns raised by prior research about whether the RRT is generally worth using, especially given its time costs and the random noise it introduces. Given a choice between RRT and DQ, the current research suggests, DQ should be the default for researchers, barring specific evidence that the RRT will provide better estimates in a particular study.

More broadly, the present research demonstrates that providing people with protections that, logically, should make them more forthcoming with information, can in fact backfire. 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. Taken together, these findings highlight how people's willingness to divulge can be at odds with the objective consequences of information revelation.

Acknowledgements

We thank Daniel McDonald for statistical consulting; Clayton Critcher and Rik Pieters for comments; and Marina Burke, Holly Howe, and Trevor Spelman for help with data collection.

Appendix 1:. A discussion of previous attempts to prevent, and correct for, non-adherence

Symmetric forced-response RRT: A way of preventing non-adherence?

Although attempts to address non-adherence have focused on post-hoc statistical correction (described earlier), Bourke's (1984) "symmetric forced-response" variant is designed to prevent intentional non-adherence. Typically, the RRT only inserts noise into stigmatized responses, preserving people's ability to unambiguously endorse the non-stigmatized behavior – an asymmetry that creates the temptation to give the non-stigmatized response, regardless of the randomizer's instruction. The symmetric forced-response variant removes this temptation by also adding noise to the non-stigmatized response. Specifically, respondents are instructed to either: respond truthfully (with probability t) or to provide one of two forced responses: either a stigmatized response (with probability s), as in the asymmetric variant, *or* a non-stigmatized response (with probability 1-t-s). However, this fix still leaves the problem, central to our account, of mis-signaling on the stigmatized response option: those forced to give a stigmatized response may fail to comply because they do not want to come across as having engaged in the stigmatizing behavior. And indeed a recent validation study using symmetric forced-response produced RRT estimates that were significantly lower than true prevalence and equivalent to DQ (Wolter & Preisendörfer, 2013).

Can statistical post-hoc correction methods eliminate the non-adherence problem?

Extant post-hoc correction methods have not been empirically validated, in the sense that non-adherence estimates gleaned by these correction methods have not been compared to true non-adherence rates. Such validation data would be important in evaluating these correction methods; in particular the Clark and Desharnais (1998) 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 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 latent class models assume adherence to be equivalent across items and across randomizer outcome. In other words, it is assumed that non-adherence is just as likely for benign questions as it is for sensitive questions (an assumption refuted by Wolter & 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 we refute in Study 3). Violations of this local independence assumption (i.e., errors of individual items are uncorrelated) can lead to biased prevalence estimates (Kreuter, Presser, & Tourangeau, 2008; Tan, Kreuter, & Tourangeau, 2012). Finally, non-adherence is estimated by counting the number of times participants give non self-incriminating responses. The method thus cannot distinguish between non-adherence versus non-engagement in the behaviors, which could make it prone to overestimating non-adherence.

Concluding, statistical post-hoc corrections for non-adherence have yet to demonstrate that they actually improve accuracy of prevalence estimates by diminishing the prevalence of non-adherence.

Appendix 2:. Instructions from RRT studies showing positive effects:

Zdep et al. (1979)

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 and 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.

Barth and 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?"

Van der Heijden et 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 h 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 3:. RRT prevalence estimator

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

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

The expected value of this estimator is given by

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

The variance is given by

$$Var[\hat{t}] = \frac{Var[Y]}{(1-p)^2}$$

$$= \frac{(p + (1-p)t)(1-p-(1-p)t)}{n(1-p)^2}$$

$$= \frac{p-p^2-p(1-p)t + (1-p)t-p(1-p)t-(1-p)^2t^2}{n(1-p)^2}$$

$$= \frac{p-p^2-p(1-p)t + (1-p)t-p(1-p)t-(1-p)^2t^2 + t(1-p)^2-t(1-p)^2}{n(1-p)^2}$$

$$= \frac{p(1-p)-2p(1-p)t-(1-p)^2t}{n(1-p)^2} + \frac{t(1-t)}{n}$$

$$= \frac{(1-p)(p(1-2t)-(1-p)t)}{n(1-p)^2} + \frac{t(1-t)}{n}$$

$$= \frac{p(1-2t)-(1-p)t}{n(1-p)} + \frac{t(1-t)}{n}$$

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

Appendix 4:. Study 2A Pretest results (presented in increasing order of intrusiveness ratings)

Question (for intrusiveness rating)	Statement (for social desirability rating)	Intrusive *M*(*SD*) 1 = not at all 2 = somewhat 3 = intrusive 4 = very	Taboo *M*(*SD*) 1 = not at all 2 = somewhat 3 = taboo 4 = very
Have you ever visited an internet dating website?	Visiting an internet dating website.	1.9 (0.8)	1.4 (0.6)
Have you ever had a crush on a co-worker?	Having a crush on a coworker.	2.0 (0.9)	1.6 (0.8)
Have you ever made out with someone in public?	Making out with someone in public.	2.0 (0.9)	1.7 (0.8)
Have you ever picked up someone at a bar?	Picking someone up at a bar.	2.1 (0.9)	1.6 (0.8)
Have you ever smoked marijuana?	Smoking marijuana.	2.1 (0.9)	1.8 (0.8)
Have you ever flatulated audibly in public?	Flatulating audibly in public.	2.1 (1.0)	2.2 (0.9)
Do you have athlete's foot?	Having athlete's foot.	2.2 (0.9)	1.7 (0.8)
Have you ever lied to a teacher in order to avoid taking an exam or handing in a term paper on time?	Lying to a teacher in order to avoid taking an exam or handing in a term paper on time.	2.2 (0.9)	2.1 (0.9)
Have you ever vomited on someone?	Vomiting on someone else.	2.2 (1.1)	2.7 (1.0)
What is your sexual orientation?	Having a preferred sexual orientation.	2.2 (0.9)	1.5 (0.9)
Have you ever been constipated?	Being constipated.	2.3 (0.9)	1.4 (0.7)
Have you ever been late in paying a bill or your rent?	Being late in paying a bill or your rent.	2.3 (0.9)	1.9 (0.8)
Have you ever cheated on a term paper?	Cheating on a term paper.	2.3 (1.0)	2.5 (1.0)
Have you ever left a restaurant without paying the bill?	Leaving a restaurant without paying the bill.	2.3 (0.9)	2.9 (1.0)
Have you ever lied about your income to someone?	Lying to someone about your income.	2.3 (0.9)	1.7 (0.8)
Have you ever showered with a partner?	Showering with a partner.	2.4 (1.0)	1.3 (0.7)
Have you ever had diarrhea?	Having diarrhea.	2.4 (0.9)	1.5 (0.7)
Do you have a trust fund?	Having a trust fund.	2.4 (1.0)	1.5 (0.9)
Do you suffer from chronic bad breath?	Suffering from chronic bad breath.	2.4 (0.9)	1.8 (0.9)
Have you ever stolen anything worth more than $1?	Stealing something worth more than $1	2.4 (0.9)	2.2 (0.9)
Do you get cold sores?	Getting cold sores.	2.4 (0.9)	1.8 (0.9)
Have you ever been audited on a tax return?	Being audited on a tax return.	2.4 (0.9)	1.8 (0.8)
Have you ever looked at pornography?	Looking at pornography.	2.4 (1.0)	1.8 (0.8)
Are you currently in a relationship that involves physical intimacy?	Being in a relationship that involves physical intimacy.	2.5 (1.0)	1.4 (0.9)
Have you ever been caught having sex?	Being caught having sex.	2.5 (0.9)	2.0 (0.9)
Have you ever bounced a check?	Bouncing a check.	2.5 (0.8)	2.2 (1.0)
Have you ever had your credit card rejected?	Having your credit card rejected.	2.5 (0.9)	2.0 (0.9)
Have you ever had a one-night stand?	Having a one-night stand.	2.5 (0.9)	1.9 (0.9)
Have you ever been turned down for a loan?	Being turned down for a loan.	2.5 (1.0)	1.8 (1.0)
Have you ever cheated on your tax return?	Cheating on your tax return.	2.6 (0.9)	2.5 (1.0)
Have you had a bowel movement within the past 24 h?	Having a bowel movement in the past 24 h.	2.6 (1.0)	1.3 (0.7)
Have you ever had a creditor come after you?	Having a creditor come after you.	2.6 (0.9)	1.9 (0.9)
Have you ever stolen anything worth more than $25?	Stealing something worth more than $25	2.6 (0.9)	2.7 (1.0)
Have you ever tried LSD?	Trying LSD.	2.6 (1.0)	2.7 (0.9)
Do you regularly have an unpaid credit card balance?	Regularly having an unpaid credit card balance.	2.6 (0.9)	1.6 (0.8)
Have you purchased books on sexuality or sexual techniques?	Purchasing books on sexuality or sexual techniques.	2.6 (1.0)	1.7 (0.9)
Have you ever masturbated?	Masturbating.	2.7 (1.0)	1.7 (0.9)
Have you ever sold marijuana to someone other than a friend or relative?	Selling marijuana to someone other than a friend or relative.	2.7 (0.9)	2.2 (1.0)
Have you ever tried heroin?	Trying heroin.	2.7 (1.0)	3.1 (1.0)
Have you ever engaged in oral sex?	Engaging in oral sex.	2.7 (1.0)	1.5 (0.8)
Have you ever had unprotected sex?	Having unprotected sex.	2.7 (0.9)	2.2 (1.0)
Do you have over $20,000 in savings?	Having over $20,000 in savings.	2.7 (1.0)	1.4 (0.8)
Do you have credit card debt above $10,000?	Having credit card debt above $10,000.	2.8 (1.0)	2.0 (1.0)
Have you ever tried cocaine?	Trying cocaine.	2.8 (0.9)	2.7 (0.9)
Have you ever participated in phone sex?	Participating in phone sex.	2.8 (0.9)	1.8 (0.7)
Have you ever taken nude pictures of yourself or your partner?	Taking nude pictures of yourself or your partner	2.8 (1.0)	1.9 (0.9)
Have you ever cross-dressed?	Cross-dressing.	2.8 (0.9)	2.4 (0.9)
Have you ever stolen anything worth more than $100?	Stealing something worth more than $100.	2.8 (0.9)	2.9 (1.0)

Have you ever had sexual contact with a member of the same sex?	Having sexual contact with a member of the same sex.	2.8 (1.0)	1.9 (0.9)
Do you have any kind of communicable disease (e.g. tuberculosis)	Having some kind of communicable disease (e.g. tuberculosis).	2.8 (0.9)	2.0 (0.9)
Have you ever declared bankruptcy?	Declaring bankruptcy.	2.8 (1.0)	2.1 (0.9)
Are you a virgin?	Being a virgin.	2.8 (0.9)	1.7 (0.9)
Have you ever had sexual thoughts about a member of your same sex?	Having sexual thoughts about a member of your same sex.	2.9 (1.0)	1.8 (0.8)
Have you ever been caught masturbating?	Being caught masturbating.	2.9 (0.9)	2.4 (0.8)
Do you maintain (e.g. shave, wax, trim) your pubic hair?	Maintaining (e.g. shaving, waxing, trimming) your pubic hair	2.9 (0.9)	1.4 (0.8)
Have you ever felt a desire to expose yourself in public?	Feeling a desire to expose yourself in public.	2.9 (1.0)	3.1 (0.9)
Have you ever engaged in an extra-partner sexual relationship?	Engaging in an extra-partner sexual relationship.	2.9 (0.9)	2.7 (1.0)
Did you reach orgasm the last time you had sex?	Reaching orgasm during sex.	3.0 (0.9)	1.3 (0.7)
Have you ever used sex toys or masturbatory aids?	Using sex toys or masturbatory aids.	3.0 (0.9)	1.8 (0.9)
Have you ever participated in group sex?	Participating in group sex.	3.0 (0.9)	2.4 (1.0)
Have you ever had sexual desires for a minor?	Having sexual desires for a minor.	3.1 (1.0)	3.5 (1.0)
Have you ever had anal sex?	Having anal sex.	3.1 (0.9)	2.2 (1.0)
Have you ever been paid to have sex?	Being paid to have sex.	3.1 (0.9)	2.9 (1.0)
Have you ever paid for sex?	Paying for sex.	3.2 (0.9)	2.8 (0.9)
Do you enjoy violent sex?	Enjoying violent sex.	3.2 (0.9)	2.8 (0.9)
Have you ever been diagnosed with a sexually transmitted disease (STD)?	Being diagnosed with an STD (sexually transmitted disease).	3.3 (0.9)	2.7 (1.0)

Appendix 5:. Study 2A selected items (each pair is equivalent in intrusiveness and different in social desirability)

Study 2a Selected items

Question Set		*M*(*SD*) Intrusive 1 = not at all 2 = somewhat 3 = intrusive 4 = very	*t*-test within question set	*M*(*SD*) Taboo 1 = not at all 2 = somewhat 3 = taboo 4 = very	*t*-test within question set
1	Have you ever picked up someone at a bar?	2.1 (0.9)	t(79) = −0.22,	1.6 (0.8)	t(79) = −4.89,
	Have you ever flatulated audibly in public?	2.1 (1.0)	p = .83	2.2 (0.9)	p < .0001
2	Have you ever showered with a partner?	2.4 (1.0)	t(79) = 0.59,	1.3 (0.7)	t(79) = −5.56,
	Have you ever been late in paying a bill or your rent?	2.3 (0.9)	*p* = .56	1.9 (0.8)	*p* < .0001
3	Are you currently in a relationship that involves physical intimacy?	2.5 (1.0)	t(79) = −0.20, p = .84	1.4 (0.9)	t(79) = −6.29, p < .0001
	Have you ever bounced a check?	2.5 (0.8)		2.2 (1.0)	
4	Do you maintain (e.g. shave, wax, trim) your pubic hair?	2.9 (0.9)	t(79) = 0.51, *p* = .61	1.4 (0.8)	t(79) = −3.62, *p* = .001
	Have you ever had sexual thoughts about a member of your same sex?	2.9 (1.0)		1.8 (0.8)	
5	Did you reach orgasm the last time you had sex?	3.0 (0.9)	t(79) = −0.90,	1.3 (0.7)	t(79) = −12.52,
	Have you ever had sexual desires for a minor?	3.1 (1.0)	p = .37	3.5 (1.0)	p < .0001

Appendix 6:. Calculations of the percentage of correct responses by guilt status and condition, studies 3 and 5B.

Respondent type	Condition	Percentage of correct responses
Innocent	Direct inquiry	100-denial rate
	RRT-tell-the-truth	100-denial rate
	RRT-respond-yes	100-admit rate
Guilty	Direct inquiry	100-admit rate
	RRT-tell-the-truth	100-admit rate
	RRT-respond-yes	100-admit rate

Appendix 7:. Other privacy-protecting techniques that may be less prone to non-adherence

A technique that seems less susceptible to protecting the wrong (i.e., non-stigmatizing) response option is the item count technique (ICT, also called the unmatched count technique, unmatched block design, or block total response; Miller, 1984; Raghavarao & Federer, 1979). The ICT is a technique that, like RRTs, introduces noise into the communication channel. Respondents are presented with a list of behaviors – one focal, sensitive, behavior amid other benign behaviors – and report how many they have engaged in. By comparing the resulting counts to those of a control group whose list contained only the benign behaviors, the prevalence of the sensitive behavior can be estimated. By design, this set-up seems less prone to the kind of non-adherence that undermines RRT effectiveness, as respondents simply report the number of behaviors they have engaged in. The ICT may hence lessen concerns of response misinterpretation because it does not force responses in any way, unlike RRTs. The ICT may also fare better in preventing other forms of non-adherence, including unintentional non-adherence, by virtue of its simplicity relative to the complex instructions required of the RRT.

Disappointingly though, a meta-analysis of studies comparing ICT to DQ (Tourangeau and Yan, 2007) found that although the ICT generally provides higher prevalence estimates of socially undesirable behaviors, the overall effect was not significant. A recent study investigating anti-gay sentiment in the US using the ICT (Coffman, Coffman, & Ericson, 2017) found that the ICT produced higher prevalence estimates than DQ, however, it also found that the ICT worked better when the sensitive answer was “no” as compared to “yes,” suggesting that concerns about self-incrimination may still play a role in the ICT. The only way to determine to what extent the ICT is instrumental in alleviating intentional and unintentional non-adherence would be to conduct validation studies that compare prevalence estimates gleaned through ICT with actual prevalence rates. Unfortunately, to our knowledge, no one has yet conducted such a study. And, any gains in alleviating non-adherence should be judged against its even greater sampling variance relative to common forms of RRT (Coutts & Jann, 2011).

Similarly, another indirect inquiry method, called the crosswise-model (Yu, et al., 2007), introduces more variance into responses than common forms of RRT, although by design it may lessen concerns over response misinterpretation relative to the RRT. In the crosswise-model, respondents are presented with two questions, each inquiring whether the respondent has engaged in a certain behavior. One of the questions is about a sensitive behavior; the other is about a non-sensitive behavior. Participants are not asked to report whether they have done the behaviors, but rather, to simply report whether their answers to the two questions would be the same (i.e., both “yes,” or both “no”) or different (i.e., one “yes,” and one “no”). Thus, this technique may reduce concerns of response misinterpretation by downplaying the act of admission, perhaps even obscuring the fact that prevalence estimates can be gleaned.

Consistent with this idea, there is some evidence that the cross-wise model produces higher prevalence estimates relative to both DQ and the RRT (Hoglinger et al., 2014; Jann, Jerke, & Krumpal, 2012; Shamsipour et al., 2014). However as with the ICT, validation studies are necessary to make definitive conclusions about the capacity for the crosswise-model to produce accurate prevalence estimates. For the crosswise-model this seems particularly important, given that with this method random responding (i.e., selecting a response option at random, as can happen when people are confused) biases prevalence estimates towards 50%. Given that sensitive behaviors tend to be uncommon (i.e., prevalence rates far lower than 50%), random responding could artifactually increase the prevalence estimates produced by this model. Moreover, to the extent that this procedure requires obfuscation of its true raison d’être – to estimate the prevalence of sensitive behaviors – it may raise ethical issues.

References

Acquisti, A., John, L. K., & Loewenstein, G. (2012). The impact of relative standards on the propensity to disclose. *Journal of Marketing Research, 49*, 160–174. https://doi.org/10.1509/jmr.09.0215.

Adjerid, I., Acquisti, A., & Loewenstein, G. (in press). Choice architecture, framing, and cascaded privacy choices. Management Science.

Adler, N. E., David, H. P., Major, B. N., Roth, S. H., Russo, N. F., & Wyatt, G. E. (1992). Psychological factors in abortion: A review. *American Psychologist, 47*(10), 1194–1204. https://doi.org/10.1037/0003-066X.47.10.1194.

Akerlof, G. A., & Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics, 115*(3), 715–753. https://doi.org/10.1162/003355300554881.

Akers, R. L., Massey, J., Clarke, W., & Lauer, R. M. (1983). Are self-reports of adolescent deviance valid? Biochemical measures, randomized response, and the bogus pipeline in smoking behavior. *Social Forces, 62*, 234–251. https://doi.org/10.2307/2578357.

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review, 99*, 544–555. https://doi.org/10.1257/aer.99.1.544.

Barth, J. T., & Sandler, H. M. (1976). Evaluation of the randomized response technique in a drinking survey. *Journal of Studies on Alcohol, 37*, 690–693.

Bégin, G., & Boivin, M. (1980). Comparison of data gathered on sensitive questions via direct questioning, randomized response technique, and a projective method. *Psychological Reports, 47*(3), 743–750.

Beldt, S. F., Daniel, W. W., & Garcha, B. S. (1982). The takahasi-sakasegawa randomized response technique. *Sociological Methods & Research, 11*, 101–111. https://doi.org/10.1177/0049124182011001006.

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), Advances in Experimental Social Psychology (vol. 6, pp. 1–62): Academic Press.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review, 96*, 1652–1678. https://doi.org/10.1257/aer.96.5.1652.

Blume, A., Lai, E., & Wooyoung , L. (2013). Eliciting private information with noise: the case of randomized response. Institute of Mathematical Economics Working paper No. 490.

Böckenholt, U., & Van der Heijden, P. G. (2007). Item randomized-response models for measuring noncompliance: Risk-return perceptions, social influences, and self-protective responses. *Psychometrika, 72*, 245–262. https://doi.org/10.1007/s11336-005-1495-y.

Boruch, R. F. (1971). Assuring confidentiality of responses in social research: A note on strategies. *The American Sociologist, 6*, 308–311.

Bourke, P. D. (1984). Estimation of proportions using symmetric randomized response designs. *Psychological Bulletin, 96*(1), 166–172.

Brewer, K. R. (1981). Estimating marijuana usage using randomized response—some paradoxical findings. *Australian Journal of Statistics, 23*(2), 139–148. https://doi.org/10.1111/j.1467-842X.1981.tb00771.x.

Buchman, T. A., & Tracy, J. A. (1982). Obtaining responses to sensitive questions: Conventional questionnaire versus randomized response technique. *Journal of Accounting Research, 20*, 263–271. https://doi.org/10.2307/2490775.

Campbell, A. A. (1987). Randomized response technique. *Science, 236*(4805), 1049. https://doi.org/10.1126/science.3576215.

Chen, X., Du, Q., Jin, Z., Xu, T., Shi, J., & Gao, G. (2014). The randomized response technique application in the survey of homosexual commercial sex among men in Beijing. *Iranian Journal of Public Health, 43*(4), 416–422.

Cialdini, R. (1966). Social influence and the triple tumor structure of organizational dishonesty. In D. Messick, & A. Tenbrunsel (Eds.). *Codes of Conduct: Behavioral research into business ethics* (pp. 44–59). New York: Russell Sage Foundation.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58*(6), 1015–1026.

Clark, S. J., & Desharnais, R. A. (1998). Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychological Methods, 3*, 160–168. https://doi.org/10.1037/1082-989X.3.2.160.

Coffman, K. B., Coffman, L. C., & Ericson, K. M. M. (2017). The size of the LGBT population and the magnitude of antigay sentiment are substantially underestimated. *Management Science, 63*, 3168–3186. https://doi.org/10.1287/mnsc.2016.2503.

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*, 169–193. https://doi.org/10.1177/0049124110390768.

Cruyff, M., Bockenholt, U., Van den Hout, A., & Van der Heijden, P. G. (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*, 316–331.

Cruyff, M., Van den Hout, A., Van der Heijden, P. G., & Böckenholt, U. (2007). Log-linear randomized-response models taking self-protective response behavior into account. *Sociological Methods & Research, 36*, 266–282. https://doi.org/10.1177/0049124107301944.

de Jong, M. G., Pieters, R., & Fox, J. (2010). Reducing social desirability bias through item randomized response: An application to measure underreported desires. *Journal of Marketing Research, 47*, 14–27. https://doi.org/10.1509/jmkr.47.1.14.

de Jong, M. G., Pieters, R., & Stremersch, S. (2012). Analysis of sensitive questions across cultures: An application of multigroup item randomized response theory to sexual attitudes and behavior. *Journal of Personality and Social Psychology, 103*, 543–564. https://doi.org/10.1037/a0029394.

Dervan, L. E., & Edkins, V. A. (2013). The innocent defendant's dilemma: An innovative empirical study of please bargaining's innocence problem. *The Journal of Criminal Law and Criminology, 103*, 1–48.

Dhar, R., & Wertenbroch, K. (2012). Self-signaling and the costs and benefits of temptation in consumer choice. *Journal of Marketing Research, 49*, 15–25. https://doi.org/10.1509/jmr.10.0490.

Duffy, J. C., & Waterton, J. J. (1988). Randomised response vs direct questioning: Estimating the prevalence of alcohol-related problems in a field survey. *Australian Journal of Statistics, 30*, 1–14. https://doi.org/10.1111/j.1467-842X.1988.tb00607.x.

Edgell, S. E., Himmelfarb, S., & Duchan, K. L. (1982). Validity of forced responses in a randomized response model. *Sociological Methods & Research, 11*, 89.

Elffers, H., Van der Heijden, P., & Hezemans, M. (2003). Explaining regulatory noncompliance: A survey study of rule transgression for two dutch instrumental laws, applying the randomized response method. *Journal of Quantitative Criminology, 19*, 409–439. https://doi.org/10.1023/B:JOQC.0000005442.96987.9e.

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature, 422*, 137. https://doi.org/10.1038/nature01474.

Ferraro, F., Pfeffer, J., & Sutton, R. I. (2005). Economics language and assumptions: How theories can become self-fulfilling. *Academy of Management Review, 30*, 8–24. https://doi.org/10.5465/amr.2005.15281412.

Forges, F. (1986). An approach to communication equilibria. *Econometrica, 54*(6), 1375–1385. https://doi.org/10.2307/1914304.

Fox, J. P., Avetisyan, M., & Van der Palen, J. (2013). Mixture randomized item-response modeling: A smoking behavior validation study. *Statistics in Medicine, 32*(27), 4821–4837. https://doi.org/10.1002/sim.5859.

Greenberg, B. G., Abul-Ela, A., Simmons, W. R., & Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association, 64*(326), 520–539. https://doi.org/10.2307/2283636.

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

Gneezy, A., Gneezy, U., Riener, G., & Nelson, L. D. (2012). Pay-what-you-want, identity, and self-signaling in markets. *Proceedings of the National Academy of Sciences, 109*(19), 7236–7240. https://doi.org/10.1073/pnas.1120893109.

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research, 35*(3), 472–482.

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

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

Hoglinger, 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 Papers.

Horvitz, D. G., Shah, B. V., & Simmons, W. R. (1967). The unrelated question randomized response model. *Social Statistics Section Proceedings of the American Statistical Association, 63* 754-754.

Houston, J., & Tran, A. (2001). A survey of tax evasion using the randomized response technique. *Advances in Taxation, 13*, 69–94. https://doi.org/10.1016/s1058-7497(01)13007-3.

Insight Central (2010). Randomized responses: More indirect techniques to asking sensitive survey questions. Retrieved September 21, 2016, from https://analysights.wordpress.com/2010/06/23/randomized-responses-more-indirect-techniques-to-asking-sensitive-survey-questions.

Jann, B., Jerke, J., & Krumpal, I. (2012). Asking sensitive questions using the crosswise model: An experimental survey measuring plagiarism. *Public Opinion Quarterly, 76*, 32–49. https://doi.org/10.1093/poq/nfr036.

John, L.K. (2015). The consumer psychology of online privacy: Insights and opportunities from behavioral decision theory. In: M. Norton, D. Rucker, C. Lamberton (Eds.), Cambridge Handbook of Consumer Psychology.

John, L. K., Acquisti, A., & Loewenstein, G. (2011). Strangers on a plane: Context-dependent willingness to divulge sensitive information. *Journal of Consumer Research, 37*(5), 858–873. https://doi.org/10.1086/656423.

Joinson, A. N., 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. https://doi.org/10.1016/j.chb.2007.10.005.

Jones, E. E. (1986). Interpreting interpersonal behavior: The effects of expectancies. *Science, 234*(4772), 41–46. https://doi.org/10.1126/science.234.4772.41.

Kim, T., Barasz, K., & John, L. K. (2018). Why am I seeing this ad? The effect of ad transparency on ad effectiveness. *Journal of Consumer Research*. https://doi.org/10.1093/jcr/ucy039.

Kipnis, D. (1972). Does power corrupt? *Journal of Personality and Social Psychology, 24*, 33–41. https://doi.org/10.1037/h0033390.

Kirchner, A. (2015). Validating sensitive questions: A comparison of survey and register data. *Journal of Official Statistics, 31*, 31–59. https://doi.org/10.1515/JOS-2015-0002.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and Web Surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly, 72*(5), 847–865.

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*(6), 1387–1403. https://doi.org/10.1016/j.ssresearch.2012.05.015.

Kulka, R. A., Weeks, M. F., & Folsom, R. E. (1981). *A comparison of the randomized response approach and the direct question approach to asking sensitive survey questions.* NC: Research Triangle Institute.

Kruglanski, A. W. (1970). Attributing trustworthiness in supervisor-worker relations. *Journal of Experimental Social Psychology, 6*(2), 214–232. https://doi.org/10.1016/0022-1031(70)90088-0.

Lamb, C. W., & Stem, D. E. (1978). An empirical validation of the randomized response technique. *Journal of Marketing Research, 15*(4), 616–621. https://doi.org/10.2307/3150633.

Lara, D., Strickler, J., Olavarrieta, C. D., & Ellertson, C. (2004). Measuring induced abortion in Mexico: A comparison of four methodologies. *Sociological Methods & Research, 32*, 529–558. https://doi.org/10.1177/0049124103262685.

Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological Bulletin, 107*, 34–47. https://doi.org/10.1037/0033-2909.107.1.34.

Lensvelt-Mulders, G., Hox, J. J., Van der Heijden, P. G., & Maas, C. J. (2005). Meta-analysis of randomized response research: Thirty-five years of validation. *Sociological Methods & Research, 33*(3), 319–348. https://doi.org/10.1177/0049124104268664.

Lingle, J. H., Brock, T. C., & Cialdini, R. B. (1977). Surveillance instigates entrapment when violations are observed, when personal involvement is high, and when sanctions are severe. *Journal of Personality and Social Psychology, 35*(6), 419–429. https://doi.org/10.1037/0022-3514.35.6.419.

Locander, W., Sudman, S., & Bradburn, N. (1976). An investigation of interview method, threat and response distortion. *Journal of the American Statistical Association, 71*(354), 269–275.

MacGregor, D. (1960). The human side of enterprise. In M. J. Handel (Ed.). *The Sociology of Organizations: Classic, Contemporary, and Critical Readings* (pp. 108–113). New York: SAGE Publications.

Marquis, K. H., Marquis, M. S., & Polich, J. M. (1986). Response bias and reliability in sensitive topic surveys. *Journal of the American Statistical Association, 81*(394), 381–389. https://doi.org/10.1080/01621459.1986.10478282.

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. https://doi.org/10.1509/jmkr.45.6.633.

Miller, J. D. (1984). *A new survey technique for studying deviant behavior.* Doctoral dissertationGeorge Washington University.

Musch, J., Bröder, A., & Klauer, K. C. (2001). Improving survey research on the World-Wide Web using the randomized response technique. *Dimensions of Internet Science,* 179–192.

Myerson, R. B. (1986). Multistage games with communication. *Econometrica, 54*(2), 323–358.

Ostapczuk, 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*, 267–287. https://doi.org/10.3102/1076998609332747.

Ostapczuk, 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, 20*(5), 489–503.

Prelec, D., & Bodner, R. (2003). Self-signaling and self-control. In G. Loewenstein, D. Read, & R. F. Baumeister (Eds.). *Time and Decision: Economic and Psychological Perspectives of Intertemporal Choice* (pp. 277–301). New York: The Russell Sage Foundation.

Raghavarao, D., & Federer, W. T. (1979). Block total response as an alternative to the randomized response method in surveys. *Journal of the Royal Statistical Society, 41*, 40–45.

Rider, R. V., Harper, P. A., Chow, L. P., & Cheng, I. C. (1976). A comparison of four methods for determining prevalence of induced abortion, Taiwan, 1970–1971. *The National Center for Biotechnology Information, 103*, 37–50.

Rohan, T. (2013). *Antidoping agency delays publication of research.* New York Times.

Rosenfeld, B., Imai, K., & Shapiro, J. N. (2015). An empirical validation study of popular survey methodologies for sensitive questions. *American Journal of Political Science, 60*(3), 783–802. https://doi.org/10.1111/ajps.12205.

Sánchez-Pagés, S., & Vorsatz, M. (2007). An experimental study of truth-telling in a sender-receiver game. *Games and Economic Behaviour, 61*, 86–112. https://doi.org/10.1016/j.geb.2006.10.014.

Schweitzer, M. E., Hershey, J. C., & Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes, 101*, 1–19. https://doi.org/10.1016/j.obhdp.2006.05.005.

Shamsipour, M., Yunesian, M., Fotouhi, A., Jann, B., Rahimi-Movaghar, A., Asghari, F., & Akhlaghi, A. A. (2014). Estimating the prevalence of illicit drug use among students using the crosswise model. *Substance Use & Misuse, 49*(10), 1303–1310. https://doi.org/10.3109/10826084.2014.897730.

Shotland, R. L., & Lynn, D. Y. (1982). The random response method: A valid and ethical indicator of the "truth" in reactive situations. *Personality and Social Psychology Bulletin, 8*, 174–179. https://doi.org/10.1177/014616728281027.

Simmons, J. (2014). MTurk vs. the lab: Either way we need big samples. Data Colada. Retrieved from http://datacolada.org/18.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632.

Singer, E., Hippler, H. J., & Schwarz, N. (1992). Confidentiality assurances in surveys: Reassurance or threat? *International Journal of Public Opinion Research, 4*, 256–268. https://doi.org/10.1093/ijpor/4.3.256.

Strickland, L. H. (1958). Surveillance and trust. *Journal of Personality, 26*, 200–215. https://doi.org/10.1111/j.1467-6494.1958.tb01580.x.

Tan, T., Kreuter, F., & Tourangeau, R. (2012). Evaluating survey questions: A comparison

of methods. *Journal of Official Statistics, 28*(4), 503–529.
Tamhane, A. C. (1981). Randomized response techniques for multiple sensitive attributes. *Journal of the American Statistical Association, 76*(376), 916–923.
Tamir, D. I., & Mitchell, J. P. (2012). Disclosing information about the self is intrinsically rewarding. *Proceedings of the National Academy of Sciences, 109*(21), 8038–8043. https://doi.org/10.1073/pnas.1202129109.
Tenbrunsel, A. E., & Messick, D. M. (1999). Sanctioning systems, decision frames, and cooperation. *Administrative Science Quarterly, 44*(4), 684–707. https://doi.org/10.2307/2667052.
Tor, A., Gazal-Ayal, O., & Garcia, S. M. (2010). Fairness and the willingness to accept plea bargain offers. *Journal of Empirical Legal Studies, 7*, 97–116. https://doi.org/10.1111/j.1740-1461.2009.01171.x.
Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859–883.
Tracy, P. E., & Fox, J. A. (1981). The validity of randomized response for sensitive measurements. *American Sociological Review,* 187–200.
Umesh, U. N., & Peterson, R. A. (1991). A critical evaluation of the randomized response method. *Sociological Methods & Research, 20*(1), 104–138. https://doi.org/10.1177/0049124191020001004.
Van der Heijden, P., Van Gils, G., Bouts, J., & Hox, J. J. (2000). A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning. *Sociological Methods & Research, 28*, 505–537. https://doi.org/10.1177/0049124100028004005.
Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60*, 63–66.
Warren, R. P. (Writer) (1946). All the king's men New York: Time Incorporated.
Weissman, A. N., Steer, R. A., & Lipton, D. S. (1986). Estimating illicit drug use through telephone interviews and the randomized response technique. *Drug and Alcohol Dependence, 18*(3), 225–233. https://doi.org/10.1016/0376-8716(86)90054-2.
Williams, B. L., & Suen, H. (1994). A methodological comparison of survey techniques in obtaining self-reports of condom-related behaviours. *Psychological Reports, 75*(3, Pt 2), 1531–1537. https://doi.org/10.2466/pr0.1994.75.3f.1531.
Wiseman, F., Moriarty, M., & Schafer, M. (1975). Estimating public opinion with the randomized response model. *Public Opinion Quarterly, 39*(4), 507–513. https://doi.org/10.1086/268247.
Wolter, F., & Preisendörfer, P. (2013). Asking sensitive questions: An evaluation of the randomized response technique versus direct questioning using individual validation data. *Sociological Methods & Research, 42*, 321–353. https://doi.org/10.1177/0049124113500474.
Yu, J.-W., Tian, G.-L., & Tang, M.-L. (2007). Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika, 67*, 251. https://doi.org/10.1007/s00184-007-0131-x.
Zdep, S. M., Rhodes, I. N., Schwarz, R. M., & Kilkenny, M. J. (1979). The validity of the randomized response technique. *Public Opinion Quarterly, 43*, 544–549.