Gaming the System: Inaccurate Responses to Randomized Response Technique Items

Ashley Richards & Elizabeth Dean, RTI International

Send all correspondence to:
Ashley Richards
RTI International
3040 East Cornwallis Road
Post Office Box 12194
Research Triangle Park, NC 27709-2194
Voice: (919) 541-8050
Fax: (919) 541-6604
E-mail: ashrichards@rti.org

ABSTRACT

The Randomized Response Technique (RRT) is used to encourage accurate responding to sensitive survey questions. When using the RRT, respondents are given two questions (one sensitive and the other nonsensitive with a known response distribution) and are instructed to answer one of them. The question to be answered is determined by the outcome of a random act with a known probability (e.g. a coin toss), that only the respondent sees. Researchers do not know which question each respondent answered, but are able to calculate proportions for each response to the sensitive question.

Though it is designed to reduce error, the RRT may actually increase measurement error if respondents implement it incorrectly. Evaluating the RRT is challenging because the outcome of its driving feature, the randomizer, is concealed from researchers. As a result, prior research has typically assumed that higher reporting of undesirable responses signals the RRT's success.

Eight RRT items were evaluated in a non-probability survey of 75 participants of the online virtual world, Second Life (SL). Participants were randomly assigned to one of three modes: face-to-face interview in SL, voice chat interview in SL, or web. The randomizer across all modes was an interactive, 3-dimensional virtual coin toss that was discreetly manipulated by the researchers in order to determine with near certainty whether participants followed the procedure.

Only 67% of participants followed the procedure for every RRT item. The greatest rate of procedural noncompliance on an item was 13%. There were no significant differences in RRT compliance by demographic characteristics or survey mode. Most participants indicated in debriefing questions that they enjoyed this method of answering questions, but their noncompliance is cause for additional skepticism about using the RRT.

INTRODUCTION

Sensitive questions tend to be a source of measurement error and nonresponse error in surveys because some respondents refuse to answer these questions or answer them inaccurately. Researchers have taken many different approaches to reducing erroneous responses to sensitive questions, focusing mainly on question wording, mode of administration, and confidentiality assurances.

One of the more novel methods of reducing error when measuring sensitive topics is the Randomized Response Technique (RRT) (Warner 1965). The RRT is intended to reduce measurement error by allowing respondents to reveal less information when answering sensitive questions. Respondents are given two questions (one sensitive, the other nonsensitive with a known response distribution) and instructed to answer only one of the questions. The question to be answered is determined by the outcome of a random act with a known probability (e.g., a coin toss) that only the respondent sees. Researchers do not know which question each respondent answers, but are able to calculate proportions for each response to the sensitive question.

The RRT's element of randomization, typically in the form of a coin toss, introduces a game-like component to the survey administration process. Survey gamification advocates argue that surveys quality and respondent attentiveness is improved when surveys are interesting, fun and interactive enough to attract respondents' attention to the task at hand. Gamification has three components:

1) The survey task is presented as a game or challenge, for example, "Describe the last meal you ate in 10 words or less."

2) Competition is invoked against other players or against the clock, such as, "Who can provide 5 brands of cigarettes the fastest?"

3) Rewards & feedback are provided, in the form of increased status on a respondent panel, badges, or instantaneously giving reports from respondents' reports back to them. An example of this kind of interactivity would be providing "How do you compare?" summary statistics to show the respondent's 30-day alcohol vs. that of their peers (Puleston 2012, Puleston 2011)

The RRT implements the first element of gamification. The requirement to flip a coin adds a surprising and interactive survey step that is dependent on chance. As in a game, there are factors outside the control of the players (the respondent and interviewer). While it's not true gamification in the sense of adding competition or making feedback or rewards available to the respondent, the RRT could be more enjoyable for respondents, despite the additional steps involved in answering question. It also has the potential to keep respondents engaged because 1) they have to flip the coin to come up with their answer and 2) it's deliberately a more challenging task than a traditional survey question.

With the introduction of more challenging tasks comes the risk of increasing measurement error. Although it is designed to reduce error, the RRT may actually *increase* measurement error if respondents implement it incorrectly. Coutts and Jann (2011), among others, have suggested that respondents do not comply with the RRT because they do not understand how to follow the procedure or because they do not understand how it protects their anonymity. Respondents may also not comply with the RRT because they do not want to appear to have endorsed a particular response.

Evaluating the true effectiveness of the RRT is difficult because researchers are blind to the outcome of the randomizing device. Even if the outcome *is* known, it is impossible to observe the cognitive processes occurring in respondents' minds, as Holbrook and Krosnick

point out (2010). Because it is impossible to confirm that the procedure was followed, researchers have typically assumed that higher reporting of undesirable attributes signal the RRT's success. Holbrook and Krosnick argue that "this calls into question interpretations of all past RRT studies and raises serious questions about whether the RRT has practical value for increasing survey reporting accuracy."

This research addresses the need, pointed out by Holbrook and Krosnick, to determine "whether the RRT does, in fact, lead to more accurate reporting in any mode or whether it simply increases rates of all reported attributes due to inevitable implementation errors" (2010). We studied the accuracy of respondent reporting by manipulating the randomizer so we knew which question respondents should answer, and by asking questions that both we and the respondents should be able to answer correctly. We first investigated whether respondents are capable of adhering to RRT procedures by pairing questions that are both innocuous. We then paired sensitive and innocuous questions to examine the influence of social desirability on accuracy of responses. Several items directed respondents to answer an innocuous question in a manner that would be undesirable if that response were given for the sensitive question in the pair. In comparison, other items had an answer that would be desirable if given for the sensitive question.

BACKGROUND

Despite, or perhaps because of its seeming success at increasing reporting of many different types of sensitive behaviors (see Holbrook and Krosnick (2010) for a nearly exhaustive list of topics to which the RRT has been applied), researchers have scrutinized the RRT extensively. They have taken three primary approaches to evaluating the validity of RRT

methods: evaluation by "more reporting is better" assumption, comparison to external benchmarks, and validation against respondent data (Lensvelt-Mulders et al. 2005).

The "more is better" approach relies on the assumption that if the RRT results in higher reporting of sensitive behaviors than direct questioning, then it is working, since sensitive behaviors tend to be underreported anyway (Sudman and Bradburn 1982; Tracy and Fox 1981). When applied to a wide variety of substantive areas, the RRT has been found to increase reporting of sensitive behaviors. Some examples are increased reporting of: illegal drug use (Zdep et al. 1979); abortion in rural Africa (Chow, Gruhn, and Chang 1979), Taiwan (Rider et al. 1976), and North America (Shimizu and Bonham 1978); and child abuse perpetration (Zdep and Rhodes 1977). On the other hand, one study of U.S. college students (convenience sample) found that students reported engaging in unprotected sex or socially stigmatized sexual behaviors more frequently with the direct questioning technique than with the RRT (Williams and Suen 1994).

More sophisticated evaluations of the RRT compare findings from RRT-administered questions to external datasets for benchmarks and to respondent-specific data captured in another form (such as administrative records). A 2005 meta-analysis concluded that studies that present validation against external datasets do not find a perfect correlation between an RRT answer and the externally recorded sensitive data point. That is, the RRT improves accuracy and increases reporting of sensitive items relative to direct questioning, but still does not always produce accurate or honest answers (Lensvelt-Mulders et al. 2005). RRT application for the National Survey of Family Growth resulted in married women reporting abortions at four times the national rate reported by the Centers for Disease Control and Prevention (Shimizu and Bonham 1978). Yet, Wiseman et al. compared RRT answers to sensitive questions on politics, morals, and race to answers to in-person interviews and answers to self-administered mail surveys, and

concluded that the method did not improve validity because the estimates generated from the RRT module were more similar to the estimates generated from interviewer-administered items than to mail survey (self-administered) items (Wiseman, Moriarty, and Schafer 1976).

When college students' academic records were compared to survey results, the RRT resulted in more accurate estimates of the frequency of receiving failing grades than a conventional question (Lamb and Stern 1978). In another study an experiment testing the RRT was staged so that research participants overheard characteristics of the lab test described before the test actually took place. Later, during the lab procedures, when participants were asked whether they had heard anything in advance, 64% of those queried using the RRT reported receiving the information vs. 25% who were queried during a conventional face-to-face interview (Shotland and Yankowski 1982). A third validation against respondent data found that comprehension of the RRT has a big impact on accuracy of response. In a study of people known to have committed social welfare system fraud, respondents who understood how the RRT worked reported trusting it more and also reported their fraud more honestly. Respondents who did not grasp the RRT exhibited less trust and reported less accurate answers (Landsheer, van der Heijden, and van Gils 1999).

Even if respondent-specific benchmarks are not available, other respondent data can be used to assess the RRT. The RRT method takes about 2 minutes longer to administer than a direct question and is likely to result in an item nonresponse rate between 5 and 10%, presumably because the task is confusing or somehow off-putting. Compared to the Unmatched Count Technique (UCT), in which one sample of respondents is given a question about whether they have done any of a list of nonsensitive activities and another sample is given the same question and response list with the sensitive behavior added, respondents are less likely to

understand how the RRT works and less likely to trust that the anonymity of their answers are protected. Less trust and less understanding results in lower reported estimates for sensitive behaviors (Coutts and Jann 2011). Understanding or comprehension of the RRT may be critical to its success. One study found that it may be too complicated for lower literacy populations in Taiwan to follow the instructions correctly (Rider et al. 1976). Cause for some concern is that lab testing of the RRT using cognitive interviews generated feedback from respondents such as laughter, joking, and comments that the requirement of a coin flip in the middle of a survey interview was not serious (Hubbard, Caspar, and Lessler 1989).

The cheating detection approach develops a statistical model for controlling for the tendency of respondents to the RRT questions to cheat. Cheating on the RRT is exemplified by disregarding the instructions to answer the question specified by the randomized procedure and answering the question according to personal choice. Rather than assuming that respondents always follow the rules, this approach assumes that a given probability is associated with *not* following the rules, as well as with applying a specific rule (Ostapczuk et al. 2009; Lakshmi and Raghavarao 1992; van den Hout, Böckenholt, and van der Heijden 2010; Moshagen, Musch, and Erdfelder 2011) The cheating detection mechanism indicated that up to approximately 47% of respondents disobey the RRT instructions (Ostapczuk, Musch, and Moshagen 2010).

Few studies have attempted to evaluate the RRT's success and functioning by controlling the outcome of the randomizer. However, two studies by Edgell et al. attempted such an evaluation by developing a nonrandom randomizer that they controlled and applied to RRT interviews with college students. The first study used the standard balanced response RRT that Warner developed. The second study used the unrelated question technique, a variation of the RRT that applies one sensitive question and one completely unrelated and nonsensitive question

such as “Were you born in an even numbered year?” In the first study using the standard RRT, 26% of participants inaccurately followed procedures on the homosexual behaviors question (1982). In the second study using the unrelated question RRT, only 10% of participants inaccurately followed procedures (1992).

Because our study manipulates the randomizer, it is similar to the Edgell et al. approach. However, some key differences exist. Both of Edgell et al.’s studies used college students enrolled in introductory-level psychology classes. Interviews were done face-to-face, with a graduate student as the interviewer. On the other hand, our test examines different modes and a somewhat more general population recruited through Second Life. Additionally, the literature suggests that a measure of respondent comprehension is critical to assessing the RRT, since misunderstanding of how the RRT works results in inaccurate answers. Our research examines respondents’ understanding of the technique. We conducted the same analyses for two sets of respondents: all respondents and only those respondents who demonstrated they did not understand the procedure. When we examined only those who demonstrated that they understood the RRT procedures, we saw that they did not continue to answer correctly when asked to give socially undesirable responses. Our study investigates possible causes of incorrect responses—whether they responded incorrectly intentionally or unintentionally.

METHODS

This research was conducted in the online virtual world Second Life (SL). SL provides a 3D graphical, interactive environment that represents (and often expands upon) real life. Because SL content is completely user-generated, researchers can design and manipulate environments and avatars to test social interactions with a consistency that is not possible in the real world (Dean et al. 2009; Murphy et al. 2010). In this case, the setting allowed us to discreetly control

the outcome of the randomizer, a coin toss, which is a standard RRT randomizer but is uncontrollable outside the virtual world.

The sample was a non-probability sample of SL users recruited through SL classifieds, the SL forum, craigslist, the SL Facebook page, and word of mouth in SL. Recruitment ads directed respondents to an online screening survey that asked for the respondents' demographic information in real life (not the information of their SL avatar).

In order to be eligible to participate in the study, respondents needed to be U.S. residents at least 18 years of age, and they needed to be locatable in SL using the avatar name they provided in the screening survey. To lessen the likelihood that an individual could participate in the study more than once under the guise of different avatars, we monitored geodata from the screening survey and usually ruled eligible only the first person from a specific city to complete the screener, assuming the person met the eligibility guidelines described above. We made one exception and accepted two respondents from the same large city (population greater than 500,000) because, given the size of the city and the timing of their responses, they were likely to be different respondents. As an additional precaution, we verified that the IP addresses for these two cases differed. Of the 352 respondents who completed the screening survey, 108 were eligible and were invited to participate in the study. Of the 76 respondents who replied to the invitation, all but one agreed to participate, for a total sample size of 75 and a participation rate[1] of 69%.

Selected respondents were interviewed between August 9 and October 11, 2011. Respondents were randomly assigned to one of three modes: in-person, voice, or Web. In-person and voice interviews were conducted at RTI International's virtual office in SL. For both the in-

[1] As suggested by AAPOR, the participation rate is used in lieu of a response rate because the sample was a non-probability sample (The American Association for Public Opinion Research 2011).

person and voice modes, the interviewer and respondent communicated through the voice chat feature in SL, enabling them to talk aloud to each other. Respondents were given a coin to flip during the survey. Before asking each question, the interviewer set the coin to determine whether it would result in heads or tails. Respondents could not see that the interviewer had control over the coin. In the in-person mode, the respondent sat across a table from the interviewer. The coin was on the table in front of the respondent, and a divider was placed on the table to make it appear that the coin flip was private (see Figure 1). In the voice mode, the interviewer and respondent were in different rooms and could not see each other, but they could talk to each other. This mode was meant to resemble a telephone interview. To make it even more like a phone interview, the respondent and interviewer never saw each other's avatars.

Figure 1. In-Person Interview Configuration

The Web mode differed from the in-person and voice modes in that respondents were sent a link to the survey that they completed in their Web browser. To ensure the coin was

synced properly with the Web survey, Web respondents did not flip a coin in SL but instead watched a video of a coin flipping. The video was embedded in the survey and displayed the same coin that respondents in the in-person and voice modes flipped. To minimize skepticism about whether the Web survey coin toss was truly random, the survey did not have a back button and the coin did not reflip if the page was refreshed.

In the Web mode, each pair of RRT questions was displayed on the screen. In the in-person and voice modes, respondents were guided through three game-like steps to complete the RRT survey. First, they were handed a virtual note card for each item that displayed the questions. Second, they were asked to flip the virtual coin. Third, the note card instructed them to answer, based on the coin flip result, one of two questions. Each note card contained a traditional survey question and an atypical question which either included a picture followed by a question like, "Is the triangle red?" or asked a question like, "Do you like pepperoni pizza?" At the beginning of the survey, respondents were guided through two example RRT items that explained how the survey would work. The series of RRT items are described below, with their full text in the Appendix.

- Two items paired two innocuous questions to check whether respondents would follow the procedure without the confound of social desirability bias. These items are henceforth referred to as "understanding items" because their purpose was to check the respondents' understanding of the RRT procedure. These understanding items used two types of innocuous questions. The first type was not actually a question, but rather an instruction to "Respond *Yes*" or "Respond *No*." This type of question is henceforth referred to as an "automatic response." The second type of question asked about an image on the note card or Web survey screen. For instance, one question

showed a picture of three pigs and one cow and asked, "Are there two cows?" The questions in each item were paired such that they had different answers (yes and no) so we could determine which question the respondent answered.

- One item, henceforth referred to as the "birthday item," also checked for understanding of the procedure. On this item, because of the outcome of the coin flip, respondents were directed to answer the question "Were you born in January?" The other question, "Do you like pepperoni pizza?" was innocuous. We assumed that if respondents followed the procedure, roughly 1/12 would respond "Yes" to this item.
- Four "undesirable response" items paired a sensitive question with an innocuous question. The outcome of the coin flip directed respondents to answer the innocuous question; the answer to this question would be socially undesirable as a response to the sensitive question. The innocuous questions included automatic response questions as well as questions referring to images on the note cards.
- Two "desirable response" items were constructed just as the undesirable response items, except that the answer to the innocuous question (that respondents were supposed to answer) would be socially desirable as a response to the sensitive question.
- One item paired two sensitive questions; we did not know the respondents' correct answers to either question. This item was used to conceal the manipulated design of the survey.

The 10 RRT items were followed by a series of debriefing questions to assess respondents' thoughts about the interview, including their understanding of the procedure, the ease or difficulty of determining which question to answer, and their perceptions of accuracy. At

the end of the interview, they were paid L$500 Linden Dollars (the SL currency), which is roughly equivalent to $2.

Results

Mode Effects

An analysis of variance (ANOVA) showed no significant effects of mode on overall accuracy of responding to RRT items, $F(2, 72) = 1.34$, $p = .27$, and one-sample t-tests showed no significant differences in accuracy by mode for individual RRT items. Subsequent analyses group all modes together due to the similarity in responding across modes.

Understanding and Birthday Items

The understanding and birthday items were designed to check whether respondents understood how to follow the RRT procedure. No respondents answered both of the understanding items incorrectly, but 12% answered one of the two questions incorrectly, which was significant, $t(74) = -3.18$, $p < .01$.

For the birthday question, we can assume that roughly 8.5% of respondents were born in January and should have responded yes to this item. Instead, 11% of respondents responded yes, but this rate was not significantly higher than was expected, $t(74) = 0.61$, $p = .73$.

Undesirable Response Items

The percentage of respondents answering incorrectly on the undesirable response items varied depending on the item, and ranged from 7% to 13%. One-sample t-tests showed that the percentage of incorrect responses was significant for each item. These items remained significant when the nine respondents who answered one or both of the understanding items were excluded from analysis (see Table 1).

Table 1. Respondents' Answers to Undesirable Response Items

	All respondents (n=75)		**Respondents who answered understanding items correctly (n=66)**	
Sensitive Question	**% Incorrect on Innocuous Question**	***t***	**% Incorrect on Innocuous Question**	***t***
Q4. Have you ever stolen something valuable that did not belong to you?	7 (.25)	-2.30*	5 (.21)	-1.76*
Q5. Did you vote in the presidential election hold on November 4, 2008?	7 (.25)	-2.30*	5 (.21)	-1.76*
Q8. Do you believe in God?	13 (.34)	-3.37***	12 (.33)	-2.99**
Q9. Have you ever had sex with a person you paid or who paid you for sex?	7 (.25)	-2.30*	6 (.24)	-2.05*

* = p < .05, ** = p < .01, *** = p < .001. Standard deviations appear in parentheses below percentages.

Desirable Response Items

Few respondents answered the desirable response items incorrectly. One-sample t-tests showed that the percentage of incorrect responses was *not* significant for each item (see Table 2).

Table 2. Respondents' Answers to Desirable Response Items

	All respondents (n=75)		Respondents who answered understanding items correctly (n=66)	
Sensitive Question	**% Incorrect on Innocuous Question**	**t**	**% Incorrect on Innocuous Question**	**t**
Q6. Do you think of yourself as homosexual, gay, lesbian, or bisexual?	1 (.12)	-1.00	1 (.12)	-1.00
Q10. Has a doctor or other health care professional ever told you that you had a sexually-transmitted disease, or STD? Some examples of STDs are genital herpes, genital warts, Chlamydia, gonorrhea, Human Papillomavirus (HPV), and syphilis.	3 (.16)	-1.42	3 (.17)	-1.43

Standard deviations appear in parentheses below percentages.

Debriefing Items

Nearly all respondents (97%) said they understood the procedure for selecting which question to answer. The same number reported that the process of determining which question to answer was "very easy" or "somewhat easy."

When asked how much of the time respondents thought they answered the question indicated by the coin, 84% said always, 7% said usually, 8% said sometimes, 1% said rarely, and no one said never. Three of the six respondents who said they only *sometimes* answered the specified question actually answered *all* of the items correctly. Overall, responses to this

question were indicative of overall accuracy on the sensitive RRT items.[2] Respondents who selected "always" or "usually" for this item answered an average of 95% of the RRT items correctly, compared to 83% of respondents who selected "sometimes" or "rarely." The relationship between mean RRT accuracy and response to this item was significant, $F(3, 71) = 6.14$, $p < .001$.

Next, respondents were asked how often they *purposely* answered the wrong question: 92% said never, 7% said rarely, 1% said sometimes, and no one said usually or always. Respondents who selected "never" answered an average of 95% of the sensitive RRT items correctly, compared to 78% of those who selected "rarely" or "sometimes." The relationship between mean accuracy and response to this item was significant, $F(2, 72) = 10.70$, $p < .001$.

The final debriefing question asked participants to share any thoughts they had about the survey. Two main themes emerged: 1) respondents did not understand why we would ask questions using the RRT, and 2) they enjoyed the process of flipping the coin to decide which question to answer. Regarding the second theme, ten participants described the survey as "fun" and/or "interesting." One participant explained, "I found it to be an interesting format. I routinely take online polls and surveys for a couple organizations. Yours seemed to involve me more, and require me to pay closer attention than those I'm used to." Another said, "I actually thought it was a pretty cool survey. I liked the virtual coin thing."

Discussion

One limitation of this research is the low cell counts for some statistics, particularly the desirable response items: only 1 respondent answered the sexuality question incorrectly, and

[2] Overall accuracy was calculated as the percentage of desirable response and undesirable response items answered correctly. The understanding and birthday items were not counted as "sensitive items" and were excluded from this calculation.

only 3 respondents answered the STD question incorrectly. Another limitation is the possibility of interviewer error when setting the coin to flip as heads or tails. The interviewer followed a script that specified when the coin should be set to heads or tails, so mistakes setting the coin are unlikely. Nonetheless, the research is limited by the fact that the coin introduced an opportunity for human error.

The RRT was tested in SL with a virtual coin, but its main use is in an actual real-life setting with a real randomizer. Doing this research in SL allowed us to bypass real-life barriers and to take a closer look at how people responded to RRT items. Respondents who demonstrated on the understanding items that they did not understand the procedure in SL are unlikely to understand it in real life. The research setting is more likely to have affected how respondents answered the social desirability questions, but the finding that respondents purposely responded incorrectly in SL suggests that we should remain skeptical of the RRT in *any* setting, including real life.

Debriefing items suggest that respondents really enjoyed answering RRT items. Respondent enjoyment, however, should not factor into the decision about whether to use the RRT because engaging respondents does not necessarily make them better (i.e., more accurate) respondents. Consistent with prior research, our findings suggest that some respondents do not understand how to implement the RRT: 12% of respondents incorrectly answered one of the two understanding items. Furthermore, three of six respondents who indicated they answered correctly *sometimes* actually answered the RRT items correctly *all* of the time. It appears that these respondents did not fully understand the RRT. This finding alone is not necessarily cause for concern, as improved instructions could increase understanding of the procedure.

What is troubling, however, is the rate of inaccurate responses across the undesirable response items. The number of respondents answering these questions incorrectly was significant. Conversely, a nonsignificant number of respondents answered the two desirable response items incorrectly. These findings suggest that respondents gave intentionally inaccurate responses to avoid providing socially undesirable responses. Furthermore, respondents with lower accuracy scores were significantly more likely to confirm in the debriefing that their inaccuracies were intentional.

If the problem with RRT items were simply that respondents do not understand the procedure, we could explain the procedure to respondents in more detail to increase their understanding. However, it appears that in addition to their problems understanding the RRT, respondents deliberately responded incorrectly to avoid providing undesirable responses. The extent to which deliberate misreporting can be reduced is unclear, but given previous findings of respondents' reactions to the RRT, some degree of deliberate misreporting may be a nearly certain pitfall of using the RRT. We look forward to additional research on the RRT as well as similar but alternative methods (e.g. the crosswise model). In particular, we look forward to research into the cognitive processes of respondents as they answer RRT items to identify aspects of the RRT that are particularly confusing. We also look forward to research on improving the design of this promising, but pitfall-laden method.

REFERENCES

The American Association for Public Opinion Research. 2011. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 7th edition.* AAPOR.

Chow, L. P., Walter Gruhn, and Wen Pin Chang. 1979. "Feasibility of the Randomized Response Technique in Rural Ethiopia." *American Journal of Public Health* 69:273-6.

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

Dean, Elizabeth, Sarah Cook, Michael Keating, and Joe Murphy. 2009. "Does this Avatar Make me Look Fat? Obesity and Interviewing in Second Life." *Journal of Virtual Worlds Research* 2(2).

De Ruyck, T. and E. Veris (2011). Play, interpret together, play again and create a win-win-win. Retrieved on May 2, 2012 from: http://insites.be/media/62353/10_play,%20interpret%20together,%20play%20again%20and%20create%20a%20win-win-win.pdf

Edgell, Stephen E., Karen L. Duchan, and Samuel Himmelfarb. 1992. "An Empirical Test of the Unrelated Question Randomized Response Technique." *Bulletin of the Psychonomic Society* 30:153-6.

Edgell, Stephen E., Samuel Himmelfarb, and Karen L. Duchan. 1982. "Validity of Forced Responses in a Randomized Response Model." *Sociological Methods & Research* 11:89.

Holbrook, Allyson L. and Jon A. Krosnick. 2010. "Measuring Voter Turnout by Using the Randomized Response Technique: Evidence Calling Into Question the Method's Validity." *Public Opinion Quarterly* 74:328-43.

Hubbard, Michael L., Rachel A. Caspar, and Judith T. Lessler. 1989. "Respondent Reactions to Item Count Lists and Randomized Response." In *Proceedings of the American Statistical Association, Section on Survey Research Methods* 544-8.

Lakshmi, Damaraju V. and Damaraju Raghavarao. 1992. "A Test for Detecting Untruthful Answering in Randomized Response Procedures." *Journal of Statistical Planning and Inference* 31:387-90.

Lamb, Charles W. and Donald E. Stem. 1978. "An Empirical Validation of the Randomized Response Technique." *Journal of Marketing Research* 15:616-621.

Landsheer, Johannes A., Peter Van der Heijden, and Ger van Gils. 1999. "Trust and Understanding, Two Aspects of Randomized Response." *Quality & Quantity* 33:1-12.

Lensvelt-Mulders, Gerty J. L. M., Joop J. Hox, Peter G. M. van der Heijden, and Cora J. M. Haas. 2005. "Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation." *Sociological Methods & Research* 33:319-48.

Moshagen, Morten, Jochen Musch, and Edgar Erdfelder. (2011). "A Stochastic Lie Detector." *Behavior Research Methods* 2011 Aug 20. [Epub ahead of print].

Murphy, Joe, Elizabeth Dean, Sarah Cook, and Michael Keating. 2010. "The Effect of Interviewer Image in a Virtual-world Survey." RTI Press (RR-0014-1012). http://www.rti.org/publications/rtipress.cfm

Ostapczuk, Martin, Morten Moshagen, Zengmei Zhao, and Jochen Musch. 2009. "Assessing Sensitive Attributes Using the Randomized Response Technique: Evidence for the Importance of Response Symmetry." *Journal of Educational and Behavioral Statistics* 34:267-87.

Ostapczuk, Martin, Jochen Musch, and Morten Moshagen. 2010. “Improving Self-Report Measures of Medication Non-Adherence Using a Cheating Detection Extension of the Randomised-Response-Technique.” *Statistical Methods in Medical Research* 2010 Jul 16. [Epub ahead of print].

Puleston, J. (2012) Gamification 101 - from theory to practice. Quirk's Marketing Research Media. Retrieved on May 2, 2012 from: http://www.quirks.com/articles/2012/20120126-1.aspx

Puleston, J. (2011). Game Theory – turning surveys into games. Retrieved on May 2, 2012 from http://question-science.blogspot.com/2011/02/game-theory-turning-surveys-into-games.html

Rider, Rowland V., Paul A. Harper, L. P. Chow, and Chi I-Cheng. 1976. “A Comparison of Four Methods for Determining Prevalence of Induced Abortion, Taiwan, 1970-1971.” *American Journal of Epidemiology* 103:37-50.

Shimizu, Iris M. and Gordon S. Bonham. 1978. “Randomized Response Technique in a National Survey.” *Journal of the American Statistical Association* 73:35-9.

Shotland, Lance R. and Lynne David Yankowski. 1982. “The Random Response method: A Valid and Ethical Indicator of Truth in Reactive Situations.” *Personality and Social Psychology Bulletin.* 8(1):174-179.

Sudman, Seymour and Norman M. Bradburn (1982). *Asking Questions: a Practical Guide to Questionnaire Design.* San Francisco: Jossey-Bass.

Tracy, Paul E. and James A. Fox (1981). “The Validity of Randomized Response for Sensitive Measurements.” *American Sociological Review, 46*(2), 197-200.

van den Hout, Ardo, Ulf Böckenholt, and Peter van der Heijden. 2010. "Estimating the Prevalence of Sensitive Behaviour and Cheating with a Dual Design for Direct Questioning and Randomized Response." *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 59:723-36.

Warner, Stanley. L. 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60:63-9.

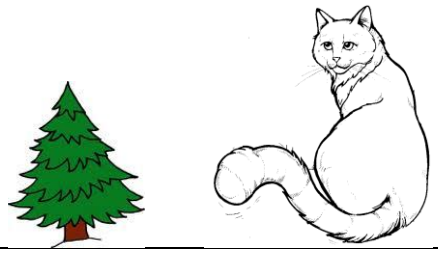
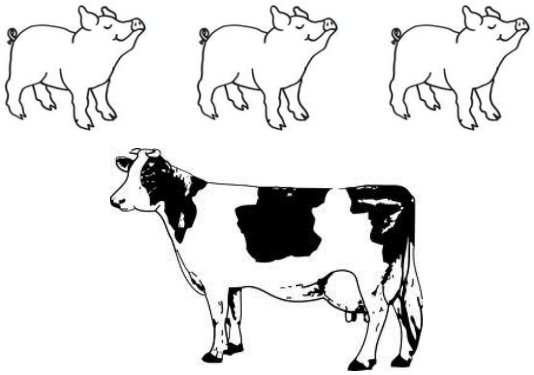
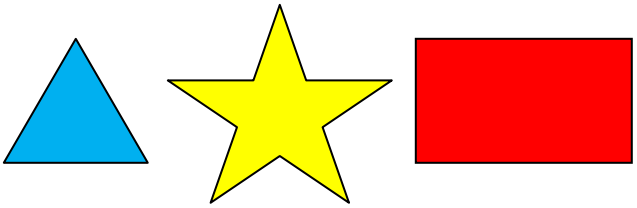
Williams, Bryan L. and Hoi Suen. 1994. "A Methodological Comparison of Survey Techniques in Obtaining Self-Reports of Condom-Related Behaviors." *Psychological Reports* 75:1531-7.

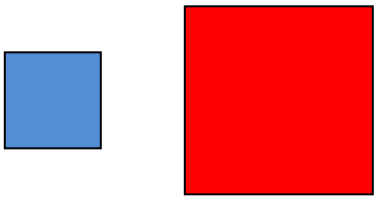
Wiseman, Frederick, Mark Moriarty, and Marianne Schafer. 1975. "Estimating Public Opinion with the Randomized Response Model." *Public Opinion Quarterly* 39:507-13.

Zdep, Stanley M. and Isabelle N. Rhodes. 1977. "Making Randomized Response Technique Work." *Public Opinion Quarterly* 40:531-7.

Zdep, Stanley M., Isabelle N. Rhodes, R. M. Schwarz, and Mary J. Kilkenny. 1979. "The Validity of the Randomized Response Technique." *Public Opinion Quarterly* 43:544-9.

Appendix: RRT Items

	Coin Toss	Notecard	Question Type
1.	Tails	Heads: Does the cat have orange fur? Tails: Is the cat larger than the tree? 	Understanding Item
2.	Heads	Heads: Respond "Yes" Tails: Are there two cows? 	Understanding Item
3.	Tails	Heads: Do you like pepperoni pizza? Tails: Were you born in January?	Birthday Item
4.	Heads	Heads: Respond "Yes" Tails: Have you ever stolen something valuable that did not belong to you?	Undesirable Response
5.	Tails	Heads: Did you vote in the presidential election held on November 4, 2008? Tails: Is the triangle red? 	Undesirable Response

	Coin Toss	Notecard	Question Type
6.	Tails	Heads: Do you think of yourself as homosexual, gay, lesbian, or bisexual? Tails: Respond "No"	Desirable Response
7.	Tails	Heads: Have you ever cheated on your taxes? Tails: Have you attended religious services in the last 7 days?	Unknown Answer
8.	Heads	Heads: Are three different types of balls shown below? Tails: Do you believe in God? 	Undesirable Response
9.	Heads	Heads: Respond "Yes" Tails: Have you ever had sex with a person you paid or who paid you for sex?	Undesirable Response
10.	Tails	Heads: Has a doctor or other health care professional ever told you that you had a sexually-transmitted disease, or STD? Some examples of STDs are genital herpes, genital warts, Chlamydia, gonorrhea, Human Papillomavirus (HPV), and syphilis. Tails: Is the blue square larger than the red square? 	Desirable Response