Variations of the Greenberg Unrelated Question Model-a Field Test

David P. Suarez, Sat Gupta, Emily Johnson, Padma Manthena

Abstract

In recent years, many complex survey designs have been introduced whose mathematical properties are very nice but such surveys may be very difficult to implement. The primary purpose of this study is to examine how efficiently some recently introduced variations of the famous Greenberg Unrelated Question Binary RRT model can be executed in a field survey. The first variation involves multiple independent responses using the original binary unrelated question model of Greenberg. The second variation involves using the original model with inverse sampling stopping at the first "yes" response. The data was collected via a survey conducted by the authors on a sample of undergraduate students in the age group of 18+ years who were enrolled in Mathematics/Statistics classes at the University of North Carolina at Greensboro, NC in September 2017. The binary response research question was "Have you ever been told by a healthcare professional that you had STD?" and the unrelated binary question was "Were you born during the months of January-April?". Our results are in line with the mathematical properties of these models.

Key words: Efficiency, Inverse Sampling, RRT models, Fieldwork Validation

AMS Subject Classification: 62D05

1 Introduction

Social desirability response bias (SDB) refers to the tendency of research subjects to give socially desirable responses instead of responses that are representative of their true feelings [7]. The bias in responses due to this personality trait becomes a major concern when the survey involves either sensitive topics such as politics, religion, and environment, or personal issues such as drug use and domestic violence [7]. A number of methods that can deal with this issue are suggested in the literature. These include the SDB scale of Reynolds et al. (1982) [14] and item response count technique of Coutts and Jann (2011) [3]. Another method is the randomized response technique (RRT), which was introduced originally by Warner [1965] [17] and then generalized by other researchers such as Greenberg et al. [1969, 1971] [5, 6], Warner [1971] [18], Klein and Spady [1993] [11], and Gupta et al. [2002, 2010, 2013] [8, 9, 10].

RRT models have been used quite a lot in field surveys. Abernathy et al. (1970) [1] used RRT models to obtain estimates on induced abortion rates in urban North Carolina. From the open survey, it was noticed that woman respondents would have hesitated to give a truthful response to the sensitive question of induced abortions. Striegel et al. (2006) [15] used indirect questioning techniques to measure how prevalent doping is among elite athletes. Official doping tests showed 0.81% of athletes testing positive for doping, while 6.8% of the athletes confessed to having practiced doping based on the RRT model [15]. In order to study the effect of higher education in favorable attitudes towards foreigners in Germany, Ostapczuk et al. (2009) [12] used two survey methods: direct questioning and RRT. The results obtained by these two survey methods demonstrated great variation. For the respondents who used RRT, the results obtained showed a sharp decline in the estimates for the proportion of xenophiles among both the less educated and highly educated Germans. Gill et al. (2013) [4] conducted a survey which involved a RRT model to estimate the prevalence of some risky sexual behaviors among students at the University of North Carolina at Greensboro. The binary question of interest in that study was "Have you been told by a healthcare professional that you have a sexually transmitted disease?", whereas the quantitative question of interest was "How many sexual partners have you had in the last 12 months?" [4]. The survey was conducted using three methods: optional RRT, direct face-to-face interviewing, and anonymous check-box survey method. It was observed that the optional unrelated question RRT method's estimates were closer to the anonymous check-box survey method's estimates. More recently, Chhabra et al. (2016) [2] used RRT models to estimate the prevalence of sexual abuse of female college students by either a friend or an acquaintance.

This study presents a field test of some variations of the Greenberg et al. (1969) model. A field survey of students was conducted at the University of North Carolina at Greensboro in which the research question was "Have you ever been told by a healthcare professional that you had STD?" and the unrelated question was "Were you born during the months of January-April?". The estimates of the prevalence from the anonymous group were compared with those given by the original Greenberg et al. (1969) model, the multiple independent response Greenberg model and the Greenberg model based on inverse sampling. The results of this fieldwork were reported in Suarez (2018) [16].

2 Models Used for the Fieldwork Validation

For this fieldwork validation, we used four groups. An anonymous survey was conducted in Group 1. The regular Greenberg et al. (1969) model was used in Group 2. The Greenberg model with multiple independent responses and the Greenberg model with inverse sampling were used in Groups 3 and 4, respectively. Theoretical details of these models are described in Suarez and Gupta (2018) [16].

2.1 Anonymous Group

For the anonymous group, the respondent is given a survey sheet by the researcher. This contained the sensitive question "Have you ever been told by a healthcare professional that you had a sexually-transmitted disease?". The respondent records his/her response to the question privately and drops the survey sheet in a box.

If π_A represents the prevalence of STD among students, then an estimator of π_A is given by

$$\hat{\pi}_A = \frac{n_1}{n}, \tag{1}$$

where n_1 is the number of "yes" responses in a sample of size n.

The variance of $\hat{\pi}_A$ is given by

$$Var(\hat{\pi}_A) = \frac{\pi_A(1-\pi_A)}{n}. \tag{2}$$

2.2 Greenberg et al. (1969) model

The second survey group used was the regular Greenberg et al. (1969) model. For this group, we prepared a deck of 100 cards where 85% of the cards contain the sensitive question and 15% of the cards contain an unrelated question ("Were you born during January-April?"). A randomization device offers respondents a choice between the sensitive question or an unrelated question. The researcher shuffles this deck and the respondent picks a card and answers the associated question. The researcher then records his/her answer to the question on a datasheet.

An estimator of π_A from this group is given by

$$\hat{\pi}_G = \frac{\hat{p}_y - \pi_B(1-p)}{p}, \tag{3}$$

where $\hat{p}_y$ is the sample proportion of "yes" responses, p is the probability of picking the sensitive question card from the deck, π_B is the known prevalence of a non-sensitive attribute B in the population. For our case, $\pi_B = \frac{4}{12}$ =P(being born during January-April).

The variance of the estimator $\hat{\pi}_G$ is given by

$$Var(\hat{\pi}_G) = \frac{p_y(1-p_y)}{np^2}, \tag{4}$$

where $p_y = \pi_A p + \pi_B(1-p)$.

2.3 Greenberg-Multiple Independent Responses Model

The third group used was the Greenberg et al. (1969) model but the respondent provides 3 independent responses. The researcher shuffles this deck each time and the respondent picks a card and speaks his/her response which is recorded. The process was repeated three times for each respondent and all responses are recorded.

An estimator of π_A from this group is given by

$$\hat{\pi}_{GM} = \frac{\frac{\bar{T}}{m} - (1-p)\pi_B}{p}, \tag{5}$$

where $\bar{T} = \sum_{i=1}^{3} T_i$, and T_i is the number of "yes" responses provided by the ith respondent in the sample.

It can be verified that

$$E(\hat{\pi}_{GM}) = \pi_A Var(\hat{\pi}_{GM}) \\ = \frac{p_y(1-p_y)}{nmp^2}, \tag{6}$$

where $m = 3$, $n =$ sample size, and p_y is as defined in Section 2.2.

2.4 Greenberg-Inverse Sampling Model

The fourth group used was the original Greenberg et al. (1969) model with inverse sampling while waiting for the first "yes" response. The researcher shuffles this deck and in turn, the respondent answers the question on the card they picked. The number of responses needed to get to the first "yes" response is recorded.

If S is the total number of trials needed to get to the first "yes" response, then

$$S \sim \text{Geometric}(p_y) \tag{7}$$

with $E(S) = \frac{1}{p_y}$ and $Var(S) = \frac{1-p_y}{p_y^2}$, where p_y is as defined earlier in Section 2.2.

Then, an estimator of π_A is given by

$$\hat{\pi}_{GI} = \frac{\frac{1}{S} - (1-p)\pi_y}{p}, \tag{8}$$

where $\bar{S}$ is the sample mean of independent observations on S.

It is verified in Suarez and Gupta (2018) that, up to first order of approximation,

$$E(\hat{\pi}_{GM}) = \pi_A \tag{9}$$

and,

$$Var(\hat{\pi}_{GM}) = \frac{p_y^2(1-p_y)}{np^2}. \tag{10}$$

3 Fieldwork Validation

3.1 Survey Description

The field survey was conducted at the University of North Carolina at Greensboro. The randomization device was a deck of 100 cards in which the sensitive question "Have you ever been told by a health-care professional that you had STD?" is asked with 85% probability and the unrelated question "Were you born during the months of January-April?" is asked with 15% probability.

Before participating in the survey, students were given a lecture about the four groups involved in the study, questions to be asked within the survey, and were assured of total confidentiality of their response. There were a total of 635 students who participated in the survey with 241 students in Group 1 (the anonymous group), 178 students in Group 2 and 216 students in Group 3.

There were practical constraints in executing the survey under the conditions of Group 4 where we were to record the number of trials needed to get to the first "yes" response. This would have required a very large sample. So, we simply looked at results of Group 3 one more time and observed there were 15 sequences of responses ending with a "yes" response. For these 15 sequences, we counted the number of trials (S), and $\bar{S}$ was the average of these 15 S-values.

3.2 Survey Results

For the first group, we had 11 "yes" responses out of 241 responses. For Group 2, 15 out of 178 subjects gave a "yes" response. For Group 3, we had 64 "yes" responses, and for Group 4, we had 15 strings of responses ending with a "yes" response. The total sample size across these 15 sequences was $n = 178$. Furthermore, we kept $\pi_B = \frac{1}{3}$ since we picked 4 months (January-April) out of 12 months. This led to the results shown in Table 1.

[13]

4 Discussion

Gupta et al. (2013) [10] cite previous studies that show that STD prevalence among college students vary between 10-25%. Compar-

Table 1: Estimated Prevalence (π_A) and Corresponding Variance from the 4 Groups

Estimator	n	$\hat{\pi}$	$Var(\hat{\pi})$
$\hat{\pi}_A$	241	0.04564315	0.0001807463
$\hat{\pi}_G$	178	0.04037607	0.0006000411
$\hat{\pi}_{GM}$	216	0.05742992	0.0001901209
$\hat{\pi}_{GI}$	178	0.04093619	0.0006065249

ing the prevalence of STD from the four groups with the above prevalence, by using Table (1), we see that the estimates of prevalence of STD from the four groups is much lower than the above prevalence for STD (0.10-0.25). However, the estimates are much closer to the estimate 4%-9% reported in Gill et al. (2013)[4].

We notice from Table (1) that the estimators $\hat{\pi}_A$ and $\hat{\pi}_{GM}$ are most precise with variances of 0.00018 and 0.00019, respectively. This is on expected lines based on theoretical results in Suarez and Gupta (2018) [16]. Although inverse sampling works well, it is difficult to implement it at a practical level since the overall sample size is likely to be very large. As a practical consideration, we recommend using a small value of m in $\hat{\pi}_{GM}$ such that $m \leq 3$.

References

[1] James R Abernathy, Bernard G Greenberg, and Daniel G Horvitz. Estimates of induced abortion in urban north carolina. *Demography*, 7(1):19–29, 1970.

[2] Anu Chhabra, BK Dass, and Sat Gupta. Estimating prevalence of sexual abuse by an acquaintance with an optional unrelated question rrt model. *The North Carolina Journal of Mathematics and Statistics*, 2:1–9, 2016.

[3] Elisabeth Coutts and Ben Jann. Sensitive questions in online surveys: Experimental results for the randomized response technique (rrt) and the unmatched count technique (uct). *Sociological Methods & Research*, 40(1):169–193, 2011.

[4] Tracy Spears Gill, Anna Tuck, Sat Gupta, Mary Crowe, and Jennifer Figueroa. A field test of optional unrelated question

randomized response models: estimates of risky sexual behaviors. In *Topics from the 8th Annual UNCG Regional Mathematics and Statistics Conference*, pages 135–146. Springer, 2013.

[5] Bernard G Greenberg, Abdel-Latif A Abul-Ela, Walt R Simmons, and Daniel G Horvitz. The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, 64(326):520–539, 1969.

[6] Bernard G Greenberg, Roy R Kuebler Jr, James R Abernathy, and Daniel G Horvitz. Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*, 66(334):243–250, 1971.

[7] Pamela Grimm. Social desirability bias. *Wiley International Encyclopedia of Marketing*, 2010.

[8] Sat Gupta, Bhisham Gupta, and Sarjinder Singh. Estimation of sensitivity level of personal interview survey questions. *Journal of Statistical Planning and inference*, 100(2):239–247, 2002.

[9] Sat Gupta, Javid Shabbir, and Supriti Sehra. Mean and sensitivity estimation in optional randomized response models. *Journal of Statistical Planning and Inference*, 140(10):2870–2874, 2010.

[10] Sat Gupta, Anna Tuck, Tracy Gill, and Mary Crowe. Optional unrelated-question randomized response models. *Involve, a Journal of Mathematics*, 6(4):483–492, 2013.

[11] Roger W Klein and Richard H Spady. An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, pages 387–421, 1993.

[12] Martin Ostapczuk, Jochen Musch, and Morten Moshagen. A randomized-response investigation of the education effect in attitudes towards foreigners. *European Journal of Social Psychology*, 39(6):920–931, 2009.

[13] David Perez-Suarez. Variations of the Greenberg unrelated question binary model. Master's thesis, University of North Carolina at Greensboro, 2018.

[14] William M Reynolds. Development of reliable and valid short forms of the marlowe-crowne social desirability scale. *Journal of clinical psychology*, 38(1):119–125, 1982.

[15] Heiko Striegel, Perikles Simon, Jochen Hansel, Andreas M Niess, and Rolf Ulrich. Doping and drug use in elite sports: An analysis using the randomized response technique: 1626: Board# 265 9: 30 am–10: 30 am. *Medicine & Science in Sports & Exercise*, 38(5):S247, 2006.

[16] David P Suarez and Sat Gupta. Variations of the Greenberg unrelated question binary model. *Involve: A Journal of Mathematics*, 11(1):119–126, 2018.

[17] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[18] Stanley L Warner. The linear randomized response model. *Journal of the American Statistical Association*, 66(336):884–888, 1971.

David Perez-Suarez
Department of Mathematics and Statistics, The University of North Carolina at Greensboro, Greensboro, NC 27402, USA
E-mail Address: d_suarez@uncg.edu

Sat Gupta
Department of Mathematics and Statistics, University of North Carolina at Greensboro, 317 College Ave, Greensboro, NC 27412, USA
E-mail Address: sngupta@uncg.edu

Emily Johnson
Department of Mathematics and Statistics, The University of North Carolina at Greensboro, Greensboro, NC 27402, USA
E-mail Address: ecnance@uncg.edu

Padma Manthena
Department of Mathematics and Statistics, The University of North Carolina at Greensboro, Greensboro, NC 27402, USA
E-mail Address: ppmanthe@uncg.edu