Statistica Sinica **31** (2021), 1871-1890
doi:https://doi.org/10.5705/ss.202016.0177

RANDOMIZED RESPONSE SAMPLING WITH APPLICATIONS TO TRACKING DRUGS FOR BETTER LIFE

Shu-Ching Su, Veronica I. Salinas, Monique L. Zamora, Stephen A. Sedory and Sarjinder Singh

Texas A&M University-Kingsville

Abstract: Randomized response is an interviewing technique designed to protect an interviewee's privacy and reduce a major source of bias (evasive answers or refusing to respond) when estimating the prevalence of sensitive characteristics using surveys of human populations. The objective of this study is to introduce a new method in the field of randomized response sampling that can be used to track the addictions of people to various substances. It has been suggested that students who use the smart drug "modafinil" are potentially putting their health at risk. We also discuss studies of similar addictions, based on the proposed technique.

Key words and phrases: Estimation of proportion, randomized response techniques, smart drug users.

1. Introduction

Estimating the relative size of a subgroup of a population is one of the most important tasks in statistical surveys. When a direct survey question about membership in a subgroup is sensitive, for example, whether a student belongs to a group that takes drugs, it usually suffers from non-negligible nonresponse (or false response). Here, indirect questioning designs such as randomized response technique offer the opportunity to elicit truthful responses by protecting the privacy of the respondents. This is very important in the field of empirical sociology, despite having to depart from the customary path of asking for information directly. Examples of collecting data using personal interview surveys on sensitive issues, such as induced abortions, drug abuse, and family income, are given by Fox (2015), Fox and Tracy (1986), Gjestvang and Singh (2006), Gjestvang and Singh (2009), Chaudhuri (2011), Chaudhuri and Christofides (2013), Su (2013), Su, Sedory and Singh (2014), and Su, Sedory and Singh (2017). Warner (1965) considered the case where the respondents in a population Ω can be divided into

Corresponding author: Sarjinder Singh, Texas A & M University-Kingsville, Kingsville, TX 78363, USA. E-mail: sarjinder.singh@tamuk.edu.

two mutually exclusive groups: one group with a stigmatizing/sensitive characteristic, and the other group without it. To estimate π_A, that is, the proportion of respondents in Ω belonging to the sensitive group, we select simple random sample s of n respondents, with replacement, from the population. To collect information on the sensitive characteristic, Warner (1965) used a randomization device. One such device is a deck of cards, with each card bearing one of the following two statements: (i) "I belong to group A", and (ii) "I do not belong to group A". The statements (i) and (ii) occur in the deck with relative frequencies P and $(1 - P)$, respectively. Each respondent in the sample s is asked to select a card at random from the well-shuffled deck. Without showing the card to the interviewer, the interviewee answers the question, "Is the statement true for you?" The number of people n_w that answered "Yes" is binomially distributed with parameters n and $\theta_w = P\pi_A + (1 - P)(1 - \pi_A)$. For large sample sizes, see Lee, Sedory and Singh (2013), the maximum likelihood estimator of π_A exists for $P \neq 0.5$, and is given by

$$\hat{\pi}_w = \frac{\hat{\theta}_w - (1 - P)}{2P - 1}, \tag{1.1}$$

where $\hat{\theta}_w = n_w/n$ is the observed proportion of "Yes" answers. The estimator $\hat{\pi}_w$ in (1.1) is unbiased for π_A, and the variance of the estimator $\hat{\pi}_w$ is given by

$$V(\hat{\pi}_w) = \frac{\pi_A(1 - \pi_A)}{n} + \frac{P(1 - P)}{n(2P - 1)^2}. \tag{1.2}$$

In the Warner (1965) model, the two questions relate to groups that are perfectly negatively associated with each other; that is, one group is the complement of the other group in the population of interest. The variance $V(\hat{\pi}_w)$ in (1.2) consists of two parts. The first is the same as that of direct question surveys, and the second corresponds to the protection of the respondents. However, it is intuitively evident that to protect the confidentiality of a respondent, it is not necessary for the two questions to be complementary. For example, one might use two unrelated questions (Do you belong to group A?/Do you belong to group Y?) In fact, it is sufficient to use some unrelated nonsensitive characteristic in the randomization device, as suggested by Greenberg et al. (1969). They proposed the unrelated questions model. In their model, the respondent answers one of two unrelated questions. For example, with probability P, he/she is asked, "Do you belong to group A?" and with probability $(1 - P)$, he/she is asked, "Is the last digit of your driving license number greater than eight?" Again, each respondent selected in the sample uses a device such as a deck of cards to determine the question to which they respond.

Let π_A be the true proportion of respondents in the population who possess the sensitive characteristic A. In addition, let π_Y be the true proportion of respondents in the population who possess a nonsensitive characteristic, say Y. This method ensures the privacy of respondents during a face-to-face survey. In the unrelated question model, the true probability of a "Yes" answer θ_G is given by

$$\theta_G = P\pi_A + (1-P)\pi_Y. \tag{1.3}$$

When π_Y is known, Greenberg et al. (1969) considered an unbiased estimator of the population proportion π_A, given by

$$\hat{\pi}_{G_1} = \frac{\hat{\theta}_G - (1-P)\pi_Y}{P}, \tag{1.4}$$

where $\hat{\theta}_G = n_G/n$ is the observed proportion of "Yes" answers in (1.4), and is an unbiased estimator for θ_G in (1.3).

When π_Y is unknown, then they suggest taking two independent samples of sizes n_1 and n_2, such that $n_1 + n_2 = n$. In the first sample with n_1 respondents, they suggest using a randomization device designed to ask the sensitive question with a probability P, so that the probability of a "Yes" answer becomes

$$\theta_1 = P\pi_A + (1-P)\pi_Y. \tag{1.5}$$

In the second sample with n_2 respondents, they suggest using a different independent randomization device, with associate probability T, such that the probability of a "Yes" answer becomes

$$\theta_2 = T\pi_A + (1-T)\pi_Y. \tag{1.6}$$

Greenberg et al. (1969) solved these two linear equations for π_A and developed an unbiased estimator $\hat{\pi}_{G_2}$, given by

$$\hat{\pi}_{G_2} = \frac{(1-T)\hat{\theta}_1 - (1-P)\hat{\theta}_2}{P-T}, \quad \text{for} P \neq T, \tag{1.7}$$

where $\hat{\theta}_1 = x_1/n_1$ and $\hat{\theta}_2 = x_2/n_2$ are the observed proportions of "Yes" answers in the first and second samples, respectively, and $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased estimators of θ_1 in (1.5) and θ_2 in (1.6), respectively.

The minimum variance of the estimator $\hat{\pi}_{G_2}$ in (1.7), using optimal values of

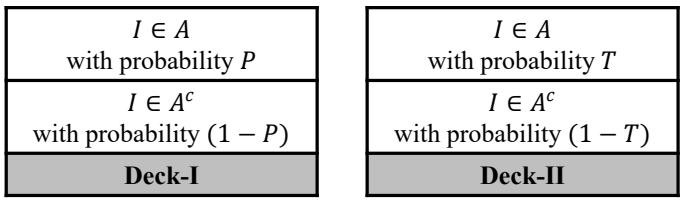

Figure 1. Two decks of cards.

n_1 and n_2, is given by:

$$\min V(\hat{\pi}_{G_2}) = \frac{[(1-T)\sqrt{\theta_1(1-\theta_1)} + (1-P)\sqrt{\theta_2(1-\theta_2)}]^2}{n(P-T)^2}, \tag{1.8}$$

Now, we discuss the efficient use of two decks of cards proposed by Odumade and Singh (2009). Each respondent in the simple random sample with replacement (SRSWR) of size n is provided with two decks of cards, marked as Deck-I and Deck-II, as shown in Figure 1.

Each respondent is requested to draw two cards, one card from each deck, and to read the statements in order. The respondent first matches his/her status with the statement written on the card drawn from the first deck. Then he/she matches his/her status with the statement written on the card drawn from the second deck. Let π_A be the true proportion of respondents in the population who possess the characteristic A. Consider a situation in which the selected respondent belongs to group A. If he/she draws a card with statement $I \in A$ with probability P from Deck-I, and a card with statement $I \in A$ with probability T from Deck-II, then he/she reports: (Yes, Yes). Now consider the situation in which the selected respondent belongs to group A^c. If he/she draws a card with the statement $I \in A^c$ with probability $(1-P)$, from Deck-I, and a card with the statement $I \in A^c$ with probability $(1-T)$ from Deck-II, then he/she also reports: (Yes, Yes). Thus, the response (Yes, Yes) can come from both types of respondents and, hence, their privacy is maintained. Therefore, the probability of a (Yes, Yes) response is given by

$$P(Yes, Yes) = \lambda_{11} = PT\pi_A + (1-P)(1-T)(1-\pi_A), \tag{1.9}$$

Similarly, the probabilities of getting (Yes, No), (No, Yes), and (No, No) responses are, respectively, given by:

$$P(Yes, No) = \lambda_{10} = P(1-T)\pi_A + (1-P)T(1-\pi_A), \tag{1.10}$$

$$P(No, Yes) = \lambda_{01} = (1-P)T\pi_A + P(1-T)(1-\pi_A), \tag{1.11}$$

and

$$P(No, No) = \lambda_{00} = (1-P)(1-T)\pi_A + PT(1-\pi_A). \tag{1.12}$$

Let $\hat{\lambda}_{11} = n_{11}/n$, $\hat{\lambda}_{10} = n_{10}/n$, $\hat{\lambda}_{01} = n_{01}/n$, and $\hat{\lambda}_{00} = n_{00}/n$ be the observed proportions of (Yes, Yes), (Yes, No), (No, Yes), and (No, No) responses, respetively. Note that $\hat{\lambda}_{11}$, $\hat{\lambda}_{10}$, $\hat{\lambda}_{01}$, and $\hat{\lambda}_{00}$ are unbiased estimators of λ_{11}, λ_{10}, λ_{01}, and λ_{00}, defined in (1.9), (1.10), (1.11), and (1.12), respectively. Odumade and Singh (2009) defined the least squared distance between the observed proportions and the true proportions as:

$$D_1 = \frac{1}{2}\sum_{i=0}^{1}\sum_{j=0}^{1}(\lambda_{ij} - \hat{\lambda}_{ij})^2. \tag{1.13}$$

They chose as their estimate the value of π_A that minimized D_1 in (1.13). Setting $\partial D_1/\partial \pi_A = 0$, they arrive at the unbiased estimator of π_A given by

$$\hat{\pi}_{OS} = \frac{1}{2} + \frac{(P+T-1)(\hat{\lambda}_{11} - \hat{\lambda}_{00}) + (P-T)(\hat{\lambda}_{10} - \hat{\lambda}_{01})}{2[(P+T-1)^2 + (P-T)^2]}. \tag{1.14}$$

The variance of the estimator $\hat{\pi}_{OS}$ in (1.14) is given by

$$\begin{aligned} &V(\hat{\pi}_{OS}) \\ &= \frac{(P+T-1)^2\{PT + (1-P)(1-T)\} + (P-T)^2\{T(1-P) + P(1-T)\}}{4n[(P+T-1)^2 + (P-T)^2]^2} \\ &\quad - \frac{(2\pi_A - 1)^2}{4n}. \end{aligned} \tag{1.15}$$

Note that if $T = P = P_0$ (say), the variance of the estimator $\hat{\pi}_{OS}$ in (1.15) becomes

$$V(\hat{\pi}_{OS})_{P=T=P_0} = \frac{\pi_A(1-\pi_A)}{n} + \frac{P_0(1-P_0)}{2n(2P_0-1)^2} = V(\hat{\pi}_w)_{q=2}(say). \tag{1.16}$$

Note too that the variance in (1.16) is the same as that when each respondent uses the Warner (1965) device twice.

The second model of Greenberg et al. (1969) with an unknown value of π_Y, is more practical in terms of increasing respondents' cooperation. However it requires two independent samples, which makes it complicated to apply in practice. In addition, the optimum sample sizes depend on the population proportion of the sensitive characteristic being estimated. Recent studies, including Lee, Sedory and Singh (2013), Abdelfatah, Mazloum and Singh (2013), Arnab, Singh and North (2012), Singh and Sedory (2011), Singh and Sedory (2012), and

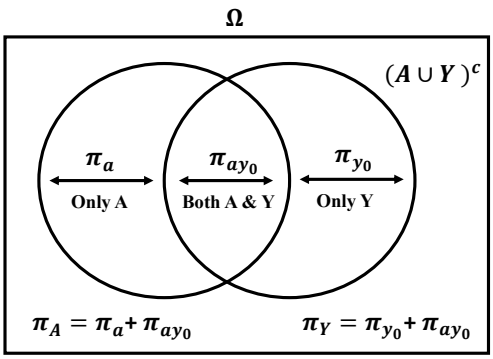

Figure 2. Pictorial representation of the population under study.

Singh and Kim (2011), as well as several papers in the special issue on randomized response sampling by Singh (2014) and others, have paid little attention to unrelated question models. This motivated us to consider improvements to this unrelated question model when π_Y is unknown.

The remainder of the paper proceeds as follows. Section 1 presents the theoretical background for the paper. In Section 2, we construct a new unrelated question model estimator using the least squared distance as well as maximum likelihood estimator. Section 3 establishes the equivalence of the least squared and maximum likelihood estimates. Here, we also use simulations to show that the proposed unrelated question model is more efficient than that of Greenberg et al. (1969). Section 4 applies the proposed estimator to a survey conducted at Texas A&M University-Kingsville. Section 5 provides black-box estimates for comparisons, and Section 6 provides a second application of the proposed model at a conference. Section 7 concludes the paper.

2. Proposed Unrelated Question Model

Let A denote the set of members of Ω possessing the sensitive attribute, and Y be the set of those with the unrelated, nonsensitive attribute. For a population Ω, it is clear that some members possess only the sensitive attribute, some possess only the nonsensitive attribute, some possess both, and some possess neither attribute. A pictorial representation of such a population is shown the Venn diagram in Figure 2.

Let π_a be the population proportion of people possessing only the sensitive characteristic, $a = A \cap Y^c$, π_{y_0} be the population proportion possessing only the nonsensitive characteristic, $y_0 = Y \cap A^c$, and π_{ay_0} be the population proportion possessing both attributes, $ay_0 = A \cap Y$. Note that $\Omega = a \cup y_0 \cup ay_0 \cup (a \cup y_0 \cup ay_0)^c$.

We then have $\pi_A = \pi_a + \pi_{ay_0}$; that is, the proportion π_A of people possessing the sensitive characteristic is the sum of the proportion π_a of people possessing only the sensitive characteristic A and the proportion π_{ay_0} of people possessing both the sensitive characteristic A and the nonsensitive characteristic Y. Similarly, $\pi_Y = \pi_{y_0} + \pi_{ay_0}$; that is, the proportion π_Y of people possessing the nonsensitive characteristic Y is the sum of the proportion π_{y_0} of people possessing only the nonsensitive characteristic Y and the proportion π_{ay_0} of people possessing both the sensitive characteristic A and the nonsensitive characteristic Y .

We select an SRSWR of n respondents from the given population Ω. Each respondent is provided with two shuffled decks of cards. We label the first deck as the green deck, and the second deck as the pink deck. The green deck consists of two types of cards, each bearing one of two questions. Let P be the proportion of cards bearing the question, "(i) Do you possess characteristic A?" and $(1-P)$ be the proportion of cards bearing question, "(ii) Are you a member of group Y?" The pink deck also consists of the same two types of cards, but in proportions T and $(1-T)$, respectively .

A selected respondent is requested to draw one card from the green deck, read the question on the card silently, and then respond truthfully either "Yes" or "No" to the question. The respondent is requested to mix the drawn card back into the deck. Next, the same respondent is requested to draw a card from the pink deck, read the question on the card silently, and then respond truthfully either "Yes" or "No" to the question. An observed response from a respondent can be classified into one of the four mutually exclusive categories: (Yes, Yes), (Yes, No), (No, Yes), or (No, No). The same process is repeated with all n respondents. The probabilities of getting (Yes, Yes), (Yes, No), (No, Yes), and (No, No) responses are, respectively, given by

$$P(Yes, Yes) = \theta_{11} = PT\pi_a + \pi_{ay_0} + (1-P)(1-T)\pi_{y_0}, \tag{2.1}$$

$$P(Yes, No) = \theta_{10} = P(1-T)\pi_a + (1-P)T\pi_{y_0}, \tag{2.2}$$

$$P(No, Yes) = \theta_{01} = P(1-T)\pi_{y_0} + (1-P)T\pi_a, \tag{2.3}$$

and

$$P(No, No) = \theta_{00} = 1 - \pi_a(P + T - PT) - \pi_{ay_0} - \pi_{y_0}(1-PT). \tag{2.4}$$

Note that θ_{11}, θ_{10}, θ_{01}, and θ_{00}, defined in (2.1), (2.2), (2.3), and (2.4), respectively, satisfy the condition $\theta_{11} + \theta_{10} + \theta_{01} + \theta_{00} = 1$.

Our aim is to estimate the unknown proportions π_a and π_{ay_0} of the respondents belonging to the groups $a = A \cap Y^c$ and $ay_0 = A \cap Y$, respectively, and

then to estimate the required proportion, $\pi_A = \pi_a + \pi_{ay_0}$, of those belonging to the group A.

Let $\hat{\theta}_{11} = n_{11}/n$, $\hat{\theta}_{10} = n_{10}/n$, $\hat{\theta}_{01} = n_{01}/n$, and $\hat{\theta}_{00} = n_{00}/n$ be the observed proportions of (Yes, Yes), (Yes, No), (No, Yes), and (No, No) responses, respectively, from the n respondents selected in the sample. Following Odumade and Singh (2009), we define the squared distance between the observed proportions and the true proportions as

$$D = \frac{1}{2}\sum_{i=0}^{1}\sum_{j=0}^{1}(\theta_{ij} - \hat{\theta}_{ij})^2. \tag{2.5}$$

We minimize the squared distance D in (2.5) with respect to the three parameters of interest π_a, π_{ay_0}, and π_{y_0}. We minimize D because this leads to simple, unbiased, and closed, form estimators of the three required proportions.

Now, we set $\partial D/\partial \pi_a = 0$, $\partial D/\partial \pi_{ay_0} = 0$, and $\partial D/\partial \pi_{y_0} = 0$.

The solution to the resulting system of linear equations leads, by the method of moments, to the three estimators in (2.6), (2.7), and (2.8), as follows:

$$\hat{\pi}_a = \frac{(P-T)(1-\hat{\theta}_{11}-\hat{\theta}_{00}) - \hat{\theta}_{10}(4PT-3P-T) - \hat{\theta}_{01}(P+3T-4PT)}{4(P-T)(P+T-2PT)}, \tag{2.6}$$

$$\begin{aligned}\hat{\pi}_{ay_0} = \frac{1}{4(P-T)(P+T-2PT)}[&(P-T)\hat{\theta}_{11}(1+2P+2T-4PT)\\ &+\hat{\theta}_{10}(2P-1)(2T^2+2PT-P-3T)\\ &+\hat{\theta}_{01}(2T-1)(3P+T-2PT-2P^2)\\ &+\hat{\theta}_{00}(2P-1)(2T-1)(P-T) - (2P-1)(2T-1)(P-T)],\end{aligned} \tag{2.7}$$

and

$$\hat{\pi}_{y_0} = \frac{(P-T)(1-\hat{\theta}_{11}-\hat{\theta}_{00}) + \hat{\theta}_{10}(4PT-P-3T) - \hat{\theta}_{01}(4PT-3P-T)}{4(P-T)(P+T-2PT)}. \tag{2.8}$$

We propose the following estimator of the required population proportion π_A:

$$\hat{\pi}_A = \hat{\pi}_a + \hat{\pi}_{ay_0}. \tag{2.9}$$

Equivalently, the estimator in (2.9) can be written as

$$\hat{\pi}_A = \frac{(P-T)(\hat{\theta}_{11}-\hat{\theta}_{00}) + (P+T-2)(\hat{\theta}_{01}-\hat{\theta}_{10}) + (P-T)}{2(P-T)}. \tag{2.10}$$

Note that π_{y_0} is not of interest, so we do not investigate further any property

of its estimator $\hat{\pi}_{y_0}$ in (2.8). For the derivations of (2.6), (2.7), (2.8), and (2.10), please see Appendix A in the online Supplementary Material. The bias and variance of the estimator $\hat{\pi}_A$ of π_A are addressed in the following theorem.

Theorem 1. *The estimator $\hat{\pi}_A$ is an unbiased estimator of the population proportion π_A, with variance given by:*

$$V(\hat{\pi}_A) = \frac{\pi_a(1-\pi_a)}{n} + \frac{\pi_{ay_0}(1-\pi_{ay_0})}{n} - \frac{2\pi_a\pi_{ay_0}}{n} + \frac{(1-P)(1-T)(P+T-2PT)(\pi_a+\pi_{y_0})}{n(P-T)^2}. \tag{2.11}$$

Proof. See Appendix A in the online Supplementary Material.

Remark 1. We suggest estimating the variance $V(\hat{\pi}_A)$ of the estimator $\hat{\pi}_A$ as follows:

$$\hat{V}(\hat{\pi}_A) = \frac{\hat{\pi}_a(1-\hat{\pi}_a)}{n-1} + \frac{\hat{\pi}_{ay_0}(1-\hat{\pi}_{ay_0})}{n-1} - \frac{2\hat{\pi}_a\hat{\pi}_{ay_0}}{n} + \frac{(1-P)(1-T)(P+T-2PT)(\hat{\pi}_a+\hat{\pi}_{y_0})}{n(P-T)^2}. \tag{2.12}$$

Remark 2. In this remark, we consider a likelihood function based on the probability mass function of the number of observed responses, as given by

$$L = \binom{n}{n_{11}, n_{10}, n_{01}, n_{00}} \theta_{11}^{n_{11}} \theta_{10}^{n_{10}} \theta_{01}^{n_{01}} \theta_{00}^{n_{00}}. \tag{2.13}$$

Taking ln of both sides of (2.13), we get

$$ln(L) = ln\binom{n}{n_{11}, n_{10}, n_{01}, n_{00}} + n[\hat{\theta}_{11} ln(\theta_{11}) + \hat{\theta}_{10} ln(\theta_{10}) + \hat{\theta}_{01} ln(\theta_{01}) + \hat{\theta}_{00} ln(\theta_{00})]. \tag{2.14}$$

It can be shown that the maximum likelihood estimates obtained by maximizing the log-likelihood function in (2.14) are the same as the least squared distance estimates. Refer to Appendix A in the online Supplementary Material.

In the next section, we compare the proposed estimators with the Greenberg et al. (1969) estimator when the population proportion of the nonsensitive characteristic is unknown, and with the Odumade and Singh (2009) estimator.

3. Relative Efficiency and Protection of Respondents

We define the relative efficiency of the proposed estimator $\hat{\pi}_A$ with respect to the Greenberg et al. (1969) estimator $\hat{\pi}_{G_2}$ as

$$RE(1) = \frac{\min V(\hat{\pi}_{G_2})}{V(\hat{\pi}_A)}, \tag{3.1}$$

where $\min V(\hat{\pi}_{G_2})$ is given in (1.8), and $V(\hat{\pi}_A)$ is given in (2.11). We define the relative efficiency of the proposed maximum likelihood estimate $\hat{\pi}_A^{mle}$ as

$$RE(2) = \frac{\min V(\hat{\pi}_{G_2})}{V(\hat{\pi}_A^{mle})}, \tag{3.2}$$

where $V(\hat{\pi}_A^{mle})$ is the Cramer–Rao lower bound of the maximum likelihood estimate (see Appendix A in the online Supplementary Material). We also define the percent relative efficiency of the proposed estimator $\hat{\pi}_A$ with respect to the Odumade and Singh (2009) estimator $\hat{\pi}_{OS}$ as

$$RE(OS) = \frac{V(\hat{\pi}_{OS})}{V(\hat{\pi}_A)}, \tag{3.3}$$

where $V(\hat{\pi}_{OS})$ is defined in (1.15).

Lee et al. (2016) used a privacy protection measure to suggest a generalization of that of Lanke (1976), given by $L = \max[P(A|Yes), P(A|No)]$. They proposed a new measure of protection for a respondent using the two-decks model proposed by Odumade and Singh (2009). For the later method or equivalently, those of Singh and Sedory (2011), and Singh and Sedory (2012), Lee et al. (2016) compute four conditional probabilities, as follows: $P[A|(Yes, Yes)] = PT\pi_A/\lambda_{11}$; $P[A|(Yes, No)] = P(1-T)\pi_A/\lambda_{10}$; $P[A|(No, Yes)] = (1-P)T\pi_A/\lambda_{01}$; and $P[A|(No, No)] = (1-P)(1-T)\pi_A/\lambda_{00}$. In (3.4), we define the least protection in the Odumade and Singh (2009) model as follows:

$$\begin{aligned} Prot(OS\ Model) = &\max\{P[A|(Yes, Yes)], P[A|(Yes, No)], \\ &P[A|(No, Yes)], P[A|(No, No)]\}. \end{aligned} \tag{3.4}$$

For the proposed unrelated questions model, we compute the same four conditional probabilities, as follows: $P^*[A|(Yes, Yes)] = (PT\pi_a + \pi_{ay_0})/\theta_{11}$; $P^*[A|(Yes, No)] = P(1-T)\pi_a/\theta_{10}$; $P^*[A|(No, Yes)] = (1-P)T\pi_a/\theta_{01}$; and $P^*[A|(No, No)] = (1-P)(1-T)\pi_a/\theta_{00}$.

Then, the least protection in the proposed unrelated question model is given in (3.5) by

$$\begin{aligned} Prot(Proposed\ Unrelated\ Model) = &\max\{P^*[A|(Yes, Yes)], P^*[A|(Yes, No)], \\ &P^*[A|(No, Yes)], P^*[A|(No, No)]\}. \end{aligned} \tag{3.5}$$

The relative protection of the proposed model over that of the Odumade and Singh (2009) model (or equivalently Singh and Sedory (2011) and Singh and Sedory (2012)) is defined as:

$$RP(OS) = \frac{Prot(OS\ Model)}{Prot(Proposed\ Unrelated\ Model)}. \tag{3.6}$$

In the case of the Greenberg et al. (1969) two-sample model, we compute the least protection level as

$$\begin{aligned} Prot(Greenberg\ Model) = \max\{&P_1[A|(Yes)], P_1[A|(No)], \\ &P_2[A|(Yes)], P_2[A|(No)]\}, \end{aligned} \tag{3.7}$$

where $P_1[A|(Yes)] = \{P + (1-P)\pi_Y\}\pi_A/\theta_1$, $P_1[A|(No)] = (1-P)(1-\pi_Y)\pi_A/(1-\theta_1)$, $P_2[A|(Yes)] = \{T + (1-T)\pi_Y\}\pi_A/\theta_2$, and $P_2[A|(No)] = (1-T)(1-\pi_Y)\pi_A/(1-\theta_2)$; have their usual meanings in (3.7).

The relative protection of the proposed model over that of the Greenberg et al. (1969) model is defined as

$$RP(G) = \frac{Prot(Greenberg\ Model)}{Prot(Proposed\ Unrelated\ Model)}. \tag{3.8}$$

Note that the relative efficiencies defined in (3.1)–(3.3) and the relative protections defined in (3.6) and (3.8) are free from the sample size. We wrote SAS code (see Appendix A in the online Supplementary Material) to compute the relative efficiency and relative protection for various values of parameters. We fixed $P = 0.686$ and $T = 0.314$, and varied the other required parameter values over the ranges $0.05 \leq \pi_a \leq 0.50$, and $0.05 \leq \pi_{ay_0} \leq 0.30$ for different choices of π_{y_0}, such that $RP(OS) > 1$ and $RE(OS) > 1$. The results are presented in Table 9 in Appendix A in the online Supplementary Material. From Table 9, note that $RE(1) = RE(2)$; that is, the proposed estimator $\hat{\pi}_A$ attains the lower bound of the variance. We prefer the proposed estimator $\hat{\pi}_A$ over the maximum likelihood estimate because it is in closed form. In addition it is unbiased, and it is easy to estimate its variance to construct confidence interval estimates. Both proposed estimators $\hat{\pi}_A$ and $\hat{\pi}_A^{mle}$ are more efficient than the Greenberg et al. (1969) estimator $\hat{\pi}_{G_2}$, but remain less (or more) protective, as indicated by the values of $RE(1) = RE(2)$ and $RP(G)$ in Table 9. In other words, in Table 9, the value of $RP(G) < 1$ shows that a single-trial question is sometimes more protective than the two trials per respondent question model, but remains significantly less efficient than the later, as indicated by $RE(1) = RE(2)$. From the simulation

study, we conclude that there are choices for the proportion of an unrelated characteristic π_Y in a population such that the proposed model performs at least as well as the Odumade and Singh (2009) model, from both protection and relative efficiency points of views.

Note that the protection criterion cannot be implemented on a given subject to determine whether he/she is a member of the sensitive group. Greenberg et al. (1969) choose π_Y close to π_A so that their model performs well. Note that if $\pi_Y \approx \pi_A$, then $P_1(A|Yes) = \{P + (1-P)\pi_Y\}\pi_A/\theta_1 \approx P + (1-P)\pi_Y$, which is free from the value of θ_1. Further note that these protection criteria divide people into two groups, those who responded "Yes" and those who responded "No"; that is, the respondents who reported "Yes" may be part of the sensitive group, with some conditional probability, as are those who reported "No", but with a different conditional probability. Thus, a protection criterion cannot be used to classify respondents into two groups.(i.e., sensitive or nonsensitive groups).

In the next section, we apply the proposed unrelated question model to investigate the prevalence of smart drugs at Texas A&M University-Kingsville.

4. Real-Data Application at Texas A&M University-Kingsville

Sky (2013) in the United Kingdom has suggested that students who use the smart drug "modafinil" are potentially putting their health at risk. Sabawi (2012) also highlights students' habit of taking smart drugs during stressful times. These articles motivated us to conduct this study during the Fall 2013 and Spring 2014 semesters at Texas A&M University-Kingsville. Two decks of cards were prepared to collect data from the students: a green deck and a pink deck. The green deck consists of 51 cards, with 35 cards bearing the question, "Have you ever used a smart drug in your college career?" and the remaining 16 cards bearing the question, "Is the last digit of your K-ID number greater than or equal to eight?" Thus $P = 0.686$ for the green deck. The pink deck also contains 51 cards, with 16 cards bearing the question, "Have you ever used a smart drug in your college career?" and the remaining 35 cards bearing the question, "Is the last digit of your K-ID number greater than or equal to eight?" Thus, $T = 0.314$ for the pink deck.

We used convenience sampling to collect data from the students. Each student who agreed to participate in the survey had to be at least 18 years old. Respondents are asked to first draw a card from the green deck. They are told to read the question on the drawn card silently, and to answer it truthfully. Lastly, the card is returned to the green deck, without showing it to anyone. The stu-

Table 1. Response from undergraduates.

Overall	Yes	No	Sum
Yes	11	8	19
No	6	102	108
Sum	17	110	127

Table 2. Response from male students.

Males	Yes	No	Sum
Yes	4	5	9
No	3	51	54
Sum	7	56	63

Table 3. Response from female students.

Females	Yes	No	Sum
Yes	7	3	10
No	3	51	54
Sum	10	54	64

dent then repeats the process for the pink deck. Each response is recorded on a response card as follows: **Gender:** male or female; **Seniority:** UG or G; **Response:** (Yes, Yes), (Yes, No), (No, Yes), or (No, No). The symbol "G" indicates a graduate student and "UG" indicates an undergraduate. No other information was collected from the students participating in the survey. An incentive of chocolate and candy induced 127 undergraduate students to participate in the survey. Only 11 graduate students participated, so those were discarded from the analysis. Although it was a convenience sample, it turned out that 63 boys and 64 girls participated in the study. The overall responses of the 127 undergraduate students were classified into a 2×2 contingency table, see Table 1.

We estimate the proportion of undergraduate students who had used a smart drug as 0.1629, with a standard error of 0.049336. The 95% confidence interval estimate is $(0.0662, 0.2596)$.

The 2×2 contingency table shown in Table 2 shows the observed responses from the 63 males who participated in the survey.

The estimate of the proportion of male students who have ever used a smart drug is 0.1696, with a standard error of 0.07355. The 95% confidence interval estimate is $(0.02548, 0.31383)$.

Similarly, the 2×2 contingency table in Table 3 shows the observed responses for the 64 females who participated in the survey.

Table 4. Black-box responses from undergraduates.

	Yes	No	Total	Prop
Overall	17	110	127	0.1339
Males	9	54	63	0.1429
Females	8	56	64	0.1250

The estimate of the proportion of female students who have ever used a smart drug is 0.1563, with a standard error of 0.06615. The 95% confidence interval estimate is $(0.02659, 0.2859)$. Note that we compute the standard errors using the square root of $\hat{V}(\hat{\pi}_A)$, given in (2.12).

5. Black-Box Technique

For comparison purposes, we also used a black box to collect data from the same students who participated in the randomized response surveys. Each respondent was given a card on which they provided the following information: **Gender**: male or female; **Seniority:** UG or G; and a response to direct question on whether or not they had ever used a smart drug: yes or no. Each respondent circled his/her response and inserted the card in a locked black box, without showing his/her response to the interviewers. A black-box technique is only effective in inducing an honest answer if respondents have confidence that the interviewer does not know the contents of the box before and after he/she has given a response. It is helpful to compare this "almost direct question" technique and the proposed randomized response technique. The black box was locked to assure respondents of their anonymity. Table 4 shows the black-box responses from the 127 students.

These responses estimate that the proportion of undergraduate smart drug users at Texas A&M University-Kingsville is 0.1339; that of male students is 0.1429, and that of female students is 0.1250. A comparison of the two estimates is given in Figure 3.

All estimates (overall, male, and female) obtained using the black-box techniques were lower than those obtained using the proposed randomized response technique. The overall randomized response estimate is 0.1629, compared with the black-box estimate of 0.1339; for males, the randomized response estimate is 0.1696, and the black-box estimate is 0.1429; for females the randomized response estimate is 0.1563, and the black-box estimate is 0.1250.

We also calculated Z_{cal} as

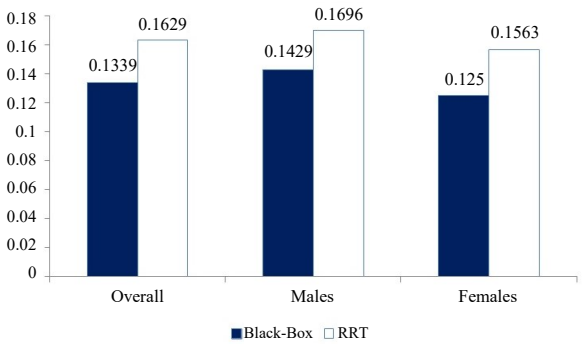

Figure 3. Estimates of smart drug users at Texas A&M University-Kingsville.

$$Z_{cal} = \frac{\hat{\pi}_A - \hat{\pi}_{BB}}{\sqrt{\hat{V}(\hat{\pi}_A) + \hat{V}(\hat{\pi}_{BB})}}, \tag{5.1}$$

where $\hat{\pi}_{BB}$ stands for the black-box estimate, with

$$\hat{V}(\hat{\pi}_{BB}) = \frac{\hat{\pi}_{BB}(1 - \hat{\pi}_{BB})}{n - 1}. \tag{5.2}$$

Here, $\hat{V}(\hat{\pi}_A)$ is given in (2.12), and in (5.2), $\hat{V}(\hat{\pi}_{BB})$ is the estimator of the variance of the sample proportion using the black box. Using (5.1), the calculated value of Z_{cal} is 0.5007 for the overall estimate, 0.3112 for males, and 0.3997 for females. These computed Z_{cal} values show that there is no significant difference between the estimates obtained from the proposed randomized response technique and those obtained from the black-box technique. We admit that a better designed survey should be conducted, making use of a probability sample in order to reach a more justifiable conclusion. However if these estimates are reasonably accurate, then students need to be taught about the adverse effects of smart drugs.

6. Real-Data Application: Booth Stat-Hawkers at Montreal, Canada

We conducted another convenience survey at the booth STAT-HAWKERS, during the Joint Statistical Meeting (JSM) 2013, Montreal, Canada. This survey had two objectives: (i) to increase awareness of randomized response techniques; and (ii) to estimate the prevalence of smart drug use. The data were collected over three days using a randomization device consisting of two decks. Again, the green deck consisted of 51 cards, with 35 cards bearing the question, "Have you ever used a smart drug?" and the remaining 16 cards bearing the question,

Table 5. Overall randomized responses at the JSM.

Overall	Yes	No	Sum
Yes	9	4	13
No	9	73	82
Sum	18	77	95

Table 6. Males randomized responses at the JSM.

Males	Yes	No	Sum
Yes	8	1	9
No	3	38	41
Sum	11	39	50

Table 7. Females randomized responses at the JSM.

Females	Yes	No	Sum
Yes	1	3	4
No	6	35	41
Sum	7	38	45

"Were you born on the first, second, third, fourth, fifth, or sixth of a month?" Thus, $P = 0.686$ in the green deck. The pink deck also consisted of 51 cards, with 16 cards bearing the first question and the remaining 35 cards bearing the second question. Thus, $T = 0.314$ in the pink deck. In other words, it was very much the same randomization device as that used in the previous application at Texas A&M University-Kingsville. Data were collected from 95 participants. The overall responses were classified into a 2×2 contingency table, shown in Table 5.

Using the proposed method, we estimated the proportion of conference attendees who had ever used a smart drug as 0.092417, with a standard error of 0.05599. 2×2 contingency table in Table 6 shows the observed responses from the 50 males who participated in the survey.

Using the proposed estimator, we estimated the proportion of male conference attendees who had used a smart drug as 0.1463, with a standard error of 0.070995. 2×2 contingency table in Table 7 shows the observed responses for the 45 females who participated.

Using the proposed estimator, we estimated that the proportion of female conference attendees who had used a smart drug as 0.032616, with a standard error of 0.087355. A pictorial presentation of conference smart drug users is given in Figure 4.

Thus, based on our theoretical study, simulations, and real-data applications,

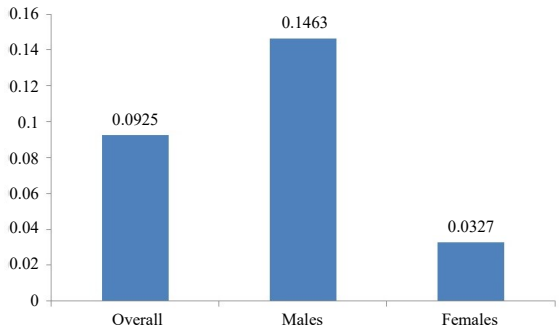

Figure 4. Estimates of smart drug users at the conference.

Table 8. Choices of parameters in the proposed model.

P	T	RP(OS)	RP(G)	RE(OS)	RE(1)	RE(2)
0.60	0.35	1.0444	1.1350	3.5289	1.0733	2.3799
0.70	0.35	1.0405	1.0237	1.8777	1.2680	2.1234
0.60	0.45	0.6787	1.0543	1.5939	1.2865	2.4341
0.70	0.45	0.7477	1.0237	1.2584	1.1998	2.2362
0.60	0.35	1.0444	1.1350	3.5289	1.0733	2.3799

we conclude that the proposed unrelated question model is more efficient than its competitors when used in real large-scale surveys containing sensitive questions.

Remark 3.

(a) We acknowledge that in the simulation study, we set $P = 0.686$ in the green deck and $T = 0.314$ in the pink deck, because we used the same values in the real-data applications. Note that in the proposed model, P cannot be equal to T. Thus, to make the proposed estimator efficient if one chooses P between 0.5 and 1.0, T should be in its complement between 0.0 and 0.50. Note that one should check for other choices of parameters by executing the SAS macro given in Appendix A to determine whether the proposed model is working efficiently for a new survey, based on a good guess of the proportions of the sensitive and unrelated attributes being used in the survey. For example, consider $\pi_a = 0.05$, $\pi_{y_0} = 0.70$, and $\pi_{ay_0} = 0.02$. Table 8 gives the results for various choices of P and T.

(b) One can increase the number of decks in a survey, but respondents may be hesitant to respond several times to the same question; thus two decks seems most appropriate.

7. Conclusion

This study considers the design of randomized response sampling when the survey estimates the prevalence of sensitive characteristics among a target population. To protect an interviewee's privacy and to reduce bias, several randomized response sampling methods have been proposed in the literature. For instance, Warner (1965) introduces a randomization device (such as a deck of cards) for two perfectly negatively associated questions. Each question occurs with a certain probability in the deck. A respondent is asked to randomly select a question (one card from the deck), and to answer the question without showing it to the interviewer. To further protect the confidentiality of a respondent, Greenberg et al. (1969) extend Warner's method to a design with two unrelated questions. This study further extends the Greenberg et al. (1969) approach to a "two deck of cards" design, where a respondent is asked to answer two questions randomly sampled from two decks. Maximum likelihood estimators of the interesting population parameters are derived, and the relative efficiency with respect to existing approaches is established. Note that we have consider an SRSWR sampling scheme, because the proposed model is compared with existing models that assume the same scheme. In addition, if the finite-population correction factor is small, SRSWR and simple random sampling without replacement designs perform similiarly. If required, one can extend the proposed model to complex survey designs by following Arnab, Singh and North (2012).

Supplementary Material

The online Supplementary Material provides detailed derivations of equations (2.6), (2.7), (2.8), and (2.10), proofs for Theorem 1 and Remark 2, SAS codes, and empirical evidence in Table 1.

Acknowledgments

The authors are thankful to the Editor-in-Chief Dr. Zhiliang Ying, Anna Chiang, Ms. Kay Mo and Yuvia Hsiang and the referee for their constructive comments. We also thank the IRB Chair Dr. Stephen D. Oller, Research Compliance Liaison Donna J. Pulkrabek, and committee members for their timely IRB approvals to collect data at Texas A&M University-Kingsville and at the JSM-2013, Montreal, Canada. All authors were either students or faculty at the Department of Mathematics, Texas A&M University-Kingsville while completing this research.

References

Abdelfatah, S., Mazloum, R and Singh, S. (2013). Efficient use of two-stage randomized response procedure. *Brazilian J. of Probability and Statistics* **60**, 63–69.

Arnab, R., Singh, S. and North, D. (2012). Use of two decks of cards in randomized response techniques for complex survey designs. *Communications in Statistics-Theory and Methods* **41**, 3198–3210.

Chaudhuri, A. (2011). *Randomized Response and Indirect Questioning Techniques in Surveys*. Chapman and Hall/CRC.

Chaudhuri, A. and Christofides, T. C. (2013). *Indirect Questioning in Sample Surveys*. Springer.

Fox, J. A. (2015). *Randomized Response and Related Methods, Surveying Sensitive Data*. 2nd Edition. SAGE Publications, California.

Fox, J. A. and Tracy, P. E. (1986) *Randomized Response, A method for Sensitive Surveys*. SAGE Publications, California.

Garza, A. (2012, December 21). Smart drugs are dumb. *The South Texan*.

Gjestvang, C. R. and Singh, S. (2006). A new randomized response model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 523–530.

Gjestvang, C. R. and Singh, S. (2009). An improved randomized response model: Estimation of mean. *Journal of Applied Statistics* **36**, 1361–1367.

Greenberg, B. G., Abul-Ela, A. L. A., Simmons, W. R. and Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association* **64**, 520–539.

Lanke, J. (1976). On the degree of protection in randomized interviews. *International Statistical Review* **44**, 197–203.

Lee, C.-S., Sedory, S. A. and Singh, S. (2013). Simulated minimum sample sizes for various randomized response models. *Communications in Statistics–Simulation and Computation* **42**, 771–789.

Lee, C. S., Su, S. C., Mondragon, K., Salinas, V. I., Zamora, M. L., Sedory, S. A. et al. (2016). Comparison of Cramer-Rao lower bounds of variances for at least equal protection of respondents. *Statistica Neerlandica* **70**, 80–99.

Odumade, O. and Singh, S. (2009). Efficient use of two decks of cards in randomized response sampling. *Communications in Statistics-Theory and Methods* **38**, 439–446.

Sabawi, F. (2012, August 30). Students turn to 'Smart Drugs' for help. *The South Texan*.

Singh, S. (2014). Randomized response techniques. *Model Assisted Statistics and Applications* **9**, 1–2.

Singh, S. and Kim, J. K. (2011). A pseudo-empirical log-likelihood estimator using scrambled responses. *Statistics and Probability Letters* **81**, 345–351.

Singh, S. and Sedory, S. A. (2011). Cramer-Rao lower bound of variance in randomized response sampling. *Sociological Methods and Research* **40**, 536–546.

Singh, S. and Sedory, S. A. (2012). A true simulation study of three estimators at equal protection of respondents in randomized response sampling. *Statistica Neerlandica* **66**, 442–451.

Sky News (2013, September 28). 'Smart Drug' Modafinil Risks Student Health. *Sky News*. `https://news.sky.com/story/smart-drug-modafinil-risks-student-health-10433081`.

Su, S.-C. (2013). *On Protection and Efficiency of Randomized Response Strategies*. Unpublished MS Thesis Submitted to the Department of Mathematics, Texas A & M University-Kingsville, Kingsville, TX.

Su, C.-S., Sedory, S. A. and Singh, S. (2014). Kuk's model adjusted for protection and efficiency. *Sociological Methods and Research* **44**, 534–551.

Su, C.-S., Sedory, S. A. and Singh, S. (2017). Adjusted Kuk's model using two non-sensitive characteristics unrelated to the sensitive characteristic. *Communications in Statistics, Theory and Methods* **46**, 2055–2075.

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

Shu-Ching Su
Texas A &M University-Kingsville, Kingsville, TX 78363, USA.
E-mail: cynthiadob@yahoo.com.tw
Veronica I. Salinas
Texas A &M University-Kingsville, Kingsville, TX 78363, USA.
E-mail: veronicas0018@gmail.com
Monique I. Zamora
Texas A &M University-Kingsville, Kingsville, TX 78363, USA.
E-mail: moniquezamora_23@yahoo.com
Stephen A. Sedory
Texas A &M University-Kingsville, Kingsville, TX 78363, USA.
E-mail: stephen.sedory@tamuk.edu
Sarjinder Singh
Texas A &M University-Kingsville, Kingsville, TX 78363, USA.
E-mail: sarjinder.singh@tamuk.edu

(Received April 2016; accepted January 2020)