Statistica Sinica Preprint No: SS-2016-0177	
Title	Randomized Response Sampling with Applications to Tracking Drugs for Better Life
Manuscript ID	SS-2016-0177
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0177
Complete List of Authors	Shu-Ching Su, Veronica I. Salinas, Monique Zamora, Stephen A. Sedory and Sarjinder Singh
Corresponding Author	Sarjinder Singh
E-mail	sarjinder.singh@tamuk.edu
Notice: Accepted version subject to English editing.	

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

Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX, USA

Abstract: **Warner (1965) proposed an interviewing technique, called randomized response, designed to protect an interviewee's privacy and to reduce a major source of bias (evasive answers or refusing to respond) when estimating the prevalence of sensitive characteristics by means of surveys of human populations. The objective of this paper is to introduce a new method in the field of randomized response sampling that could be used for tracking the addictions of people to various substances. Sky News (2013), United Kingdom, suggests that students who use the smart drug 'modafinil' are potentially putting their health at risk. A few studies of similar addictions, based on handson experience with the newly proposed technique, are discussed.**

Key words and phrases: Randomized response techniques, estimation of proportion, smart drug users.

1. Introduction

The estimation of the relative size of a certain subgroup of a population under study is one of the most important tasks in statistical surveys. When the question about membership in the subgroup is sensitive, as for example, whether a student belongs to the grouptaking drugs, the direct question on the subject usually suffers from non-negligible nonresponse (or false response). This is the point where indirect questioning designs like the 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, although one has to depart from the customary path of asking information directly. Some examples of the collection of data through 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 two mutually exclusive groups: one group with a stigmatizing/ sensitive characteristic and the other group without it. For estimating π_A, the proportion of respondents in the population Ω belonging to the sensi-

tive group, a simple random sample s of n respondents is selected using with replacement sampling from the population. To collect the information on the sensitive characteristic Warner (1965) made use of 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. 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 make use of some unrelated non-sensitive characteristic in the randomization device, as suggested by Greenberg, Abul-Ela, Simmons, and Horvitz (1969). They proposed the unrelated questions model. In their model, the respondent should answer one of two questions, which are not related to each other. 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 8?" Again each respondent selected in the sample uses a device like a deck of cards to determine the question to which they respond.

Let π_A be the true proportion of respondent in the population who possess the sensitive characteristic A. Also let π_Y be the true proportion of respondents in the population who possess non-sensitive characteristic, say Y. This method also 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:

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

If π_Y is known, then Greenberg, Abul-Ela, Simmons, and Horvitz (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.

If π_Y is unknown, then they suggested taking two independent samples of sizes n_1 and n_2 such that $n_1 + n_2 = n$. In the first sample, of 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, of n_2 respondents, they suggest using another independent randomization device, with associate probability T, such that the probability of a 'Yes' answer becomes:

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

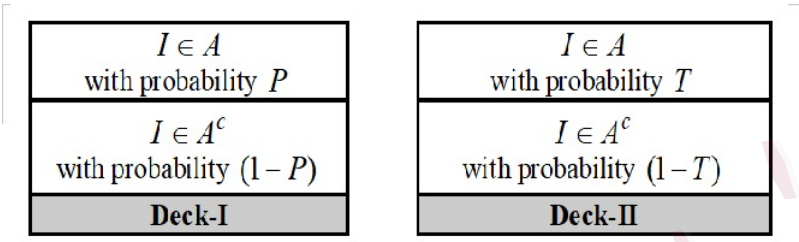

Figure 1: Two decks of cards

Greenberg, Abul-Ela, Simmons, and Horvitz (1969) solved these two linear equations for π_A and developed an unbiased estimator $\hat{\pi}_{G_2}$ given by:

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

The minimum variance of the estimator $\hat{\pi}_{G_2}$, using optimal values of 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 an efficient use of two decks of cards proposed by Odumade and Singh (2009). Each respondent in the simple random and with replacement (SRSWR) sample of 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 simultaneously, one card from each deck, and read the statements in order. The respondent first matches his/her status with the statement written on the card drawn

from the first deck, and 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 that possesses the characteristic A. Consider a situation where 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 is requested to report: (Yes, Yes). Consider another situation the selected respondent belongs to group A^c : If he/she draws a card with the tatement $I \in A^c$, 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 is also requested to report: (Yes, Yes). Thus the response (Yes, Yes) can come from both types of respondents and hence their privacy will be maintained. Thus the probability of a (Yes, Yes) response is given by

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

Likewise, 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. Odumade and Singh (2009) defined the least square 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. Setting $\frac{\partial D_1}{\partial \pi_A} = 0$, 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 $\hat{\pi}_{OS}$ is given by

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

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

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

which is same variance one would obtain if each respondent were requested to use the Warner (1965) device twice.

Greenberg, Abul-Ela, Simmons, and Horvitz (1969)'s second model, with unknown value of π_Y , is more practical in increasing the respondents' cooperation. However it requires two independent samples, which makes it complicated to apply in real practice. Also the optimum values of the sample sizes depend on the value of the population proportion of the sensitive characteristic being estimated. Recent studies by several authors, e.g. Lee, Sedory and Singh (2013), Abdelfatah, Mazloum and Singh (2013), Arnab, Singh and North (2012), Singh and Sedory (2011), Singh and Sedory (2012), Singh and Kim (2011), and several papers in the special issue on Randomized Response Sampling by Singh (2014) etc., have paid little attention to unrelated question models. This has motivated us to consider improvements to this unrelated question model when π_Y is unknown.

In this paper, section 1 presents the theoretical background for the paper; in section 2 we construct a new unrelated question model estimator using least squared distance and also construct maximum likelihood

estimators; section 3 establishes the equivalence of the least squared and maximum likelihood estimates in addition to determining the greater efficiency of the newly proposed unrelated question model over the Greenberg, Abul-Ela, Simmons, and Horvitz (1969) unrelated question model through simulation; section 4 illustrates an application of the proposed estimator to a survey conducted at Texas A & M University-Kingsville; section 5 provides black-box estimates for comparisons; section 6 illustrates a second application of the proposed model at a conference, and section 7 concludes the findings.

2. Proposed New Unrelated Question Model

Let A denote the set of members of Ω possessing the sensitive attribute, and Y the set of those with the unrelated, non-sensitive attribute. For a population Ω , under study it is clear that some members possess only the sensitive attribute, some possess only the non-sensitive 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} the population proportion of the people

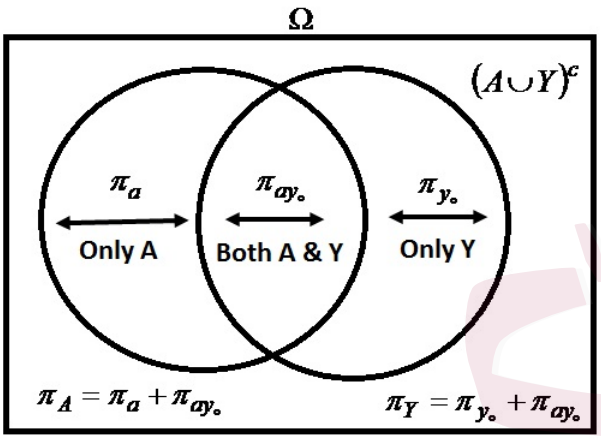

Figure 2: Pictorial representation of the population under study.

possessing only the non-sensitive characteristic $y_0 = Y \cap A^c$; π_{ay_0} the population proportion of people possessing both sensitive and non-sensitive 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 $\pi_A = \pi_a + \pi_{ay_0}$; the proportion π_A of people possessing the sensitive characteristic is the sum of the proportion π_a of people possessing only sensitive characteristic A and the proportion π_{ay_0} of people possessing both sensitive characteristic A and the non-sensitive characteristic Y. Similarly, $\pi_Y = \pi_{y_0} + \pi_{ay_0}$; the proportion π_Y of people possessing non-sensitive characteristic Y is a sum of the proportion π_{y_0} of people possessing only non-sensitive characteristic Y and the proportion π_{ay_0} of people possessing both sensitive characteristic A and the non-sensitive characteristic Y .

We consider selecting a simple random and with replacement sample 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 printed on these cards. Let P be the proportion of cards bearing the questions: "(i) Do you possess characteristic A?" and $(1-P)$ be the proportion of cards bearing questions: "(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), or (Yes, No), or (No, Yes), or (No, No). The same process is repeated with all n respondents selected in the sample. The probabilities of getting (Yes, Yes), (Yes, No), (No, Yes)

and (No, No) responses are, respectively, given by:

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

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

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

and

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

We note that these 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 ultimately 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 from the n respondents selected in the sample. Following Odumade and Singh (2009), we define a 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 with respect to the three parameters of interest π_a, π_{ay_0} and π_{y_0}. The motivation to consider the minimization of D is that it leads to simple, unbiased and closed form estimators of the three required proportions.

Now we set

$\frac{\partial D}{\partial \pi_a} = 0$, $\frac{\partial D}{\partial \pi_{ay_0}} = 0$, and $\frac{\partial D}{\partial \pi_{y_0}} = 0$.

The solution to the resulting system of linear equations leads, by the method of moments, to the following three estimators:

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

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

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 an estimator of the required population proportion π_A as :

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

or equivalently

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

Note that π_{y_0} is not of interest, so we do not investigate further any property of its estimator $\hat{\pi}_{y_0}$. For the derivations of (2.6), (2.7), (2.8) and (2.10) please see Appendix-A in Online Supplementary Document. 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:*

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

Proof. See Appendix A in Online Supplementary Document.

Remark 2.1: An estimator of variance $V(\hat{\pi}_A)$ of the estimator $\hat{\pi}_A$ is suggested as:

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

Remark 2.2: In this remark, we consider a likelihood function based on the probability mass function of the number of observed responses as given by:

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

On taking ln on both sides, 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) is the same as the least square distance estimates. One can aslo refer to Appendix-A in Online Supplementary Document.

In the next section, we consider a comparison of the proposed estimators with the Greenberg, Abul-Ela, Simmons, and Horvitz (1969) estimator when the population proportion of the non-sensitive 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, Abul-Ela, Simmons, and Horvitz (1969) estimator $\hat{\pi}_{G_2}$ by:

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

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 $\hat{\pi}_A^{mle}$ by:

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

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

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

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

Lee, Su, Mondragon, Salinas, Zamora, Sedory, and Singh (2016) utilized the measure of privacy protection to suggest a generalization of the Lanke (1976) privacy protection measure given by $L = Max[P(A|Yes), P(A|No)]$. They proposed a new measure of protection of a respondent while using the two decks model proposed by Odumade and Singh (2009). For the Odumade and Singh (2009), or equivalently Singh and Sedory (2011), Singh and Sedory (2012), method of using of two decks, Lee, Su, Mondragon, Salinas, Zamora, Sedory, and Singh (2016) compute four conditional probabilities as follows: $P[A|(Yes, Yes)] = \frac{PT\pi_A}{\lambda_{11}}$; $P[A|(Yes, No)] = \frac{P(1-T)\pi_A}{\lambda_{10}}$; $P[A|(No, Yes)] = \frac{(1-P)T\pi_A}{\lambda_{01}}$ and $P[A|(No, No)] = \frac{(1-P)(1-T)\pi_A}{\lambda_{00}}$. Then the least protection in the Odumade and Singh (2009) is given by:

$$Prot(OS\ Model) = Max[P\{A|(Yes, Yes)\}, P\{A|(Yes, No)\}, P\{A|(No, Yes)\}, P\{A|(No, No)\}] \tag{3.4}$$

In the case of the proposed unrelated questions model, we also compute the same four conditional probabilities as follows: $P^*[A|(Yes, Yes)] = \frac{PT\pi_a + \pi_{ay0}}{\theta_{11}}$; $P^*[A|(Yes, No)] = \frac{P(1-T)\pi_a}{\theta_{10}}$; $P^*[A|(No, Yes)] = \frac{(1-P)T\pi_a}{\theta_{01}}$; and $P^*[A|(No, No)] = \frac{(1-P)(1-T)\pi_a}{\theta_{00}}$;

Then the least protection in the proposed unrelated question model is given by:

$$Prot(Proposed\ Unrelated\ Model) = Max[P^*\{A|(Yes, Yes)\}, P^*\{A|(Yes, No)\}, P^*\{A|(No, Yes)\}, P^*\{A|(No, No)\}] \tag{3.5}$$

The relative protection of the proposed unrelated question model over the Odumade and Singh (2009) 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 case of Greenberg, Abul-Ela, Simmons, and Horvitz (1969) two-samples model, we compute the least protection level as:

$$Prot(Greenberg\ Model) = Max[P_1\{A|(Yes)\}, P_1\{A|(No)\}, P_2\{A|(Yes)\}, P_2\{A|(No)\}] \tag{3.7}$$

where

$P_1[A|(Yes)] = \frac{\{P+(1-P)\pi_Y\}\pi_A}{\theta_1}$; $P_1[A|(No)] = \frac{(1-P)(1-\pi_Y)\pi_A}{1-\theta_1}$;

$P_2[A|(Yes)] = \frac{\{T+(1-T)\pi_Y\}\pi_A}{\theta_2}$; $P_2[A|(No)] = \frac{(1-T)(1-\pi_Y)\pi_A}{1-\theta_2}$;

The relative protection of the proposed unrelated question model over the Greenberg, Abul-Ela, Simmons, and Horvitz (1969) model is defined as:

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

Note that the values of the relative efficiencies defined in (3.1)-(3.3) and the relative protections defined in (3.6) and (3.8) are free from the value of sample size. We wrote SAS codes, as given in Appendix-A in Online Supplementary Document, for computing the relative efficiency and relative protection for various values of parameters. In the study, 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 choice of π_{y_0} such that the value of $RP(OS) > 1$ and $RE(OS) > 1$. The

results obtained are presented in Table 9 in Appendix-A in Online Supplementary Document. From Table 9, note that $RE(1) = RE(2)$, that is, the proposed estimator $\hat{\pi}_A$ attains the lower bound of 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 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, Abul-Ela, Simmons, and Horvitz (1969) estimator $\hat{\pi}_{G_2}$ but remain sometimes 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 single trial question is sometime more protective than the two trials per respondent question model, but remains drastically less efficient than the two trial model indicated by $RE(1) = RE(2)$ values for different situation considered. From the simulation study, we conclude that there are choices for the proportion in a population of an unrelated characteristic π_Y such that the proposed unrelated question model can perform at least as good as the Odumade and Singh (2009) model, from both the protection and relative efficiency point of views.

We would like to mention that the protection criterion cannot be implemented on a given subject to determine whether he/she is a member of the

sensitive group or not. As suggested by Greenberg, Abul-Ela, Simmons, and Horvitz (1969) that choose π_Y close to π_A so that their model performs well. Note that if $\pi_Y \approx \pi_A$ then $P_1(A|Yes) = \frac{\{P+(1-P)\pi_Y\}\pi_A}{\theta_1} \approx P + (1-P)\pi_Y$ which is free from the value of θ_1. Further note that these protection criterion divide the people into two groups – those who responded "Yes" and those who responded "No", that is, the respondents who reported "Yes" may be considered in sensitive group with some conditional probability, and those who reported "No" may also be considered to be in the sensitive group, with different conditional probability. Thus a protection criterion cannot be used for classifying respondents into two groups, *viz.* sensitive or non-sensitive groups.

In the next section, we consider an application of the proposed unrelated question model to the investigation of the prevalence of use of smart drugs at Texas A & M University-Kingsville.

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

Sky News (2013) United Kingdom, suggests that students who use the smart drug 'modafinil' are potentially putting their health at risk. Sabawi's (2012) article also highlights students' bad 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 consisted of 51 cards with 35 cards bearing the question, "Have you ever once used any 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 8?" Thus $P = 0.686$ for the Green-Deck. The Pink-Deck was also made of 51 cards with 16 cards bearing the question, "Have you ever once used any smart drug in your college career?" and the remaining 35 cards were bearing the statement, "Is the last digit of your K-ID number greater than or equal to 8?" 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, and was of at least 18 years old, was asked to first draw a card from the green deck. They were told to read the question on the drawn card silently and answer it truthfully. The card was returned to the green deck without showing it to anyone. In the same way, the same student was also asked to draw a card from the pink deck, read the question silently, and respond honestly. Lastly, the card was returned to the pink deck. The response of every student participating in the survey was recorded on a response card as: **Gender:**

Table 1: Response from undergraduates.

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

Male or Female; **Seniority:** UG or G; **Response:** (Yes, Yes) or (Yes, No), or (No, Yes), or (No, No). The symbol "G" was used to indicate graduate students and "UG" was used for undergraduates. No other information was collected from any student who participated in the survey. An incentive of chocolate and candies 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. The overall responses of the 127 undergraduate students were classified into a 2x2 contingency Table 1.

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

The 2x2 contingency Table 2 shows the observed responses from the 63

Table 2: Response from male students.

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

Table 3: Response from female students.

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

males who participated in the survey.

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

Similarly the 2x2 contingency Table 3 shows the observed responses from 64 females who participated in the survey.

The estimate of proportion of female students have ever used smart drug in their career is 0.1563 with a standard error of 0.06615. The 95% confidence interval estimate is $(0.02659, 0.2859)$. It may be worth pointing out that we computed standard errors using 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 as follows: Each respondent was also given the card to provide information on: **Gender**: Male or Female; **Seniority:** UG or G; and a direct question: Have you ever once used any smart drug in your college career? **Response:** Yes or No. The respondent was requested to circle his/her response and leave such a card in a locked black-box without showing his/her response to the interviewers. A black-box technique will only be effective in inducing interviewee to answer honestly if the respondents have confidence that the interviewer does not know the content of the box before and after he/she has given their response. It is helpful to compare this "almost direct question" technique and the proposed randomized response technique. The black box was locked to assure the respondents of the anonymity of his/her response

Table 4: Black-box responses from undergraduates.

Overall	Yes	No	Total	Prop
Yes	7	3	10	0.1339
No	3	51	54	0.1429
Sum	10	54	64	01250

honestly. Table 4 shows the black box responses from 127 students.

These black box responses provide an 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 of smart drug users obtained at Texas A & M University is given in Figure 3.

All estimates (overall, males and females) 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 in comparison to the black box estimate of 0.1339; for males the randomized response estimate is 0.1696 and the black box estimate is 0.1429; and for females the randomized response estimate is 0.1563 and the black box estimate is 0.1250.

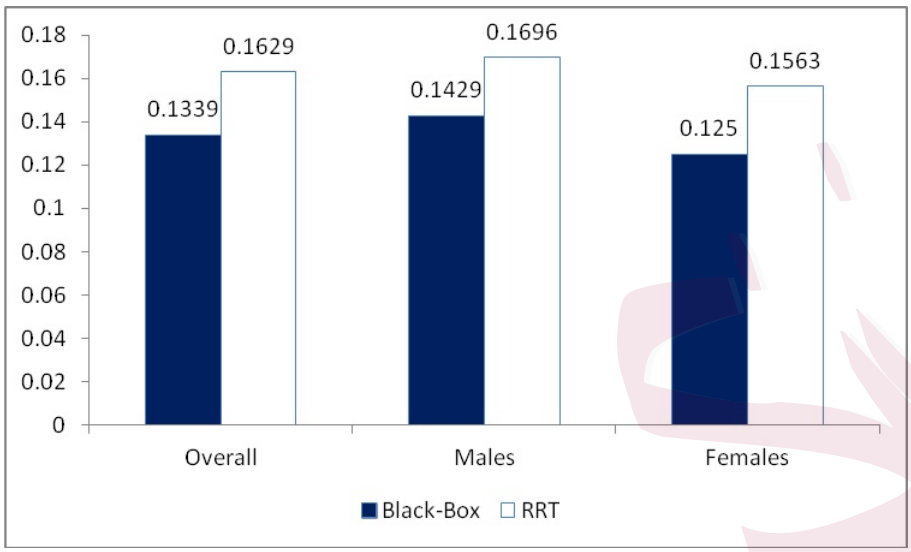

Figure 3: Estimates of smart drug users at TAMUK

We also calculated Z_{cal} as

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

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

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

and $\hat{V}(\hat{\pi}_A)$ is given in (2.12). The calculated value of Z_{cal} is 0.5007 in case of overall estimate, 0.3112 in case of male and 0.3997 in case of female estimate. 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 at Texas A&M University-Kingsville about the adverse effect of use of smart drugs on their life and career.

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

We conducted another convenient survey at the booth STAT-HAWKERS, during the Joint Statistical Meeting (JSM) 2013, Montreal, Canada. The purpose of this survey had two different objectives: (i.) To increase awareness of randomized response techniques to those statisticians who come to attend the conference, (ii) To estimate the prevalence of smart drug use among the preassembly smart community of statisticians. The data were collected for three days by 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 once used any smart drug in your career?" and the remaining 16 cards bearing the question, "Were you born on 1st, 2nd, 3rd, 4th, 5th or 6th of a month?" Thus $P = 0.686$ in the green deck.

Table 5: Overall randomized responses at the JSM.

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

The pink deck also consisted of 51 cards with 16 cards bearing the question, "Have you ever once used any smart drug in your college career?" and the remaining 35 cards were bearing the question, "Were you born on 1st, 2nd, 3rd, 4th, 5th or 6th of a month?" 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. During three days efforts, it was possible to collect data from 95 participants. The overall responses of 95 responses were classified into a 2x2 contingency Table 5.

By using the proposed method, we estimated that the proportion of conference attendees who have ever used smart drug in their career is 0.092417 with a standard error of 0.05599. The 2x2 contingency Table 6 shows observed responses from the 50 males who participated in the survey.

Table 6: Males randomized responses at the JSM.

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

Table 7: Females randomized responses at the JSM.

Overall	Yes	No	Sum
Yes	1	3	4
No	6	35	41
Sum	7	18	45

By using the proposed estimator, we estimate that the proportion of male conference attendees who have ever used smart drug in their career is 0.1463 with a standard error of 0.070995. The 2x2 contingency Table 7 shows the observed responses from 45 females who participated.

By using the proposed estimator, we estimate that the proportion of female conference attendees who have ever used smart drug in their career is 0.032616 with a standard error of 0.087355. A pictorial presentation of

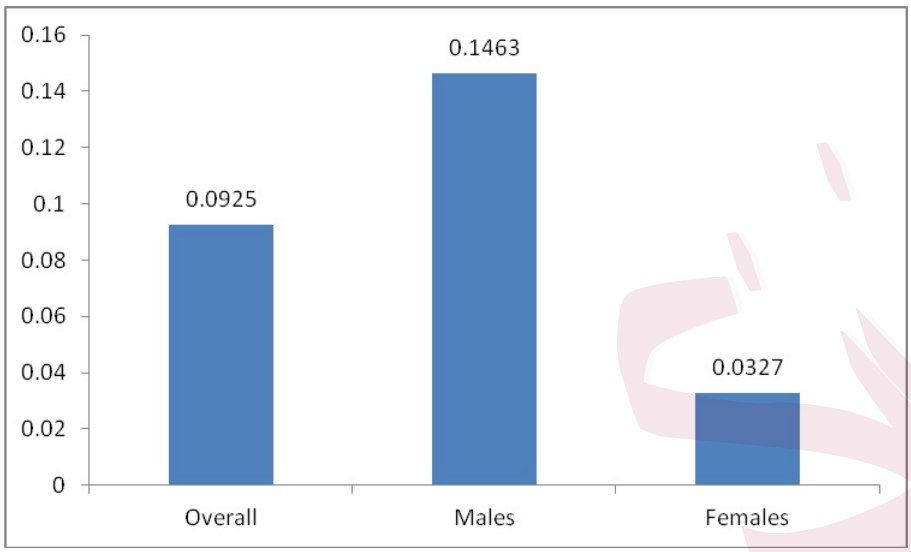

Figure 4: Estimates of smart drug users at the conference.

conference attendee estimates of smart drug users is given in Figure 4.

Thus based on our three evidences: theoretical, simulations, and real data applications; we conclude that the proposed unrelated question model should be more efficient than its competitors when used in real large scale-surveys where sensitive questions are being investigated by a social scientist.

The following remarks are answers to some general questions raised one of the reviewers:

Remarks 6.1: (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, so to make the proposed

Table 8: More 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

estimator efficient if one chooses P between 0.5 and 1.0 then T should be in its complement between 0.0 and 0.50. It may be worth pointing out that one should must check for other choice of parameters by executing the SAS macro given in Appendix-A if the proposed model is working efficiently for a new survey or not based on a good guess the proportion of sensitive attribute and unrelated attribute being used in the survey. For example consider, keeping $\pi_a = 0.05$, $\pi_{y_0} = 0.70$, and $\pi_{ay_0} = 0.02$, Table 8 gives different results for different choices of P and T.

(b) One could increase the use of number of decks in a survey, but sometime respondents are found hesitant in responding several times to the same question thus use of two-decks seems more appropriate to avoid

refusals.

7. Conclusion

This paper considers the design problem of randomized response sampling when the survey is to estimate the prevalence of some 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 and a respondent is asked to randomly select a question (one card from the deck) and answer the question without showing the question to the interviewer. To further protect the confidentiality of a respondent, Greenberg, Abul-Ela, Simmons, and Horvitz (1969) extends Warner's method to the design with two unrelated questions. This paper further extends Greenberg, Abul-Ela, Simmons, and Horvitz (1969) approach to the "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 relative efficiency with respect to the existing approaches is also established. Further, note that we

have considered the SRSWR sampling scheme, because the proposed model is compared to other existing models which assume the same scheme. Also if the finite population correction factor is small SRSWR and SRSWOR designs perform almost the same. If required, one can extend the proposed unrelated question model to complex survey designs by following Arnab, Singh and North (2012).

Acknowledgements The authors are thankful to the Editor-in-Chief Dr. Zhiliang Ying, Admin: Anna Chiang, and a learned referee for their very constructive comments which lead to substantial improvement on the original manuscript. Thanks are due to the IRB Chair Dr. Stephen D. Oller and 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 during the completion of 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.

REFERENCES

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**: 16-17, 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.*, Second 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). Smart drugs are dumb. The South Texan, December 21, 2012., *http://www.southtexannews.com/opinions/4857-smart-drugs-are-dumb.*

Gjestvang, C. R. and Singh, S. (2006). A new randomized response model, *J. Roy. Statist. Soc.*, **B**: 68, 523-530.

Gjestvang, C. R. and Singh, S. (2009). An improved randomized response model: Estimation of mean., *Journal of Applied Statistics*, **36**: 12, 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., *Int Stat Rev*, **44**: 197-203.

Lee, Cheon-Sig, Sedory, S.A. and Singh, S. (2013). Simulated minimum sample sizes for various randomized response models. , *Communications in Statistics-Simulation and Computation*, **42**(4): 771-789.

Lee, C.S., Su, S.C., Mondragon, K., Salinas, V.I., Zamora, M.L., Sedory, S.A. and Singh, S. (2016). Comparison of Cramer-Rao lower bounds of variances for at least equal protection of respondents. , *Statistica Neerlandica*, **70**(2), 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, Fares (2012). Students turn to 'Smart Drugs' for help., *The South Texan, August 30, 2012*

Singh, S. (2014). Randomized Response Techniques., *Model Assisted Statistics and Applications*, **9**(1): 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**(3): 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**(4): 442-451.

Sky News (2013). 'Smart Drug' Modafinil Risks Student Health., *Sky News: September 28, 2013, United Kingdom.*

Su, Shu-Ching (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, Ching-Shu, Sedory, S.A. and Singh, S. (2014). Kuk's model adjusted for protection and efficiency., *Sociological Methods and Research,***44**(3): 534-551.

Su, Ching-Shu, 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**(4), 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

REFERENCES

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 ???? 2016; accepted ???? 20??)