Survey Methodology, December 1989
Vol. 15, No. 2, pp. 225–235
Statistics Canada

Randomized Response Sampling from Dichotomous Populations with Continuous Randomization

LeROY A. FRANKLIN[1]

ABSTRACT

A randomized response model for sampling from dichotomous populations is developed in this paper. The model permits the use of continuous randomization and multiple trials per respondent. The special case of randomization with normal distributions is considered, and a computer simulation of such a sampling procedure is presented as an initial exploration into the effects such a scheme has on the amount of information in the sample. A portable electronic device is discussed which would implement the presented model. The results of a study taken, using the electronic randomizing device, is presented. The results show that randomized response sampling is a superior technique to direct questioning for at least some sensitive questions.

KEY WORDS: Randomized response; Randomization with continuous distributions; Computer simulation.

1. INTRODUCTION

Surveys often seek to estimate the proportion of individuals satisfying a particular condition. If the condition involves a highly personal or controversial subject (*e.g.*, seeking new employment, sexual behavior) or of an illegal nature (*e.g.* drug usage, criminal activities), survey respondents may be reluctant to answer honestly or may refuse to answer a direct question as to whether they satisfy the condition of interest. In such cases, it is difficult to make inferences about proportions on the basis of a survey in which sensitive questions are asked directly.

Randomized response sampling plans utilize a stochastic or randomizing device to enable respondents to provide answers to sensitive questions without fully revealing information regarding the sensitive issue. The actual outcome of the device for a particular respondent is observed by the respondent but not by the interviewer. However, the properties of the device are known to the experimenter, and this enables the experimenter to make inferences about the proportion of interest without knowing specifically about any single individual. The stochastic device introduces noise into the information-gathering process, but the resulting loss of information may be preferable to the uncontrollable noise introduced by nonresponse or lying when direct questions are used.

The original randomized response model was proposed by Warner (1965) and involved a dichotomous randomization for a dichotomous population. His model was studied from a Bayesian viewpoint in Winkler and Franklin (1979). The randomized response model with two or more trials per respondent was introduced by Gould, Shah and Abernathy (1969) and further developed by Liu and Chow (1976). Both papers demonstrated the superiority of the multiple trials per respondent in improving the efficiency of the estimate over the single trial model of Warner's. However, both also note that multiple trials might produce simultaneously

[1] Dr. LeRoy A. Franklin, Department of System and Decision Sciences, Indiana State University, School of Business, Terre Haute, Indiana 47809.

growing suspicion and lowered "truth telling" over the single trial model. The survey paper prepared by Horvitz, Greenberg, and Abernathy (1976) discusses several other plans with discrete randomization devices. In addition a thorough theoretical development and review of results is contained in the recent volume by Chaudhuri and Mukerjee (1988) entitled "Randomized Response: Theory and Techniques." A more general model, using either discrete or continuous randomization, is presented in Warner (1971) and these more general models were discussed from a Bayesian viewpoint by Pitz (1980), Smouse (1984), and O'Hagen (1987). A few surveys have actually been undertaken, some showing the randomized response methods are superior to direct survey methods (*e.g.* Gould *et al.* 1969 and Liu and Chow 1976) and a few others of uncertain results (*e.g.* Brewer 1981). However, only Poole (1974) developed a specific continuous randomization distribution (uniform) to estimate a continuous distribution and this was implemented by having respondents report their answer multiplied by a number chosen randomly from a random number table.

In this paper, we consider a randomized response model for sampling from a dichotomous population, but using a continuous randomization distribution. With Warner's original randomized response technique, the randomizing device determines which question the respondent answers. But with the method developed in this paper, the question for a respondent is fixed by whether or not he belongs to the sensitive group. The randomization here chooses values from two distributions (one for "yes" and the other for "no") and the respondent provides the value appropriate to his group membership. Multiple trials are incorporated into the model by having the respondent provide a single multi-digit response. This provides a potential benefit over usual multiple trial techniques in that the respondent perceives he/she has provided just one answer when in fact the multi-digit response incorporates several trials of the respondent.

The general model, for which the randomization can be handled via any type of distribution, is presented in Section 2. The special case in which the randomization involves normal distributions is discussed in Section 3, along with an approximating procedure for assessing the effect of randomization and multiple trials per respondent. Section 4 presents a computer simulation investigating the role that specific choices of means and standard deviations play in the efficiency of surveying by using normal distribution randomization with multiple trials. Section 5 presents a way of implementing normal distributions as the randomizing distribution through the use of a computerized, electronic device that generates and displays random normal values. Such a device was felt to be potentially superior to "drawing cards" or "flipping a spinner" since these methods may not be properly implemented by the respondent or the interviewer. The results of a survey taken using that electronic device to investigate five sensitive questions are examined in Section 6. Finally, a summary and a brief discussion of design issues are considered in Section 7.

2. THE MODEL

Suppose that we are interested in θ, the proportion of individuals belonging to Group A among the members of a particular population. A simple random sample of n individuals is chosen from the population with $n \geq 1$, where we assume that the population is large enough relative to n so that the sampling process can be viewed effectively as sampling with replacement. A total of k trials are conducted with each respondent, where $k \geq 1$. On trial j for respondent i, random values are drawn from the distribution functions G_{ij} and H_{ij}. The respondent sees both values and is asked to report the value from G_{ij} if he or she belongs to Group A and the value from H_{ij} otherwise. The researcher knows the exact form of G_{ij} and H_{ij} but sees only the value reported by the respondent, denoted by z_{ij}, and, thus, does not know from which distribution it came.

Inferences must be made about θ based on the kn sample observations z_{ij}, with $i = 1, \ldots, n$ and $j = 1, \ldots, k$. For convenience, we assume in the remainder of this paper that G_{ij} and H_{ij} are absolutely continuous with corresponding densities g_{ij} and h_{ij}; the development for the discrete case is analogous. The conditional density function of z_{ij} given θ is $\theta \; g_{ij} \; (z_{ij}) + (1 - \theta) \; h_{ij} \; (z_{ij})$, and the likelihood function for the entire experiment is:

$$L(\underset{\sim}{z} \mid \theta) = \prod_{i=1}^{n} \left[\theta \prod_{j=1}^{k} g_{ij} \; (z_{ij}) + (1 - \theta) \prod_{j=1}^{k} h_{ij} \; (z_{ij}) \right] \text{ for } 0 \leq \theta \leq 1, \qquad (2.1)$$

where $\underset{\sim}{z} = (\underset{\sim}{z}_1, \ldots, \underset{\sim}{z}_n)$ and $\underset{\sim}{z}_i = (z_{i1}, \ldots, z_{ik})$.

Expanding the likelihood function using the binomial theorem allows the likelihood function to be written in the form

$$L(\underset{\sim}{z} \mid \theta) = \sum_{t=0}^{n} \alpha_t \, \theta^t \, (1 - \theta)^{n-t} \text{ where } 0 \leq \theta \leq 1 \text{ and} \qquad (2.2)$$

$$\alpha_t = \sum_{s=1}^{c} \left[\prod_{i \in C_{ts}} \prod_{j=1}^{k} g_{ij} \; (z_{ij}) \right] \left[\prod_{i \notin C_{ts}} \prod_{j=1}^{k} h_{ij} \; (z_{ij}) \right], \text{ with} \qquad (2.3)$$

$C_{t1}, \ldots, C_{tc}$ representing the $c = \binom{n}{t}$ combinations of t items out of n. Here $\theta^t(1 - \theta)^{n-t}$ is the Bernoulli likelihood conditional upon exactly t respondents being in Group A, and α_t is the likelihood of $\underset{\sim}{z}$ given t. The mixture form in 2.2 arises because we are unable to observe a specific t in our sample.

A special case of (2.1) arises when we assume that the same randomizing distributions are used for all n respondents. Thus, $g_{ij} = g_j$ and $h_{ij} = h_j$ for $i = 1 \ldots n$ and thus (2.1) reduces to

$$L(\underset{\sim}{z} \mid \theta) = \prod_{i=1}^{n} \left[\theta \prod_{j=1}^{k} g_i \; (z_{ij}) + (1 - \theta) \prod_{j=1}^{k} h_j \; (z_{ij}) \right] \text{ for } 0 \leq \theta \leq 1. \qquad (2.4)$$

Whichever the form, in order to find the maximum likelihood estimates, a direct computer grid search must be made. This is feasible since θ is only a one-dimensional quantity and is restricted to the interval from 0 to 1. This can be easily accomplished by using well-known search techniques applied to the log of the likelihood function. (See, for example, Kennedy and Gentle 1980).

3. RANDOMIZATION WITH NORMAL DISTRIBUTIONS

Although any continuous distribution (*e.g.* Weibull, uniform, *etc.*) can be used as the randomizing distribution in the model discussed in Section 2, in this section only the normal distribution will be examined. Furthermore, suppose that the same randomization distributions are used for all respondents, so that form (2.4) is the appropriate likelihood. Thus, g_j and h_j are normal densities with means μ_{gj} and μ_{hj} and standard deviations σ_{gj} and σ_{hj}, respectively. Then the likelihood function in Section 2 can be related to these normal densities.

The amount of information that can be obtained about θ obviously depends on the means and standard deviations that are chosen. At one extreme, if $\mu_{gj} = \mu_{hj}$ and $\sigma_{gj} = \sigma_{hj}$ for $j = 1, \ldots, k$, then θ drops out of the likelihood function and $\underset{\sim}{z}$ (the sample) will provide no information about θ. At the other extreme, if $| \mu_{gj} - \mu_{hj} | \to \infty$ for any j with σ_{gj} and σ_{hj} fixed **or** if $\sigma_{gj} \to 0$ and $\sigma_{hj} \to 0$ for any j with a fixed $| \mu_{gj} - \mu_{hj} | \neq 0$, then we are effectively able to determine which group each respondent belongs to and the sampling process thus approaches Bernoulli sampling in θ.

An approximation to $L(\underset{\sim}{z} \mid \theta)$ as developed by Winkler and Franklin (1979) makes it easier to assess the effect of randomization and multiple trials with the choice of specific means and standard deviations. That is, for each sample, we can approximate the actual likelihood function given by (2.4) with an approximate likelihood function of the form

$$L^*(r^*, n^* \mid \theta) = \theta^{r^*} (1 - \theta)^{n^* - r^*}. \qquad (3.1)$$

Taking the first and second derivations of the log of the approximating likelihood (3.1) and solving to find the maximum $(\hat{\theta})$ and the curvature at that maximum yields:

$$\hat{\theta} = \frac{r^*}{n^*} \qquad (3.2)$$

and

$$\left[\frac{\partial^2 \log L^*(r^*, n^* \mid \theta)}{\partial \theta^2} \right]_{\theta = \hat{\theta}} = - \frac{n^*}{\hat{\theta}(1 - \hat{\theta})}. \qquad (3.3)$$

Next taking the first derivative of the log of the exact likelihood (2.4) and setting it to equal zero gives the equation that will yield the exact maximum likelihood estimate for θ:

$$\sum_{i=1}^{n} \frac{\gamma_i - \eta_i}{\theta\gamma_i + (1-\theta)\eta_i} = 0 \text{ where } \gamma_i = \prod_{j=1}^{k} g_j\,(z_{ij}),\ \eta_i = \prod_{j=1}^{k} h_j\,(z_{ij}). \qquad (3.4)$$

A grid search produces for (3.4) its solution $(\hat{\theta}_r)$. Taking the second derivative of the log of the exact likelihood (2.4) yields:

$$\left[\frac{\partial^2 \log L\,(\underset{\sim}{z} \mid \theta)}{\partial \theta^2} \right] = - \sum_{i=1}^{n} \frac{[\gamma_i - \eta_i]^2}{[\theta\gamma_i + (1 - \theta)\eta_i]^2}. \qquad (3.5)$$

Substituting $\hat{\theta}_r$ into (3.5) gives the curvature of the actual log likelihood at $\hat{\theta}_r$ (the maximum). Equations (3.2) and (3.3) are two equations in two unknowns, r^* and n^*. Setting (3.2) $= \hat{\theta}_r$ and (3.3) $=$ (3.5) allows us to solve for r^* and n^* so that the approximating log likelihood has the same maximum $\hat{\theta} = \hat{\theta}_r$, and curvature at that maximum as does the actual log likelihood. Thus, the randomized response sample outcome of $\underset{\sim}{z}$ can be thought of as approximately equivalent to a non-randomized response sample (*i.e.* regular Bernoulli sampling) with r^* members out of n^* in the sensitive group. In this sense, n^* can be thought of as a rough measure of the amount of information in the randomized response sample which is of size n.

4. A COMPUTER SIMULATED INVESTIGATION OF THE CHOICE OF MEANS AND STANDARD DEVIATIONS

To investigate the impact of a given set of means and standard deviations for the normal randomizing distributions as well as the impact the size of θ and k (the number of trials) has upon r^* and n^* the randomized response sampling process was simulated by generating, via computer, repeated samples from a Bernoulli process with parameter θ and k sets of two-digit responses for each sample. In our simulation, we let $\mu_{gj} = 50$, $\mu_{hj} = 40$, and $\sigma_{gj} = \sigma_{hj} = \sigma$ for $j = 1, \ldots, k$. We considered two values of θ (.10 and .25), two values of σ (6 and 9), three values of n (50, 200, and 500), and three values of k (1, 2, and 3). Such values were chosen since they will register two-digit deviates that would overlap in distribution considerably and provided then a bench mark for later choices in the actual survey environment. For each of the 36 combinations of parameters, we replicated the sampling procedure 25 times. The solutions of r^* and n^* were found numerically for each sample, and the average values of n^* for the 25 replications with each set of parameter values are given in Table 1.

The average values of n^* vary considerably. At the worst extreme, when $\sigma = 9$, $\theta = .10$, and only one trial per respondent is used, n^* tends to be only 10-15 percent of n. On the other hand, when $\sigma = 6$, $\theta = .25$, and three trials are used per respondent, n^* is about 75 percent of n. As expected, the average value of n^* (the effective sample size) increases as n (the number of respondents) increases or as k (the number of trials per respondent) increases. In addition, decreasing σ or increasing θ also leads to a higher n^*.

For each combination of parameters, the mean and variance of $\hat{\theta}$ over the 25 trials were determined. The average values of $\hat{\theta}$ are very close (within 5%) to the corresponding values of θ, and the variance of $\hat{\theta}$ tends to increase as the average n^* decreases and, hence, tends to validate the simulation.

Table 1

Average Values of the Effective Sample Size (n^*) for Various Sample Sizes (n) and the Number of Trials per Respondent (k)

		$\theta = .10$		$\theta = .25$	
n	k	$\sigma = 6$	$\sigma = 9$	$\sigma = 6$	$\sigma = 9$
50	1	16.2	7.0	17.3	9.2
	2	27.3	13.1	30.6	17.8
	3	32.6	18.1	38.2	23.6
200	1	58.3	24.8	79.0	41.2
	2	103.1	49.6	124.4	72.9
	3	136.6	77.7	151.0	97.7
500	1	148.4	59.6	196.9	103.6
	2	261.1	129.3	309.5	181.2
	3	345.8	193.1	375.6	242.7

5. A PORTABLE, COMPUTERIZED RANDOMIZING DEVICE

Randomized-response sampling, using randomization with normal distributions and multiple trials, provides flexibility to the experimenter, who can select means and variances as well as the number of respondents and the number of trials per respondent. However, this flexibility is not of any value, unless the sampling scheme actually can be implemented in practice. The sampling scheme utilizing Bernoulli randomization can be implemented in a number of ways (*e.g.*, with cards or colored beads). However, the scheme developed in this paper requires generation of random normal values by some portable device.

A computerized, electronic device was built around the Intel 8080 microprocessor to generate and display random normal values. Each value is obtained by summing 16 uniformly distributed random numbers and transforming that sum to achieve a normal deviate with the desired mean and standard deviation. From the Central Limit Theorem, the resulting values should be approximately normally distributed, and extensive tests indicate that the values produced by the device do indeed behave like random normal values. This technique was chosen over other possible methods of generating normal deviates due to the simplicity of programming such a method in machine instructions for this specific microprocessor. For more details concerning the generation of the random normal values and the testing of the device, see Franklin (1977), Kennedy and Gentle (1980), as well as Knuth (1969).

The final, resulting device was approximately the size of a cigar box and is easily held in the hand. Power can be supplied either by a battery pack or by an extension cord.

For display purposes, the random normal values are truncated to two digits, and the device is designed to display six such two-digit numbers simultaneously in "windows" of six digits each. One window displays values chosen from g_1, g_2, and g_3 which appears as a single six-digit number in the "Yes" window. The other window displays values chosen from h_1, h_2, and h_3 which also appears as a single six-digit number for "No". The six means and standard deviations are stored permanently in the device, but they can be changed easily by using a small, detachable keyboard.

The actual surveying process is accomplished in the following manner. First, the interviewer asks the respondent a sensitive question about Group A. The respondent then pushes a button to activate the device, and two six-digit numbers appear in the windows within about one quarter of a second. If the respondent is a member of Group A, the number in the first window (the "Yes" window) is reported; otherwise, the number in the second window (the "No" window) is reported. To convince the respondent of the "randomness" of the values, he or she is encouraged to press the button several times and to observe the resulting numbers before the sensitive question is actually asked. Note that although $k = 3$, the respondent perceives a response as a single six-digit number, and we are thus actually obtaining three trials with a single six digit response. Hence, the advantage of multiple trials per respondent is exploited without the usual accompanying disadvantages coming into play.

6. SURVEY RESULTS AND CONCLUSIONS

Two simultaneous, but independent, surveys were conducted on the campus of a large urban university of students enrolled in that university. The first asked five sensitive questions of a respondent by the direct question method. The second asked the same five sensitive questions of a different respondent but using Randomized Response Sampling with continuous randomization implemented by the electronic device presented in the previous section. For the study $k = 3$ and $\mu_{g_1} = \mu_{g_2} = \mu_{g_3} = 40$ and $\mu_{h_1} = \mu_{h_2} = \mu_{h_3} = 50$ with $\sigma_{g_j} = \sigma_{h_j} = 5$ for $j = 1, 2, 3$. These values were chosen in accordance to the finding of the computer simulation discussed in Section 4. A different group of students was systematically selected (one in five) for each of the two surveys from students on the campus and individually interviewed. Each student surveyed was given a brief introduction as to the purpose of the survey and asked if they wished to participate. Less than 10% of all individuals stopped by both survey teams declined to participate. If the individual was willing to participate, he/she was then asked to provide his/her social security number to verify that he/she was, indeed, enrolled in the university. All respondents of both surveys had their social security number checked against an administrative master list of students and those not recorded as enrolled students were eliminated from the study (less then 5 percent of those surveyed).

Requiring their social security number also deliberately injected the element of associating the individual's identity with his responses. For many surveys (*i.e.* telephone, mail-in questionnaires, house-to-house surveys, *etc.*), this is the case and plays a significant role in the willingness of a respondent to answer truthfully. It was felt that it was precisely in such "revealing" circumstances that randomized response sampling can benefit the researcher most. The resulting sample sizes for the direct and randomized response methods were $n_1 = 473$ and $n_2 = 477$. The five sensitive questions were:

Q1 — "Have you ever cheated on an exam here at this university?"
Q2 — "Would you ever cheat on your income tax?"
Q3 — "Would you ever steal from an employer?"
Q4 — "Have you smoked any marijuana in the last 30 days?"
Q5 — "Have you ever participated in a homosexual act?"

All five questions were felt to be sufficiently sensitive so that any gains by randomized response sampling over direct sampling could be easily apparent. In addition, as a final question, the respondents in the randomized response group were asked "Do you think your friends would be more willing to tell the truth if they were asked sensitive questions by this technique?" This was asked in an effort to measure the acceptance and confidence of the person being interviewed that this particular randomized response technique did provide personal protection and anonymity.

The estimates of the proportion of respondents who are in the sensitive group are presented in Table 2 for both direct ($\hat{\theta}_{id}$) and randomized response ($\hat{\theta}_{ir}$) for question i along with the estimate of n_i^* (the effective sample size) for the randomized response method using the method discussed in Section 3. Also is presented the z value of a one-sided test of hypothesis H_o: $\theta_{id} - \theta_{ir} = 0$ vs H_a: $\theta_{id} - \theta_{ir} < 0$, along with the observed p-values. The tests were conducted using n_1 and n_i^* as sample sizes and hence give a much more conservative result than if n_1 and n_2 were utilized.

It is noteworthy that the randomized response method gave a higher estimate of θ for each of the five sensitive questions than the direct survey method. Furthermore, for Questions 1, 2, and 5, the randomization response method gave statistically significantly higher estimates of θ (p-values $< .001$ for all three) than the direct survey method. Hence, there seems to be conclusive evidence that, at least for some sensitive issues, the randomized response method with continuous randomization does provide better estimates of population proportions. It should also be noted that by our choices of μ_{g_j}, μ_{h_j}, σ_{g_j} and σ_{h_j} and $k = 3$ that n_i^* typically was 75 to 85 percent of the original sample size n_2 and thus most of the information was "recovered" by our randomized response method.

Table 2

Estimates of θ and Results of Testing Equality of θ's for Direct and Randomized Response Sampling with Respective Sample Sizes of $n_1 = 473$ and $n_2 = 477$

Question i	$\hat{\theta}_{id}$	$\hat{\theta}_{ir}$	Effective sample size n_i^*	z-value	p-value
1	.0634	.2013	394.5	6.098	<.0001
2	.1797	.2941	408.1	3.997	<.0001
3	.1078	.1207	384.8	.583	.2810
4	.1882	.1942	409.5	.234	.4091
5	.0042	.0355	339.0	3.341	.0004

Furthermore, it is instructive to consider the nonsignificant results for Questions 3 and 4. This information (if the three significant results are ignored) could lead an observer to conclude that randomized response techniques are not particularly advantageous over direct questioning. However, in the light of the three significant differences revealed, this lack of significance perhaps could be interpreted as the question really was not "sensitive enough" to lead to dramatic differences in θ's or even that the question was "so sensitive" that the respondent chose to lie even with the randomized response technique. In addition, Question 1 "Have you ever cheated on an exam?" seemed to the experimenter to be relatively "unsensitive" but in retrospect the answer to this question when tied to the social security number of the respondent (given before the questioning process started) presented a much more threatening circumstance than was initially realized. Thus, perhaps some of the confusion about the efficacy of the randomized response technique is related to the "true sensitivity" of the question for the interviewee as opposed to the "perceived sensitivity" by the interviewer or experimenter. These aspects need further examination.

Finally, 88.9% (424 of the 477) felt "their friends would be more likely to answer truthfully sensitive questions by this randomized response technique." While some reservations may be expressed by the respondents' "desire to please the interviewer," nevertheless, this overwhelming percentage coupled with the significant differences already discussed seem strong evidence that this technique was accepted and felt to be protective of the interviewee.

7. DISCUSSION

The model developed in this paper permits the use of continuous, as well as discrete, randomizing distributions in utilizing randomized response sampling from a dichotomous population. In order to implement the model using randomization with normal distributions, a computerized, electronic device was also developed and discussed. The device is portable, has programmable means and standard deviations for the six normal distributions and provides from a single six digit response, three separate two digit trials. Such a system has both potential advantages and disadvantages over other randomized response techniques.

First, as alluded to in the introduction, a computerized randomizing device could be superior to the standard randomized response methods of "drawing cards" or "flipping a spinner" since these methods may not be properly implemented by either the respondent or the interviewer which would induce uncontrolled error. (See Abernathy, Greenberg and Horvitz (1970) for a discussion of the problems of "insufficient card shuffling" and "card loss" as well as insufficient interviewer training). Since the production of the randomizing values is computerized, the distributional problems that can and have accompanied the use of cards, beads, and spinners are eliminated because the problem of "random selection of values" is taken out of the hands of the interviewer *and* respondent and placed in the "hands" of the computer. If the computerized device fails, it is usually a complete, catastrophic crash of the whole chip which is readily apparent and very, very rare.

The second (and perhaps greatest) advantage is in the ability of the device to present a choice of two numbers each six digits in length from which the respondent chooses to answer "yes" or "no". But what seems to the respondent as a single six digit answer is in fact three separate two digit answers and in effect provides three trials per respondent. Thus, the benefits of multiple trials per respondent are gained but, since the respondent is unaware of the multiple trials format, without the usual accompanying disadvantages (noted by Liu and Chow 1976) coming into play.

In addition, the freedom to choose the six means and six standard deviations provides the experimenter with additional flexibility over standard randomized response techniques. For instance, if it is felt that the differences in the first two digits are most noticeable to respondents, the experimenter can make μ_{h_1} and σ_{h_1} close to (or even equal to) μ_{g_1}, and σ_{g_1}, respectively. Similarly, if the middle two digits might receive the least attention, the experimenter could attempt to gain the most information from these values by separating μ_{h_2} and μ_{g_2} the furthest. It is also possible to wire the displays in other than the obvious manner. For instance, the two digits of the first random normal value could appear as the fifth and second digits of the six digit number instead of the first and second digits. This flexibility in wiring, together with the the choices of parameters should provide a sampling scheme that is quite informative to the researcher without seemingly to threaten the respondent.

It should also be noted that while for this particular microprocessor it was convenient to utilize randomization with normal distributions, several other continuous distributions (*e.g.* uniform, Weibull) or even multi-valued discrete distributions (*e.g.* multinomial or poisson) could have been used. Further investigation into newer microprocessors as well as different randomizing distributions is recommended.

There are, however, some potential disadvantages associated with this particular randomized response technique. The cost of such a randomizing device since it involves a microprocessor is the order of fifteen hundred to two thousand dollars to produce. However, its versatility in wiring and programming would hopefully allow a device to be used in many investigations over several years and thus help to defray its rather high cost.

More difficult to quantify is the respondent's perception of the computerized device and the degree of confidence or suspicion he/she might have about the device. Do respondents fear that the computerized device is somehow "storing" their answer that somehow later can be deciphered to expose them? From the survey results, it seems that greater truth telling was secured by using the computerized randomizing devices over the direct survey method. Nevertheless, further study is recommended to compare this randomized response technique which uses the computerized device with other more standard randomized response techniques.

In practice, several matters are relevant in the consideration of design issues (*i.e.*, the selection of means and standard deviations for the device). In order to gain more information for a given sample size, we should increase $|\mu_{g_j} - \mu_{h_j}|$ and decrease σ_{g_j} and σ_{h_j} for $j = 1, 2, 3$. However, as this is done, it will become clearer to the respondent that, despite the randomization, the response is very revealing concerning the respondent's group membership. As a result, the respondent may not answer honestly or may refuse to answer. Additional study is needed to determine optimal values for choice of means and standard deviations. The results in Table 1 give some indication of the effects of varying a common standard deviation. But from a practical viewpoint, the field survey seemed to indicate that the choice of means separated by two standard deviations was able to both gain the confidence of the respondent and (with the multiple trials) to gain back from 75 to 85 percent of the original sample size without the usual "loss of confidence" that accompanies multiple trial techniques.

In particular, the field trial compared the direct survey techniques with the randomized response using the electronic device discussed with $\mu_{h_j} = 40$ and $\mu_{g_j} = 50$ and $\sigma_{h_j} = \sigma_{g_j} = 5$ for $j = 1, 2, 3$ for the normal, randomizing distributions. Of the five sensitive questions which were asked of the two (independent) groups, the randomized response method provided significantly greater estimates ($p < .001$) than the direct method for three of the questions. In addition, 88.9% of the subjects interviewed by the randomized response technique felt "their friends would be more likely to tell the truth if they were asked sensitive questions by this technique". Thus, it seems that (for at least certain questions), this randomized response sampling technique achieved greater honesty in response than the direct sampling method.

The question of protection of the respondent's privacy needs to be discussed. It is not ethical to tell the respondent that his or her group membership is disguised by the randomization, if, in fact, the disguise is transparent to the researcher (*e.g.* for example, by recording only even numbers for "YES" and only odd numbers for "NO"). With the electronic device that has been discussed, it seems indeed possible to provide true privacy without losing much information. If the means and standard deviations are programmed into the device and are not provided to an interviewer, the interviewer will find it very difficult to discriminate between group members and non-group members in the interviewing process, particularly if the wiring is "scrambled". Thus, the flexibility that enables us to gain information without threatening the respondent also helps to disguise the actual group membership from the interviewer.

ACKNOWLEDGEMENTS

The author wishes to thank the referees and an Associate Editor for their constructive comments.

REFERENCES

ABERNATHY, J.R., GREENBERG, B.G., and HORVITZ, D.G. (1970). Estimates of induced abortion in urban North Carolina. *Demography*, 7, 19-29.

BARNARD, G.A. (1976). Discussion on the invited and contributed papers. *International Statistical Review*, 44, 226.

BREWER, K.R.W. (1981). Estimating marijuana usage using randomized response some parodoxical findings. *Australian Journal of Statistics*, 23, 139-148.

CAMPBELL, C., and JOINER, B.L. (1973). How to get the answer without being sure you've asked the question. *The American Statistician*, 27, 229-231.

CHAUDHURI, A., and MUKERJEE, R. (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker, Inc..

CHOW, L.P., LIU, P.T., and MOSELY, W.H. (1973). A new randomized response technique for study of contemporary social problems. Presented at the 101st Annual Meeting of the American Public Health Association, Statistics Section.

FRANKLIN, L.A. (1977). A Bayesian approach to randomized response sampling. Unpublished doctoral dissertation, Indiana University, Bloomington, IN.

GOULD, A.L., SHAH, B.U., and ABERNATHY, J.R. (1969). Unrelated question randomized response techniques with two trials per respondent. *Proceedings of the Section on Social Statistics American Statistical Association*, 351-359.

HORVITZ, D.G., GREENBERG, B.G., and ABERNATHY, J.R. (1976). Randomized response: A data-gathering device for sensitive questions. *International Statistics Review*, 44, 181-196.

KENNEDY, W.J., and GENTLE, J.E. (1980). *Statistical Computing*. New York: Marcel Dekker, Inc..

KNUTH, D.E. (1969). *Semi Numerical Algorithms*, (Volume 2). New York: Addison Wesley.

LIU, P.T., and CHOW, L.P. (1976). The efficiency of the multiple trial randomized response technique. *Biometrics*, 32, 607-618.

O'HAGAN, A. (1987). Bayes linear estimates for randomized response models. *Journal of the American Statistical Association*, 82, 580-585.

PITZ, G.F. (1980). Bayesian analysis of randomized response models. *Psychological Bulletin*, 87, 209-212.

POOLE, W.K. (1974). Estimation of the distribution function of a continuous type random variable through randomized response. *Journal of the American Statistical Association*, 69, 1002-1005.

SMOUSE, E.P. (1984). A note on Bayesian lease squares inference for finite population models. *Journal of the American Statistical Association*, 79, 390-392.

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

WARNER, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.

WINKLER, R.L., and FRANKLIN, L.A. (1979). Warner's randomized response model: A Bayesian approach. *Journal of the American Statistical Association*, 74, 207-214.