RANDOMIZED RESPONSE DESIGNS

By

SORACHAI BHISALBUTRA

Bachelor of Science
Chulalongkorn University
Bangkok, Thailand
1964

Master of Science
Iowa State University
Ames, Iowa
1968

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the Degree of
DOCTOR OF PHILOSOPHY
December, 1976

RANDOMIZED RESPONSE DESIGNS

Thesis Approved:

W. M. Wade
Thesis Adviser

Lyle Broemeling

Don Holbert

Thomas D. Killion

Norman N. Durham
Dean of the Graduate College

ACKNOWLEDGEMENTS

I wish to express my sincere thanks and appreciation to my major adviser, Dr. William D. Warde, for suggesting the problem and for his guidance and assistance throughout the various stages in the preparation of this thesis.

Special appreciation is also expressed to each member of my advisory committee, Dr. Lyle D. Broemeling, Dr. David L. Weeks, Dr. Don Holbert and Dr. Thomas G. Kielhorn, for their interest and helpful suggestions.

Deepest appreciation is expressed to my mother, my father and my wife, Ladda, for their encouragement and love.

Finally, I would like to express my thanks to Mrs. Joyce Gazaway who typed this thesis.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Statement of the Problem and Literature Review

In surveys of human populations on sensitive or highly personal matters, the respondent often refuses to respond or intentionally gives incorrect answers. The bias produced by these two sources of non-sampling error is sometimes large enough to make the sample estimates seriously misleading.

Warner (1) developed an interviewing procedure designed to reduce or eliminate these biases. He called the technique "randomized response" because the respondent selects a question from two or more questions in which at least one is sensitive using a probability basis by using a randomizing device without revealing to the interviewer which of the alternative questions has been chosen. The individual reply, which must be "Yes" or "No" to each question, is of no certain meaning for a specific respondent, but a batch of replies provides useful information for estimating the proportion of the population that has the "sensitive characteristic". To apply the Warner model, a simple random sample of n people will be drawn with replacement from the population and each person interviewed. Before the interviews, the respondent is provided a random device in order to choose one of two statements of the form:

I belong to the sensitive group S.

I do not belong to the sensitive group S.

Without showing to the interviewer which statement has been chosen, the respondent is required only to answer "yes" or "no" according to the statement selected and to his actual status with respect to the sensitive group S.

Let P_S represent the true proportion of respondents who belong to the sensitive group S.

Let p represent the probability that the random device shows S.

If n_1 denotes the number in the sample who report "yes", then the maximum likelihood estimator of P_S and its variance are

$$\hat{P}_S = \frac{n_1}{(2p - 1)n} + \frac{p - 1}{2p - 1}, \quad p \neq \frac{1}{2}$$

$$\text{Var}(\hat{P}_S) = \frac{P_S(1 - P_S)}{n} + \frac{p(1 - p)}{n(2p - 1)^2} .$$

Warner also shows that the proportion of respondents who would answer a direct question untruthfully need not be too great before the mean square error of the usual estimate would exceed the variance of the randomized response estimate.

Abul-Ela et al. (2) extended Warner's design to the trichotomous case to estimate the proportions of three related mutually exclusive groups, one or two of which are sensitive. In order to apply this extension, two independent non-overlapping simple random samples of size n_1 and n_2 must be drawn with replacement from the population. In each of the two samples, a different random device must be used to obtain information concerning the group to which a respondent belongs. Suppose that each random device consists of a deck of cards. Deck 1 is used in the first sample; deck 2 in the second. Each deck contains three different types of cards. One type of card says "I belong to group I"; the second, "I belong to group II"; the third, "I belong to group III". In every deck within one survey, the proportions in deck 1 must be different from those in deck 2, and the proportions within any deck must not be one third for each group.

A variation of the Warner technique has been suggested by Walt R. Simmons [see Horvitz et al. (3)] and is designed to increase further the cooperation of the respondents. It requires the respondents to randomly select one of two unrelated questions, so that the mutually exclusive and complimentary properties of the Warner technique no longer apply. Two independent, non-overlapping simple random samples of size n_1 and n_2 are required. Every respondent in the two samples is asked to reply with only a "Yes" or "No" answer to one of two statements selected on a probability basis, where one question refers to a non-sensitive attribute, say N, unrelated to the sensitive attribute, S. In this particular model, two sets of the randomizing device need to be used. Set 1 is used for respondents in the first sample, and set 2 is used for the respondents in the second sample, and the two sets must also be different with respect to the probability that statement S will be selected.

Let P_S and P_N represent the proportion in the population with the sensitive attribute S and non-sensitive attribute N respectively.

Let p_1 and p_2 represent the probability that the randomizing device shows S in sample set 1 and sample set 2, respectively.

If n_{11} and n_{21} denote the number of "Yes" answers in the two corresponding samples, then the maximum likelihood estimator of P_S and its variance are

$$\hat{P}_S = \frac{\frac{n_{11}}{n_1}(1 - p_2) - \frac{n_{21}}{n_2}(1 - p_1)}{p_1 - p_2}, \quad p_1 \neq p_2,$$

$$\widehat{\text{Var}}(\hat{P}_S) = \frac{1}{(p_1 - p_2)^2}\left[\frac{\frac{n_{11}}{n_1}\left(1 - \frac{n_{11}}{n_1}\right)(1 - p_2)^2}{n_1} + \frac{\frac{n_{21}}{n_2}\left(1 - \frac{n_{21}}{n_2}\right)(1 - p_1)^2}{n_2}\right].$$

It is noted that if P_N is known, then a single sample is sufficient to estimate P_S. This estimator and its variance are

$$\hat{P}_S = \frac{\frac{n_{11}}{n_1} - (1 - p_1)P_N}{p_1}$$

$$\widehat{\text{Var}}(\hat{P}_S) = \frac{1}{n_1 p_1^2}\left[\frac{n_{11}}{n_1}\left(1 - \frac{n_{11}}{n_1}\right)\right].$$

It has been shown by Greenberg et al. (13) that the unrelated questions design with P_N unknown is slightly less efficient than the unrelated questions design with P_N known. Moors (4) showed that, for optimally allocated sample size for the two samples, the unrelated question randomizing device can be used in the first sample only, while the second sample is used to estimate the proportion in the population with the non-sensitive characteristic. If the proportion in the population with the non-sensitive characteristic is known in advance, only one sample is required to estimate the proportion with the sensitive characteristic. Moors further showed that with optimal choice of the two sample sizes, the unrelated questions design will be more efficient than the Warner design, regardless of the probability of choosing the sensitive question in the first sample and regardless of the choice of the proportion with non-sensitive characteristics in the population.

To improve efficiency when two samples are required and the proportion in the population with the non-sensitive characteristic is not known beforehand, Donald T. Campbell suggested a two alternative questions design [see Folsom et al. (5)] which consists of using two non-sensitive alternative questions in conjunction with the sensitive question. The respondents in both samples answer a direct question on a non-sensitive topic and also one of two questions selected by the randomizing device. The non-sensitive question used in the randomized response part of the first sample will be the question which the respondents are required to answer directly in the second sample, and vice versa. It can be shown that the two alternative questions design will never be any less efficient than the estimator with Moors' optimized version of the standard two sample one alternate question design and the two alternate questions design will never be more efficient than the single alternate question design with the proportion in the population with non-sensitive characteristic known.

The variance introduced by the random selection of questions may be reduced by repeated trials with each respondent. The use of two trials per respondent has been discussed by Horvitz et al. (3). A gain in efficiency can be achieved by using the additional information provided by the individual response sequences (yes, yes; yes, no; no, yes; and no, no) rather than pooling data for the two trials. Liu and Chow (10) have suggested the use of a special randomizing device which, in a single trial, yields an estimate with variance roughly equivalent to the variance of the Warner estimate with five trials per respondent. The special randomizing device consists of a spherical bottle with a thin narrow neck. The bottle contains red and white beads, at least six of each color. The respondent is first told that the red beads refer to the sensitive category and the white beads to the non-sensitive category. The respondent is then asked to shake the device thoroughly before turning it upside down, permitting exactly five beads to move into the neck of the bottle, which is frosted on one side so that the interviewer cannot observe the result of the trials. Without mentioning color, respondents who belong to the sensitive class report the number of red beads in the neck of the bottle, and respondents who do not belong to the sensitive class report the number of white beads.

Let Z_i represent the number of beads reported by the i^{th} respondent; T is the total number of beads used in the randomizing device; k is the number of beads sampled in each trial.

If p denotes the proportion of red beads (sensitive category) in the bottle, then it follows that this multiple trials design unbiased estimator of the true proportion of respondents who belong to the sensitive group, S, P_S, is

$$\hat{P}_S = \frac{\bar{Z} - (1 - p)}{(2p - 1)} , \quad p \neq \frac{1}{2}$$

where

$$\bar{Z} = \frac{1}{kn} \sum_{i=1}^{n} Z_i$$

with

$$\text{Var}(\hat{P}_S) = \frac{P_S(1 - P_S)}{n} + \frac{T - k}{T - 1}\left[\frac{p(1 - p)}{(2p - 1)^2}\right].$$

To the present, practically all research in the field of randomized response has been concerned with refining the technique for use with questions of a qualitative nature requiring only a "yes" or "no" response. The technique need not be restricted to nominal scale data. It has wide application in the field of quantitative response data, and study is being directed toward development of the method for use in this area. Greenberg et al. (7) discuss the extension of the randomized response technique to the case of obtaining information on the distribution of quantitative data. They utilize the unrelated questions concept in estimating only the mean and variance of the distribution of the quantitative measure. Greenberg et al. also discussed the choice of the probability of selecting the sensitive questions in the two samples, the selection of the non-sensitive characteristic and the allocation of the sample size into two samples. Eriksson (8) has discussed a new randomized response model for obtaining information on the distribution of a quantitative variable. By using this new method, the i^{th} respondent, who has the true value X_i, is asked to make a random choice of a card from a deck which contains cards saying "Give a true answer" and "Say that your value is Y_j" and then answers in accordance with the instruction on the card. The value given by the i^{th} person is considered as an observation on a variable z_{ij} with sample space $(X_i, Y_1, \ldots, Y_L)$. The values are taken on with probabilities $p, p_1, \ldots, p_L$ respectively. Given the sample of n persons the estimator μ_x and its variance can be found.

To estimate the entire distribution of the quantitative variable, not just the mean and variance, Poole (9) suggested a new technique. This technique is different in that instead of asking the respondent to answer one of two randomly chosen questions, he is asked to multiply the true response by a random number and tell the interviewer only the result. More recently, Liu and Chow (10) have discussed a new discrete quantitative randomized response model by using a predetermined combination of balls in the randomizing device instead of asking an innocuous question. It has been shown that the procedures for administering this method are simple and that its efficiency of estimations is higher than in the other currently available models.

Objectives of the Study and Organization of Thesis

It is the objective of this thesis to develop new randomized response models that increase the cooperation of the respondent, simplify the estimation of the parameters and at the same time decrease the variances of the randomized response estimators. Two randomized response models for proportions and five randomized response models for quantitative data have been proposed in this paper. The description of the models and estimation of the parameters including the models based on repeated trials per respondent have been discussed in Chapter II and Chapter III. In particular, a brief discussion comparing estimates obtained by both direct and randomized response models for proportions under different assumptions is made to see where a new model has potential advantages over the direct interviewing process and other available models and is presented in Chapter II. In Chapter III, we compare the efficiencies of the multiplicative and additive models for quantitative data. The entire study of these comparisons was carried out on the basis of the empirical investigation. The extension of the randomized response models for proportions to the case when some sampled respondents do not report truthfully or refuse to answer the questions and to the multi-proportions situation, the randomized response model for cluster sampling and the combining of randomized response estimators have also been discussed in Chapter II.

The methods of determining the sample size for each model and the extra cost in terms of sample size for the randomized response model for proportions as compared with the regular model are described in Chapter IV which includes examples and applications of some of the results.

Chapter V gives a summary and conclusions of the results obtained in this study.

CHAPTER II

RANDOMIZED RESPONSE MODELS FOR PROPORTIONS

Model I

Description of the Model and Estimation of the Parameters

Based on randomized response research to date, we may say that the unrelated questions design when the proportion in the population with the non-sensitive characteristic is known in advance is always preferable to the other currently available models for proportions. The reasons for this are that the design requires only a single sample and the efficiency of estimation is higher than for the other designs. To get a new model in which the procedure for administering the model is simpler and more efficient than the unrelated questions model (e.g., time spent explaining how to use the randomizing device and hence the interviewing cost, the likelihood of truthful answers), we should consider the procedure in which the knowledge of the population with the non-sensitive characteristics can be achieved by incorporating it in the randomizing device.

Suppose that every person in a population belongs to either the sensitive group S or non-sensitive group N and it is required to estimate the proportion belonging to the sensitive group S. A simple random sample of n people is drawn with replacement from the population and each person is interviewed. Before the interviews, each interviewer is furnished with an identical randomizing device, say a box of balls. Each box contains two different types of balls. Each type of ball says either "Give the true answer" ("yes" or "no" to the sensitive question) or "Yes". In every box being used as a randomizing device, the proportions of these two kinds of balls are identical. Then in each interview the respondent is asked to shake the box and draw one ball unobserved by the interviewer. The respondent only answers the specified question without telling the interviewer which question is being answered.

Let P_S represent the true proportion of respondents who belong to the sensitive group S.

Let m_1 and m_2 represent the number of balls saying "Give the true answer" and "Yes", respectively, then, assuming truthful answers by each respondent, the probability that a respondent will answer "Yes" is

$$p = P_S\left(\frac{m_1}{m_1 + m_2}\right) + \frac{m_2}{m_1 + m_2} . \tag{2.1}$$

Let n_1 denote the number of respondents answering "Yes". Under the assumption of completely truthful reporting, the likelihood equation of the sample is

$$L(P_S) = \left[P_S\left(\frac{m_1}{m_1 + m_2}\right) + \frac{m_2}{m_1 + m_2}\right]^{n_1} \left[1 - \left\{P_S\left(\frac{m_1}{m_1 + m_2}\right) + \frac{m_2}{m_1 + m_2}\right\}\right]^{n-n_1} .$$

Then the maximum likelihood estimate of P_S is

$$\hat{P}_S = \left(\frac{m_1 + m_2}{m_1}\right)\left(\frac{n_1}{n} - \frac{m_2}{m_1 + m_2}\right) . \tag{2.2}$$

The expected value of the estimate is

$$\begin{aligned} E(\hat{P}_S) &= \left(\frac{m_1 + m_2}{m_1}\right)\left[\frac{1}{n} E(n_1) - \frac{m_2}{m_1 + m_2}\right] \\ &= \left(\frac{m_1 + m_2}{m_1}\right)\left[\frac{1}{n} (np) - \frac{m_2}{m_1 + m_2}\right] \\ &= P_S \end{aligned} \tag{2.3}$$

and the variance of $\hat{P}_S$ is

$$\begin{aligned} \operatorname{Var}(\hat{P}_S) &= \left(\frac{m_1 + m_2}{m_1}\right)^2 \frac{\operatorname{Var}(n_1)}{n^2} \\ &= \left(\frac{m_1 + m_2}{m_1}\right)^2 \frac{np(1 - p)}{n^2} \\ &= \frac{P_S(1 - P_S)}{n} + \frac{m_2(1 - P_S)}{m_1 n} . \end{aligned} \tag{2.4}$$

Expression (2.3) and (2.4) show that $\hat{P}_S$ is an unbiased estimate of the true population proportion P_S and the variance of $\hat{P}_S$ can be expressed as the sum of the variance due to sampling plus the variance due to the random device. The estimator of $\operatorname{Var}(\hat{P}_S)$ is

$$\widehat{\operatorname{Var}}(\hat{P}_S) = \frac{1}{n}\left(\frac{m_1 + m_2}{m_1}\right)^2 \left(\frac{n_1}{n}\right)\left(1 - \frac{n_1}{n}\right). \tag{2.5}$$

It is clear that the minimum of $\operatorname{Var}(\hat{P}_S)$ or $\widehat{\operatorname{Var}}(\hat{P}_S)$ can be obtained by choosing m_2/m_1 as small as possible. That is, m_1 is very large when compared with m_2.

Comparison of the Model with Some Other Available Models

For purposes of comparing the efficiency of estimation in the two models, we will first assume that the sample sizes are the same in both models and all respondents are reporting truthfully (except in the regular model). The model effect was computed as the ratio of the variance of the estimator for Model I to the variance of the estimates obtained in the other models. The cases in which Model I is more efficient are shown in the tables or figures by those ratios that are less than one.

Regular Model (Direct Question Model). Suppose that in a regular survey members of group S tell the truth only with probability p_S and members of the non-sensitive group tell the truth with probability one.

The estimator of P_S and its mean square error are

$$\hat{P}_S = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{2.6}$$

$$MSE(\hat{P}_S) = \frac{1}{n}(P_S p_S)(1 - P_S p_S) + [P_S(p_S - 1)]^2 \tag{2.7}$$

where

$$Y_i = \begin{cases} 1 & \text{if the } i^{th} \text{ member reports "Yes",} \\ 0 & \text{if the } i^{th} \text{ member reports "No".} \end{cases}$$

The results of an empirical investigation for $n = 100$, $P_S = 0.1$, 0.3, 0.5, 0.7 and 0.9, and $n = 1000$, $P_S = 0.1$ are given in Table XVI to Table XXI (see Appendix). A plot of the data in Table XVI to Table XXI is shown in Figure 1.

The data in Table XVI to Table XXI and the graph in Figure 1 illustrate the following:

(i) Model I can be more efficient than the regular estimates even with sample sizes as small as 100, depending on the parameter P_S, p_S and m_2/m_1.

(ii) The efficiency of Model I relative to the regular estimates increases as P_S increases or as m_2/m_1 decreases.

(iii) Model I loses efficiency relative to the regular estimates as the probability of telling the truth by the respondents, p_S, increases.

(iv) Model I is more efficient than the regular estimates for all p_S and m_2/m_1 provided $P_S \geq 0.6$ and $n \geq 100$.

(v) If the respondents who belong to the sensitive group tell the truth with probability less than 0.7, then Model I is more efficient than the regular estimates for all P_S and m_2/m_1 (except for some m_2/m_1 when $P_S = 0.1$) even with sample sizes as small as 100.

<u>Warner Model</u>. For the Warner model described in Chapter I, the empirical investigation is shown in Table XXII (see Appendix) and Figure 2. The data and a plot of the data give the following results.

(i) The Warner randomized response model is far less efficient than Model I for $p = m_2/(m_1 + m_2)$.

(ii) Model I, with $m_2/m_1 = 0.1$, is more efficient than the Warner model for all P_S and p.

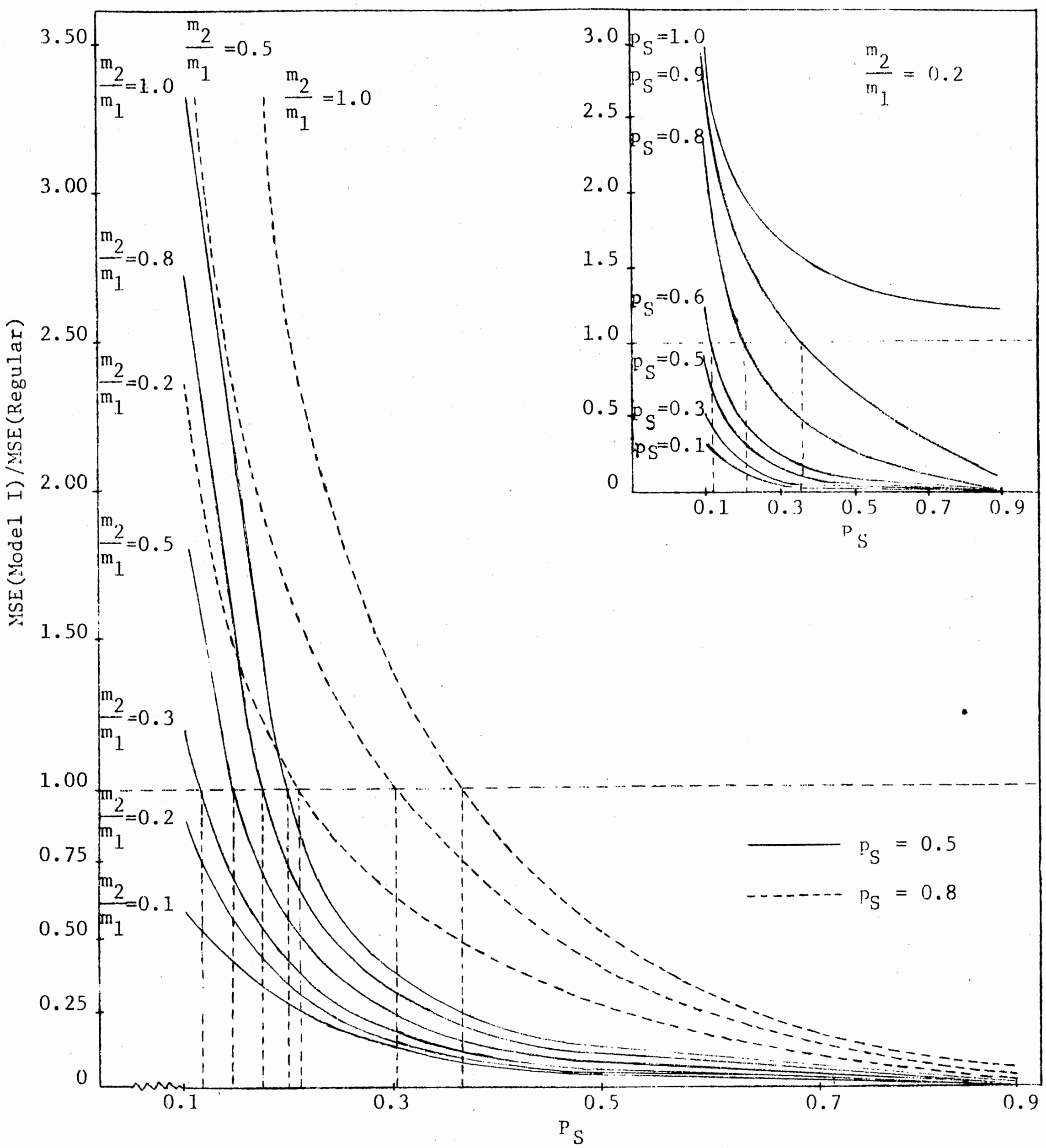

Figure 1. Comparison of Model I and the Regular Model with n = 100

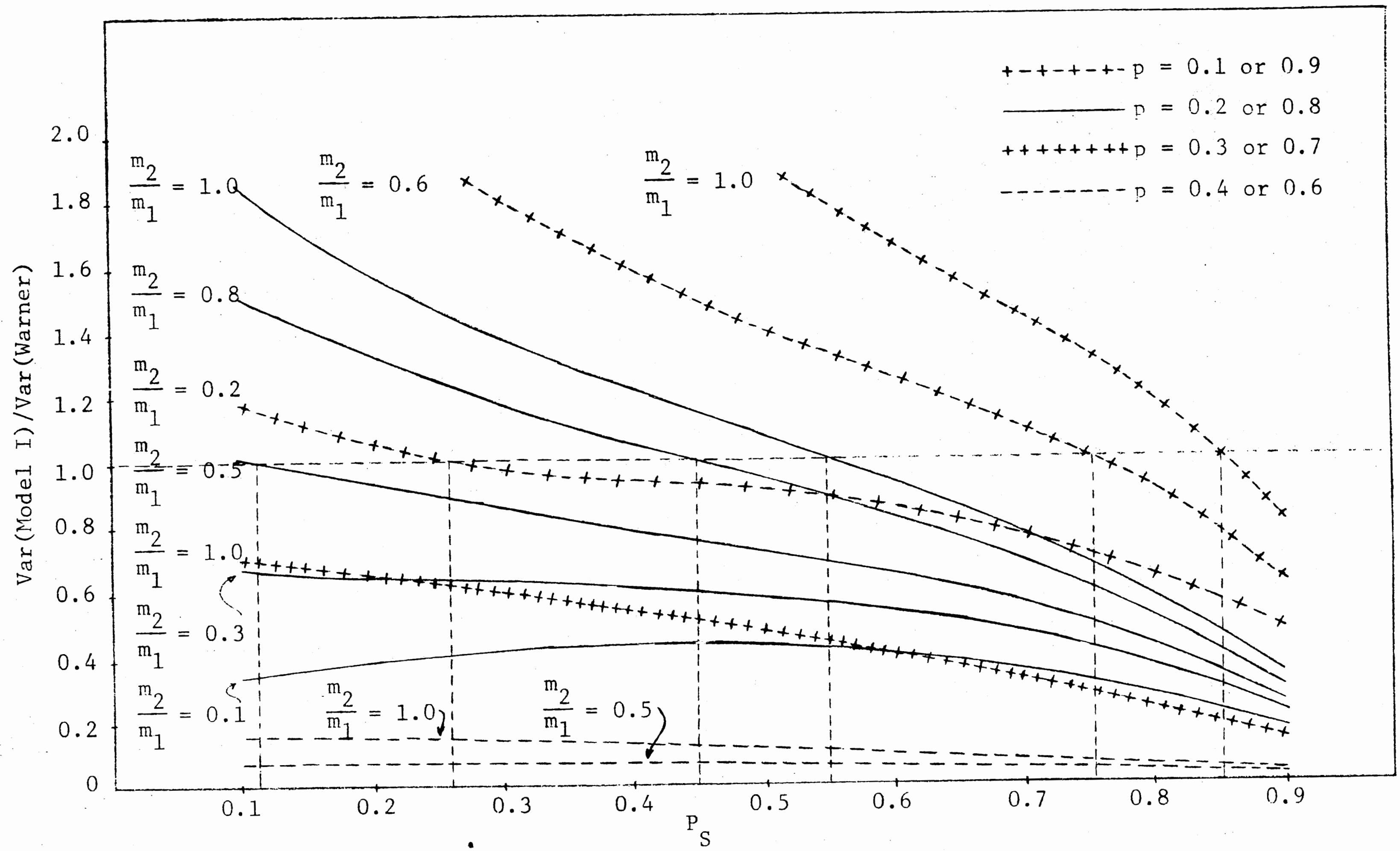

Figure 2. Comparison of Model I and the Warner Model

(iii) For $P_S = 0.9$, Model I is more efficient than the Warner model for all p and m_2/m_1.

(iv) The efficiency of Model I relative to the Warner model increases as P_S increases.

(v) Model I loses efficiency relative to the Warner model as the probability, p, of selecting the sensitive question decreases.

(vi) The variance of Model I and the Warner model are sensitive to m_2/m_1 and p respectively.

(vii) To increase the cooperation of the respondents, let $p = 0.2$ or 0.8. Model I is more efficient than the Warner model for all values of P_S and m_2/m_1 less than 0.5.

(viii) In general, we may say that Model I is more efficient than the Warner model.

Unrelated Questions Model. For the unrelated questions model described in Chapter I, we will only compare Model I and the unrelated questions design with P_N known. The results of empirical sampling for every combination of $P_S = 0.1$, 0.5 and 0.9 and $P_N = 0.1$, 0.5 and 0.9 are given in Table XXIII to Table XXVII (see Appendix). The graph of these data are shown in Figure 3. The results are summarized below:

(i) As P_S increases, Model I is more efficient than the unrelated questions model for all P_N.

(ii) Model I is more efficient than the unrelated questions model for all P_N and m_2/m_1 provided that

$P_S = 0.9$ and p_1 is less than approximately 0.7.

or $P_S = 0.1$ and p_1 is less than approximately 0.3.

or $P_S = 0.5$ and p_1 is less than approximately 0.5.

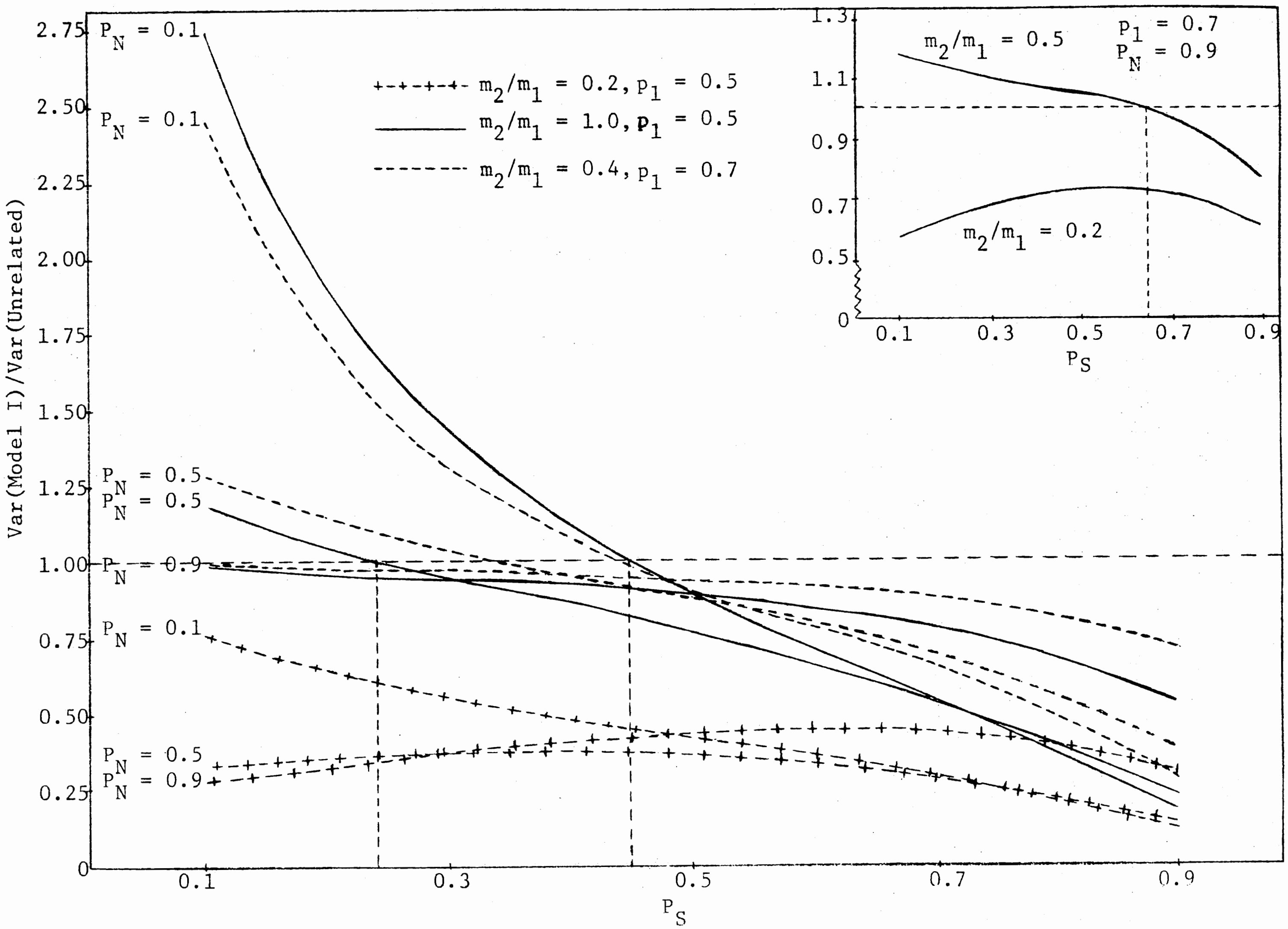

Figure 3. Comparison of Model I and the Unrelated Questions Model

(iii) Model I is more efficient than the unrelated questions model for all P_S, P_N and m_2/m_1, provided that p_1 is less than approximately one third.

(iv) In order to avoid raising suspicion in the respondents, let $p_1 = 0.5$ and $m_2/m_1 = 1$. Model I is more efficient than the unrelated questions model for all P_S and P_N except for very small values of P_S and P_N (e.g., 0.1).

(v) For small values of m_2/m_1, Model I is more efficient than the unrelated questions model except for small values of P_S and large values of p_1 (e.g., 0.9).

Multiple Trials Model. For the multiple trials model described in Chapter I, the ratio of the variance of Model I to that of the multiple trials model for different values of P_S, p, k and m_2/m_1 with $T = 100$ are given in Table XXVIII to Table XXXII (see Appendix). The graph of these data is shown in Figure 4 and the results are summarized below:

(i) Model I is more efficient than the multiple trials model with $T = 100$, $k = 5$, $p = 0.4$ for all m_2/m_1 and P_S or $T = 100$, $k = 10$, $p = 0.4$ for all m_2/m_1 and $P_S \geq 0.4$.

(ii) With $m_2/m_1 \leq 0.1$, Model I is more efficient than the multiple trials model with $T = 100$, $k = 10$ (also 5) for all p and $P_S \geq 0.9$.

(iii) With $m_2/m_1 < 0.3$, Model I is more efficient than the multiple trials model with $T = 100$, $k = 5$ for all P_S and $p > 0.3$.

(iv) With $p = m_2/m_1 = 0.1$, 0.2, 0.3 and 0.4, Model I is more efficient than the multiple trials model with $T = 100$, $k = 5$ for P_S greater than 0.9, 0.7, 0.5 and 0.1, respectively.

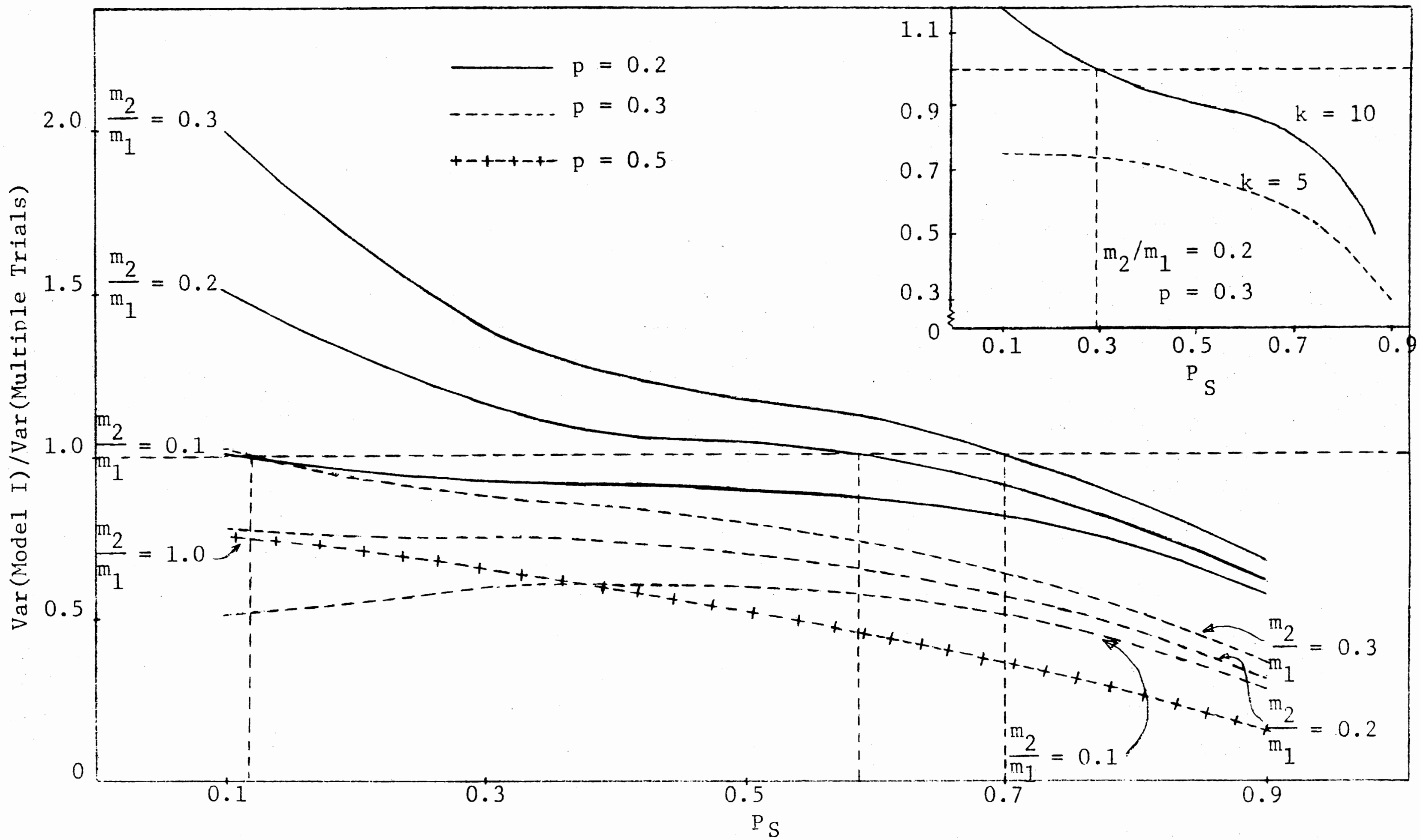

Figure 4. Comparison of Model I and the Multiple Trials Model with T = 100, k = 5, p = 0.2, 0.3, 0.5 and m_2/m_1 = 0.1, 0.2, 0.3, 1.0

(v) As P_S or p increases, the ratio of the variance of Model I to the multiple trials model decreases.

(vi) As k increases, the ratio of the variance of Model I to the multiple trials model increases.

(vii) Use Model I to estimate P_S if we expect that the true proportion of the sensitive group in the population is greater than 0.5 and let m_2/m_1 be less than 0.3.

Extension of Model I in the Case When Some Sampled Respondents Do Not Report Truthfully

Suppose that members of the sensitive group tell the truth only with probability p_{IS} and members of the non-sensitive group tell the truth with probability p_{IN}, then the probability that a respondent will answer "Yes" is

p = Pr (A respondent is a member of the sensitive group and answers "Yes" to the statement "Give the true answer")

+ Pr (A respondent is a member of the non-sensitive group and answers "Yes" to the statement "Give the true answer")

+ Pr (A respondent is a member of the sensitive group and answers "Yes" to the statement "Yes")

+ Pr (A respondent is a member of the non-sensitive group and answers "Yes" to the statement "Yes")

$$= \left(\frac{m_1}{m_1 + m_2}\right) P_S p_{IS} + \left(\frac{m_1}{m_1 + m_2}\right)(1 - P_S)(1 - p_{IN}) + \left(\frac{m_2}{m_1 + m_2}\right) P_S p_{IS}$$

$$+ \left(\frac{m_2}{m_1 + m_2}\right)(1 - P_S) p_{IN}$$

$$= P_S p_{IS} + (1 - P_S)\left[\left(\frac{m_2 - m_1}{m_1 + m_2}\right) p_{IN} + \left(\frac{m_1}{m_1 + m_2}\right)\right] . \qquad (2.8)$$

Let n_1 denote the number of respondents answering "Yes", then the estimate for p is n_1/n.

The estimator of P_S and its variance are

$$\hat{P}_S = \left[\frac{n_1}{n} + \left(\frac{m_1 - m_2}{m_1 + m_2}\right) p_{IN} - \left(\frac{m_1}{m_1 + m_2}\right)\right] \Big/ \left[p_{IS} + \left(\frac{m_1 - m_2}{m_1 + m_2}\right) p_{IN} - \left(\frac{m_1}{m_1 + m_2}\right)\right] \qquad (2.9)$$

$$\text{Var}(\hat{P}_S) = \frac{p(1 - p)}{n\left[p_{IS} + \left(\frac{m_1 - m_2}{m_1 + m_2}\right) p_{IN} - \left(\frac{m_1}{m_1 + m_2}\right)\right]^2} \qquad (2.10)$$

with

$$\widehat{\text{Var}}(\hat{P}_S) = \frac{\frac{n_1}{n}\left(1 - \frac{n_1}{n}\right)}{n\left[p_{IS} + \left(\frac{m_1 - m_2}{m_1 + m_2}\right) p_{IN} - \left(\frac{m_1}{m_1 + m_2}\right)\right]^2} . \qquad (2.11)$$

For $m_1 = m_2$, the estimator of P_S and its variance will be reduced to

$$\hat{P}_S = \left[\frac{n_1}{n} - \frac{1}{2}\right] \Big/ \left[p_{IS} - \frac{1}{2}\right], \quad p_{IS} \neq 1/2 \qquad (2.12)$$

$$\mathrm{Var}(\hat{P}_S) = p'(1 - p') \Big/ \left[n \left(p_{IS} - \frac{1}{2} \right)^2 \right] \tag{2.13}$$

where

$$p' = P_S\left(p_{IS} - \frac{1}{2}\right) + \frac{1}{2}$$

with

$$\widehat{\mathrm{Var}}(\hat{P}_S) = \frac{1}{n\left[p_{IS} - \frac{1}{2}\right]^2}\left(\frac{n_1}{n}\right)\left(1 - \frac{n_1}{n}\right). \tag{2.14}$$

Since those people in a population who belong to the non-sensitive group have no reason to give an incorrect answer, let $p_{IN} = 1$, then the estimator of P_S and its variance are

$$\hat{P}_S = \left[\frac{n_1}{n} - \frac{m_2}{m_1 + m_2}\right] \Big/ \left[p_{IS} - \frac{m_2}{m_1 + m_2}\right] \tag{2.15}$$

$$\mathrm{Var}(\hat{P}_S) = \frac{p''(1 - p'')}{n\left[p_{IS} - \frac{m_2}{m_1 + m_2}\right]^2} \tag{2.16}$$

where

$$p'' = P_S\left[p_{IS} - \frac{m_2}{m_1 + m_2}\right] + \frac{m_2}{m_1 + m_2}$$

with

$$\widehat{\text{Var}}(\hat{P}_S) = \frac{\left(\frac{n_1}{n}\right)\left(1 - \frac{n_1}{n}\right)}{n\left[p_{IS} - \frac{m_2}{m_1 + m_2}\right]^2} . \tag{2.17}$$

A comparison of Model I with the regular model in the case when members of the sensitive group tell the truth only with probability 0.5, 0.7 or 0.9, and members of the non-sensitive group tell the truth with probability one is shown in Table I. The sample size used in each case is 100 and $m_2/m_1 = 0.2$. A plot of the data in Table I is also shown in Figure 5. The data in Table I and the graph in Figure 5 suggest that we should not use Model I if

(i) $p_S > 0.23$ and $p_{IS} = 0.7$ provided that $P_S = 0.1$.

(ii) $p_S > 0.45$ and $p_{IS} = 0.9$ provided that $P_S = 0.1$.

(iii) $P_S < 0.26$ and $p_{IS} = 0.7$ provided that $p_S = 0.5$.

The range of P_S for another combination of p_{IS} and p_S can also be found directly from Figure 5.

Extension of Model I in the Case When Some Sampled Respondents Refuse to Answer the Questions

Some sampled respondents, especially the respondents who belong to the sensitive group, may not want to answer the sensitive questions even using the randomizing device. By forcing the respondents to answer the questions, they may give false information which commonly causes large systematic errors when estimating the parameters of interest. The following is the model developed for the sample in which some sampled respondents refuse to answer the sensitive questions.

TABLE I

COMPARISON OF MODEL I AND THE REGULAR MODEL IN THE CASE WHEN SOME SAMPLED RESPONDENTS DO NOT REPORT TRUTHFULLY, $n = 100$, $m_2/m_1 = 0.2$

P_{IS}	P_S	MSE (Model I)/MSE (Regular)				
		P_S=0.1	P_S=0.3	P_S=0.5	P_S=0.7	P_S=0.9
0.5	0.5	4.8485	0.7404	0.3105	0.1731	0.1093
	0.4	3.6181	0.5260	0.2183	0.1211	0.0762
	0.3	2.7746	0.3918	0.1616	0.0893	0.0561
	0.2	2.1878	0.3026	0.1243	0.0686	0.0431
	0.1	1.7561	0.2405	0.0985	0.0543	0.0341
	0.0	1.4400	0.1955	0.0800	0.0441	0.0276
0.7	0.7	3.8903	0.7921	0.3483	0.1873	0.1067
	0.6	2.7917	0.4868	0.2050	0.1080	0.0608
	0.5	2.0269	0.3252	0.1340	0.0700	0.0392
	0.4	1.5151	0.2310	0.0942	0.0489	0.0273
	0.3	1.1618	0.1721	0.0697	0.0361	0.0201
	0.2	0.9142	0.1329	0.0536	0.0277	0.0154
	0.1	0.7354	0.1056	0.0425	0.0219	0.0122
	0.0	0.6030	0.0859	0.0345	0.0178	0.0099
0.9	0.9	3.6848	1.5360	0.9316	0.5602	0.2759
	0.8	2.9737	0.8136	0.3734	0.1836	0.0773
	0.7	2.1871	0.4518	0.1869	0.0869	0.0353
	0.6	1.5694	0.2777	0.1100	0.0501	0.0201
	0.5	1.1414	0.1855	0.0719	0.0324	0.0130
	0.4	0.8509	0.1318	0.0505	0.0227	0.0090
	0.3	0.6532	0.0982	0.0374	0.0167	0.0067
	0.2	0.5136	0.0758	0.0288	0.0129	0.0051
	0.1	0.4134	0.0602	0.0228	0.0102	0.0040
	0.0	0.3390	0.0490	0.0185	0.0083	0.0033

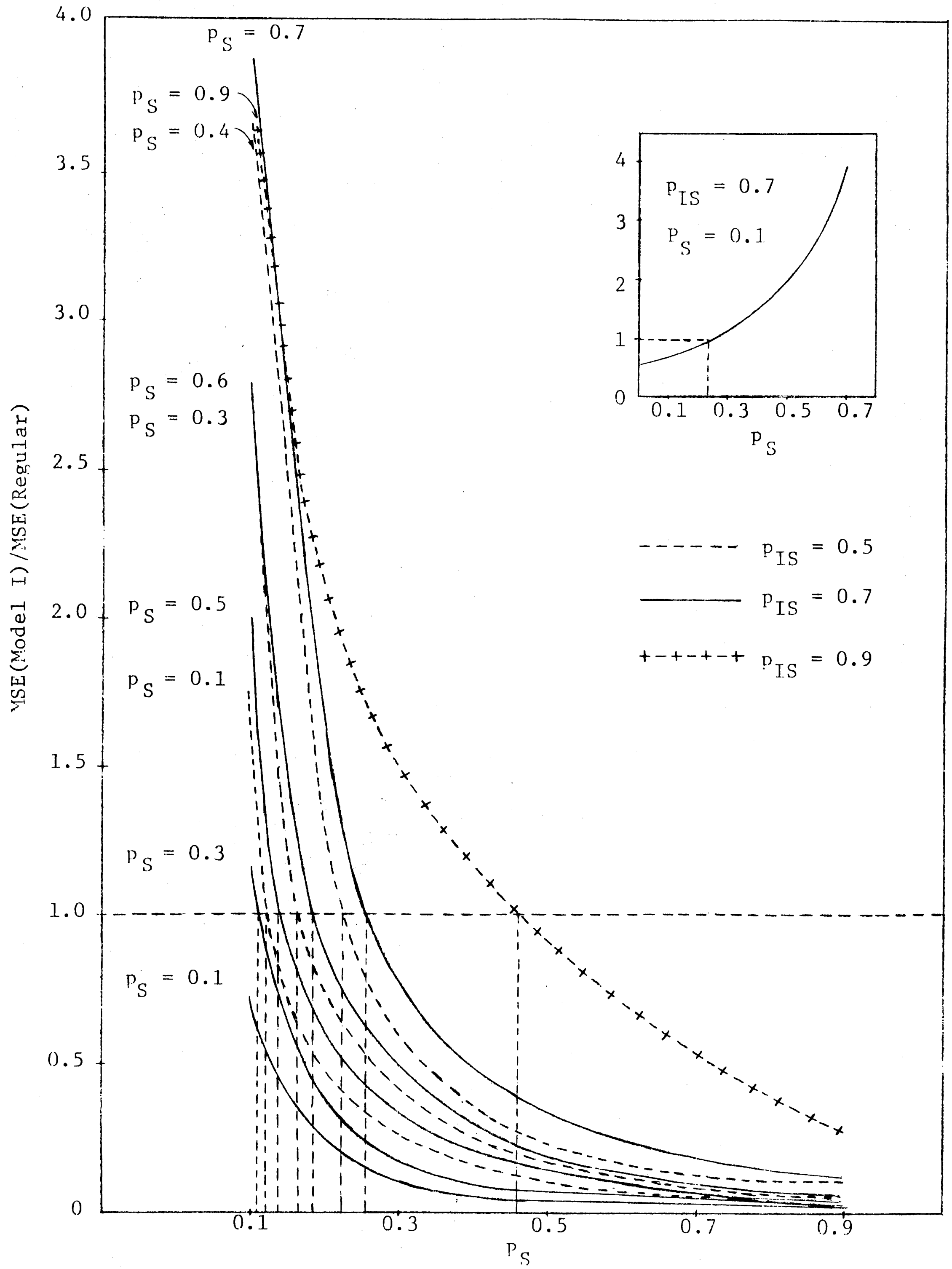

Figure 5. Comparison of Model I and the Regular Model in the Case When Some Sampled Respondents Do Not Report Truthfully with n = 100 and m_2/m_1 = 0.2

Let n be the sample size, n_r be the number of respondents who refuse to answer the questions, then $n_c = n - n_r$ will be the number of respondents who answer the questions.

Thus, the probability that a respondent who answers the questions will answer "Yes" is

$$p_c = P_{Sc}\left(\frac{m_1}{m_1 + m_2}\right) + \frac{m_2}{m_1 + m_2} \tag{2.18}$$

where P_{Sc} is the true population proportion of the people who answer the questions and also belong to the sensitive group.

Since the estimate for p_c is $\hat{p}_c = \frac{n_{c1}}{n_c}$ where n_{c1} is the number of respondents answering "Yes", then

$$\hat{P}_{Sc} = \left[\frac{n_{c1}}{n_c}(m_1 + m_2) - m_2\right]\frac{1}{m_1}$$

$$= \left(\frac{m_1 + m_2}{m_1}\right)\frac{n_{c1}}{n_c} - \frac{m_2}{m_1}\ . \tag{2.19}$$

Let k denote the proportion of respondents who refuse to answer the questions and also belong to the sensitive group, then the adjusted value for $\hat{P}_{Sc}$ is

$$\hat{P}_{Sc}(\text{Adjusted}) = \frac{n_c\hat{P}_{Sc} + kn_r}{n}\ . \tag{2.20}$$

Substituting identity (2.19) in (2.20), the adjusted value for $\hat{P}_{Sc}$ may be written

$$\hat{P}_{Sc}(\text{Adjusted}) = \frac{n_c}{n}\left(\frac{m_1 + m_2}{m_1}\right)\frac{n_{c1}}{n_c} - \frac{n_c m_2}{nm_1} + \frac{kn_r}{n} . \qquad (2.21)$$

We note that (2.21) depends on an unknown parameter k. Various techniques to estimate k are considered later.

It is evident that the persons in a population which do not belong to the sensitive group have no reason to refuse to answer the questions, so let $k = 1$. The estimator for P_{Sc} will be reduced to

$$\hat{P}_{Sc}(\text{Adjusted, } k=1) = \frac{(m_1 + m_2)n_{c1}}{m_1 n} - \frac{n_c m_2}{nm_1} + \frac{n_r}{n} \qquad (2.22)$$

with

$$\text{Var}[\hat{P}_{Sc}(\text{Adjusted},k=1)] = \frac{1}{n^2}\left(\frac{m_1 + m_2}{m_1}\right)^2 [\text{Var}(n_{c1}) + \text{Var}(n_c) - 2\text{Cov}(n_{c1},n_c)] \qquad (2.23)$$

and

$$\widehat{\text{Var}}[\hat{P}_{Sc}(\text{Adjusted},k=1) = \frac{1}{n^2}\left(\frac{m_1 + m_2}{m_1}\right)^2 \left[n\left(\frac{n_{c1}}{n}\right)\left(1 - \frac{n_{c1}}{n}\right) + n\left(\frac{n_c}{n}\right)\left(1 - \frac{n_c}{n}\right) + 2n\left(\frac{n_{c1}}{n}\right)\left(\frac{n_c}{n}\right)\right]$$

$$= \frac{1}{n^3}\left(\frac{m_1 + m_2}{m_1}\right)^2 [n_{c1}(n - n_{c1}) + n_c(n - n_c) + 2n_{c1}n_c]. \qquad (2.24)$$

A further consideration is that, what would be the estimate of P_S for the case when $k \neq 1$. This brings up the question of how can we estimate the value of k from our sample. There are at least two alternative approaches, both of which are approximate solutions:

(1) Using $\hat{k} = \frac{n_r}{n} \hat{P}_{Sc}$. (2.25)

(2) Using $\hat{k} = \frac{n_r}{n} \hat{P}_{Sc}$(Adjusted). (2.26)

Alternative 1. This estimator is good only when the value of P_S is the same for both those who answer and those who refuse to answer. There is some evidence that the value of a parameter will not be the same for respondents as for non-respondents [Finkner (16), Hendricks (17)]. In our case, the value of P_S in the sampled respondents who answer the question is likely to be less than the value of P_S in those sampled respondents who refuse to answer.

Alternative 2. Since the value of P_{Sc}(Adjusted) depends upon the value of k, then a direct solution of k is not available. However, the value of k can be found by a simple iterative procedure, illustrated as follows:

(i) Find the first approximation

$$\hat{k}_1 = \frac{n_r}{n} \hat{P}_{Sc}$$

and then

$$\hat{P}_{Sc}(\text{Adjusted})_1 = \frac{n_c}{n}\left(\frac{m_1 + m_2}{m_1}\right)\frac{n_{c1}}{n_c} - \frac{n_c m_2}{n m_1} + \frac{\hat{k}_1 n_r}{n} .$$

(ii) Find the second approximation

$$\hat{k}_2 = \frac{n_r}{n} \hat{P}_{Sc}(\text{Adjusted})_1$$

and then

$$\hat{P}_{Sc}(\text{Adjusted})_2 = \frac{n_c}{n}\left(\frac{m_1 + m_2}{m_1}\right)\frac{n_{c1}}{n_c} - \frac{n_c m_2}{nm_1} + \frac{\hat{k}_2 n_r}{n} .$$

(iii) Find the third approximation

$$\hat{k}_3 = \frac{n_r}{n} \hat{P}_{Sc}(\text{Adjusted})_2$$

and then

$$\hat{P}_{Sc}(\text{Adjusted})_3 = \frac{n_c}{n}\left(\frac{m_1 + m_2}{m_1}\right)\frac{n_{c1}}{n_c} - \frac{n_c m_2}{nm_1} + \frac{\hat{k}_3 n_r}{n} .$$

(iv) If the i^{th} approximation of $\hat{k}$, $\hat{k}_i$, is equal or approximately equal to the $(i - 1)^{th}$ approximation, then $\hat{k}_i$ is the estimator of k we want.

By using $\hat{k} = \left(\frac{n_r}{n}\right) \hat{P}_{Sc}$ [method (1)] as an estimator of k, the adjusted value for $\hat{P}_{Sc}$ will be

$$\hat{P}_{Sc}(\text{Adjusted},\hat{k}) = \frac{1}{n}\left(n_c + \frac{n_r^2}{n}\right)\left[\left(\frac{m_1 + m_2}{m_1}\right)\frac{n_{c1}}{n_c} - \frac{m_2}{m_1}\right] . \qquad (2.27)$$

If we assume n_c fixed instead of being a random variable, then the variance of $\hat{P}_{Sc}$(Adjusted) can be written in the form

$$\text{Var}[\hat{P}_{Sc}(\text{Adjusted},k=1,n_c \text{ fixed})] = \left(\frac{n_c}{n}\right)^2\left(\frac{m_1 + m_2}{m_1}\right)^2 \text{Var}\left(\frac{n_{c1}}{n_c}\right)$$

$$= \frac{n_c}{n}\left[\frac{P_{Sc}(1 - P_{Sc})}{n} + \frac{m_2(1 - P_{Sc})}{m_1 n}\right] . \qquad (2.28)$$

Since the expression (2.28) without $\frac{n_c}{n}$ is $\mathrm{Var}(\hat{P}_{Sc})$ [see (2.4)], and $\frac{n_c}{n}$ is less than or equal to 1, then we can conclude that

$$\mathrm{Var}[\hat{P}_{Sc}(\text{Adjusted}, k=1, n_c \text{ fixed})] \leq \mathrm{Var}(\hat{P}_{Sc}). \qquad (2.29)$$

Again, the variance of $\hat{P}_{Sc}(\text{Adjusted})$ when $\hat{k} = \frac{n_r}{n}\hat{P}_{Sc}$ can also be written in the form

$$\mathrm{Var}[\hat{P}_{Sc}(\text{Adjusted}, \hat{k}, n_c \text{ fixed})] = \frac{1}{n^2}\left(n_c + \frac{n_r^2}{n}\right)^2\left(\frac{m_1 + m_2}{m_1}\right)^2 \mathrm{Var}\left(\frac{n_{c1}}{nc}\right)$$

$$= \left[\frac{n_c}{n} + \frac{n_r^2}{n^2}\left(2 + \frac{n_r^2}{nn_c}\right)\right]\left[\frac{P_{Sc}(1 - P_{Sc})}{n} + \frac{m_2(1 - P_{Sc})}{m_1 n}\right]$$

$$= \mathrm{Var}[\hat{P}_{Sc}(\text{Adjusted}, k=1, n_c \text{ fixed})]$$

$$+ \frac{n_r^2}{n^2}\left(2 + \frac{n_r^2}{nn_c}\right)\left[\frac{P_{Sc}(1 - P_{Sc})}{n} + \frac{m_2(1 - P_{Sc})}{m_1 n}\right]. \qquad (2.30)$$

Combining Model I with the Unrelated Questions Model (P_N Known)

In general Model I and the unrelated questions model seem to be more efficient than other available models. Unfortunately, the efficiency of these two models over the other ones depends upon the value of P_S. For some surveys we may not have any idea about the value of P_S in the population at all. In order to get a new model which is more efficient than either of these two models, the best way is to combine the good points of the two models, using the inverse of the variance of the estimators in each model as weights. For this purpose, only a single sample is needed. But each sampled respondent has to use two different randomizing devices, one set for Model I and another set for the unrelated questions model, to answer the questions given by each model. A diagram of the design is shown in Table II below.

TABLE II

DIAGRAM OF THE COMBINED DESIGN

Sampled Respondent	Model I	Unrelated Questions Model
1	——	——
2	——	——
3	——	——
.	.	.
.	.	.
.	.	.
n	——	——
$\hat{P}_S$	$\hat{P}_{SI}$	$\hat{P}_{SU}$
$Var(\hat{P}_S)$	$Var(\hat{P}_{SI})$	$Var(\hat{P}_{SU})$

We shall now formulate the combined estimate of P_S in Model I and the unrelated questions model in mathematical terms.

$$\hat{P}_S(\text{Combined}) = \frac{\dfrac{\hat{P}_{SI}}{Var(\hat{P}_{SI})} + \dfrac{\hat{P}_{SU}}{Var(\hat{P}_{SU})}}{\dfrac{1}{Var(\hat{P}_{SI})} + \dfrac{1}{Var(\hat{P}_{SU})}}$$

$$= \frac{Var(\hat{P}_{SU})\hat{P}_{SI} + Var(\hat{P}_{SI})\hat{P}_{SU}}{Var(\hat{P}_{SI}) + Var(\hat{P}_{SU})} . \qquad (2.31)$$

The variance of $\hat{P}_S(\text{Combined})$ is approximately

$$Var[\hat{P}_S(\text{Combined})] = \frac{[Var(\hat{P}_{SU})]^2 Var(\hat{P}_{SI}) + [Var(\hat{P}_{SI})]^2 Var(\hat{P}_{SU})}{[Var(\hat{P}_{SI}) + Var(\hat{P}_{SU})]^2}$$

$$= \frac{Var(\hat{P}_{SI})Var(\hat{P}_{SU})}{Var(\hat{P}_{SI}) + Var(\hat{P}_{SU})} . \qquad (2.32)$$

The expression (2.32) implies that $Var[\hat{P}_S(\text{Combined})]$ is always less than or equal to the minimum of $Var(\hat{P}_{SI})$ and $Var(\hat{P}_{SU})$.

Model II

In order to extend Model I to a multi-proportions randomized response model, we will look at a special version of Model I which we will call Model II. In this situation, there is no difference between Model I and Model II except for the randomizing device used in these two models.

Description of the Model and Estimation of the Parameters

The randomizing device used in this model is a box containing m balls; m_1 are white and $m - m_1$ are black $(m - m_1 \geq 2)$. The proportion of balls with different colors must be predetermined. At the interview, the respondent is asked to shake the box and draw two of the balls in that box with or without replacement. The respondent is required to tell the truth if he gets at least one white ball (i.e., "Yes" or "No"). If the respondent gets all black balls, he is required to say "Yes".

Let P_S represent the true proportion of respondents who belong to the sensitive group S, then the probability that a respondent will answer "Yes" for sampling without replacement is

$$p = P_S\left(\frac{\binom{m_1}{2}\binom{m-m_1}{1}+\binom{m_1}{2}}{\binom{m}{2}}\right)+\frac{\binom{m-m_1}{2}}{\binom{m}{2}}$$

$$= P_S\left(1-\frac{\binom{m-m_1}{2}}{\binom{m}{2}}\right)+\frac{\binom{m-m_1}{2}}{\binom{m}{2}} \ . \qquad (2.33)$$

If a random sample of size n is taken, let n_1 denote the number of respondents answering "Yes", then the estimator of P_S is

$$\hat{P}_S = \left[\frac{n_1}{n}\binom{m}{2}-\binom{m-m_1}{2}\right]\Big/\left[\binom{m}{2}-\binom{m-m_1}{2}\right] \qquad (2.34)$$

with

$$\mathrm{Var}(\hat{P}_S) = \left(\frac{\binom{m}{2}}{\binom{m}{2}-\binom{m-m_1}{2}}\right)^2 \mathrm{Var}\left(\frac{n_1}{n}\right)$$

$$= \left(\frac{\binom{m}{2}}{\binom{m}{2}-\binom{m-m_1}{2}}\right)^2 \frac{1}{n}\, p(1-p). \qquad (2.35)$$

Substituting identity (2.33) in (2.35), the variance of $\hat{P}_S$ may be written

$$\mathrm{Var}(\hat{P}_S) = \frac{1}{nk_1^2}\left[k_1 P_S\left\{1 - k_1 P_S - 2\binom{m - m_1}{2}\right\} + \binom{m - m_1}{2}\left\{1 - \binom{m - m_1}{2}\right\}\right] \tag{2.36}$$

where $k_1 = \binom{m}{2} - \binom{m - m_1}{2}$.

Since the estimator for p is $\frac{n_1}{n}$, then the estimator for $\mathrm{Var}(\hat{P}_S)$ is

$$\widehat{\mathrm{Var}}(\hat{P}_S) = \left(\frac{\binom{m}{2}}{k_1}\right)^2 \frac{1}{n}\left(\frac{n_1}{n}\right)\left(1 - \frac{n_1}{n}\right). \tag{2.37}$$

For sampling with replacement, a ball is drawn out, its color noted and then replaced. This is done two times. In this case, the probability of drawing a white ball is constant and is equal to m_1/m. By applying the probability density function of the binomial distribution, the probability that a respondent will answer "Yes" is

$$p = P_S\left[\binom{2}{1}\left(\frac{m_1}{m}\right)\left(1 - \frac{m_1}{m}\right) + \left(\frac{m_1}{m}\right)^2\right] + \left(1 - \frac{m_1}{m}\right)^2. \tag{2.38}$$

Proceeding exactly as in the sampling without replacement case we obtain

$$\hat{P}_S = \left[\frac{n_1}{n} - \left(1 - \frac{m_1}{m}\right)^2\right] \Big/ \left[1 - \left(1 - \frac{m_1}{m}\right)^2\right] \tag{2.39}$$

with

$$\mathrm{Var}(\hat{P}_S) = \frac{1}{nk_2^2}\left[k_2P_S\left\{1 - k_2P_S - 2\left(1 - \frac{m_1}{m}\right)^2\right\} + \left(1 - \frac{m_1}{m}\right)^2 \left\{1 - \left(1 - \frac{m_1}{m}\right)^2\right\}\right] \tag{2.40}$$

where

$$k_2 = \left[1 - \left(1 - \frac{m_1}{m}\right)^2\right] = \frac{m_1}{m}\left(2 - \frac{m_1}{m}\right)$$

and

$$\widehat{\mathrm{Var}}(\hat{P}_S) = \frac{1}{nk_2^2}\left(\frac{n_1}{n}\right)\left(1 - \frac{n_1}{n}\right). \tag{2.41}$$

It can be shown that sampling with replacement is generally more precise than sampling without replacement but for a large population size, the precision of the two methods tends to be very similar.

Let us turn now to the efficiency of Model II. The expression (2.37) and (2.41) suggest that if a large number of balls are drawn out from the box and the ratio of the white balls to the black balls is very large, then the efficiency of Model II will be increased significantly. For the purpose of inducing more cooperation in the respondents, the ratio of the white balls to the black balls should not be too large (it should probably be no larger than 4) and the number of balls drawn from the box should not be greater than three.

A Multi-Proportions Randomized Response Model

In the case where every person in a population belongs to one of t mutually exclusive groups $(1, 2, \ldots, t)$ and it is desired to estimate the proportion in each group, the technique of Model II can be extended.

We shall now illustrate this technique by using a box containing m balls $(m \geq 2t)$ where $m_1 \geq t$ are white and $(m - m_1) \geq t$ are black, as our randomizing device. The proportion of balls with different colors will be predetermined. At the interview, the respondent is asked to shake the box and draw at least t of the balls in that box with or without replacement. The respondent is required to answer according to the number of white balls he gets and the probability of getting that number of white balls. For example, suppose we use $m = 30$, $m_1 = 20$ to estimate the proportion of 3 mutually exclusive groups in the population and let the respondent draw 5 of the balls from the box without replacement. In this case there are six possible numbers of white balls drawn from the box. They are 0, 1, 2, 3, 4 and 5. The assigned groups to each number of white balls and the corresponding probabilities are shown in Table III.

If our interests do not center on any groups, the groups that are assigned to $r = 0$, 1 and 2 can be interchanged. In the case where our interest is primarily in a particular group and where the proportion of respondents belonging to that group is expected to be high, we should assign that group to the value of r which has the largest probability among the three values which have been assigned the response

"Answer Group i". If, however, the proportion belonging to that group is expected to be small, then the value of r assigned should be the one having the smallest of these three probabilities. For example, if we are interested in group 1 and want to get a good estimator for P_1, the assigned number of white balls for "Group 1" should be two if the expected P_1 is high, and zero if it is low.

TABLE III

THE ASSIGNED GROUPS TO THE NUMBER OF WHITE BALLS AND THE CORRESPONDING PROBABILITIES

Number of White Balls (r)	Pr(r)	"Response"
0	0.00177	"Group 1"
1	0.02947	"Group 2"
2	0.15999	"Group 3"
3	0.35998	"Give the true answer"
4	0.33999	"Give the true answer"
5	0.10880	"Give the true answer"

It is evident that the main emphasis in assigning answers to values of r is to maximize the probability of getting "Give the true answer" and minimize the probability of getting "Group i".

Now, let P_i represent the true proportion of respondents who belong to "Group i", then the probability that a respondent will answer

"Group i" is

$$p_i = P_i\left[1 - \sum_{i=1}^{t} Pr(r_i)\right] + Pr(r_i) \qquad (2.42)$$

where $Pr(r_i)$ is the probability of drawing "Group i".

If a random sample of size n is taken, let n_i denote the number of respondents answering "Group i", and then the estimator for p_i is n_i/n.

Substituting $p_i = \frac{n_i}{n}$ in (2.42) and solving for P_i we will get

$$\hat{P}_i = \left[\frac{n_i}{n} - Pr(r_i)\right] \Big/ \left[1 - \sum_{i=1}^{t} Pr(r_i)\right] \qquad (2.43)$$

with

$$Var(\hat{P}_i) = \frac{1}{n} p_i(1 - p_i) \Big/ \left[1 - \sum_{i=1}^{t} Pr(r_i)\right]^2 \qquad (2.44)$$

and

$$Cov(\hat{P}_i, \hat{P}_j) = \frac{Cov\left(\frac{n_i}{n}, \frac{n_j}{n}\right)}{\left[1 - \sum_{i=1}^{t} Pr(r_i)\right]^2} . \qquad (2.45)$$

Now,

$$\mathrm{Cov}\left(\frac{n_i}{n}, \frac{n_j}{n}\right) = \frac{1}{n^2}\,\mathrm{Cov}(n_i, n_j)$$

$$= \frac{1}{n^2}\,\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) \quad \text{where} \quad X_i, X_j = 1 \text{ or } 0,$$

$$= \frac{1}{n}\,\mathrm{Cov}(X_i, X_j)$$

$$= \frac{1}{n}[E(X_iX_j) - E(X_i)E(X_j)]$$

$$= -\frac{1}{n}\,E(X_i)E(X_j)$$

$$= -\frac{1}{n}\,p_ip_j. \tag{2.46}$$

As before, since $\hat{p}_i = \frac{n_i}{n}$, then the estimator of $\mathrm{Var}(\hat{P}_i)$ and $\mathrm{Cov}(\hat{P}_i, \hat{P}_j)$ are

$$\widehat{\mathrm{Var}}(\hat{P}_i) = \frac{1}{n}\left(\frac{n_i}{n}\right)\left(1 - \frac{n_i}{n}\right) \Bigg/ \left[1 - \sum_{i=1}^{t} \Pr(r_i)\right]^2 \tag{2.47}$$

$$\widehat{\mathrm{Cov}}(\hat{P}_i, \hat{P}_j) = -\frac{1}{n}\left(\frac{n_i}{n}\right)\left(\frac{n_j}{n}\right) \Bigg/ \left[1 - \sum_{i=1}^{t} \Pr(r_i)\right]^2. \tag{2.48}$$

In order to minimize the variance of the estimator, the expression (2.44) suggests that

(i) The ratio m_1/m should tend to one and hence the probability of getting a large number of white balls be high.

(ii) The number of balls drawn from the randomizing device each time should be equal or close to the number of black balls in the randomizing device.

(iii) The number of white balls that correspond to the t

smallest probabilities should be assigned to the t groups, the rest of them assigned to the answer "Give the true answer".

We note, however, that $\sum_{i=1}^{t} Pr(r_i)$ should be about 0.2 or more to encourage cooperation among the respondents.

Randomized Response Model for Proportions in Cluster Sampling

According to a randomized response model for proportions, the sampling units are classified into the sensitive group or the non-sensitive group so that $\hat{P}_S$ is the ratio of the number of units in the sensitive group in the sample to the total number of units in the sample. The formulae for the variance and the estimated variance of $\hat{P}_S$ derived for this case can not be used in the case when each sampling unit is composed of a group of elements, and it is the elements that are classified as S or N. For example, if the sampling unit is a family and the elements are members of the family or the sampling unit is a school and elements are students in that school.

In order to find $\hat{P}_S$ and the estimated variance of $\hat{P}_S$ for this kind of sampling unit, suppose the size of the sampling unit is not constant.

Let m_i be the number of elements in the i^{th} sampling unit, then the estimator of P_S in the i^{th} sampling unit, $\hat{P}_{Si}$, is

$$\hat{P}_{Si} = m_{Si}/m_i \tag{2.49}$$

where m_{Si} is the number of elements belonging to the sensitive group.

The proportion of units falling in the sensitive group in the sample is

$$\hat{P}_S = \sum_{i=1}^{n} m_{Si} \Big/ \sum_{i=1}^{n} m_i ,$$

$$= \hat{R} \tag{2.50}$$

where n is the sample size.

Structurally, this is a typical ratio estimate. Let us consider the variance of a ratio estimator. If variates x_i and y_i are measured on each unit of a simple random sample of size n, assumed large, the variance of $\hat{R} = \bar{y}/\bar{x}$ is approximately:

$$\text{Var}(\hat{R}) \doteq \frac{1}{n\bar{X}^2} \sum_{i=1}^{N} \frac{(y_i - Rx_i)^2}{N - 1} \tag{2.51}$$

where $R = \bar{Y}/\bar{X}$ is the ratio of population means.

If we put m_{Si} for y_i and m_i for x_i in (2.51), the approximate variance of $\hat{P}_S$ is

$$\text{Var}(\hat{P}_S) \doteq \frac{1}{n\bar{M}^2} \sum_{i=1}^{N} \frac{(m_{Si} - P_S m_i)^2}{N - 1}$$

$$= \frac{1}{n} \sum_{i=1}^{N} \left(\frac{m_i}{\bar{M}}\right)^2 \frac{(\hat{P}_{Si} - P_S)^2}{N - 1} \tag{2.52}$$

where P_S is the proportion of elements in the sensitive group in the population and $\bar{M} = \frac{1}{N} \sum_{i=1}^{N} m_i$ is the average number of elements per sampling unit in the population.

For the estimated variance of $\hat{P}_S$

$$\widehat{\text{Var}}(\hat{P}_S) \doteq \frac{1}{n} \sum_{i=1}^{n} \left(\frac{m_i}{\bar{m}}\right)^2 \frac{(\hat{P}_{Si} - \hat{P}_S)^2}{n - 1} \qquad (2.53)$$

where $\bar{m} = \frac{1}{n} \sum_{i=1}^{n} m_i$ is the average number of elements per sampling unit in the sample.

In summary, a randomized response model may be used in conjunction with a ratio estimator of the population proportion P_S in the case when each sampling unit is composed of a group of elements, and it is the elements that are classified as S or N. The formula for estimating P_S is the same as the formula used in the randomized response model for proportions with one respondent as a sampling unit, but the formulae for the variance and estimated variance of $\hat{P}_S$ have to be changed by using the approximate variance of a ratio estimator with $m_{Si} = y_i$ and $m_i = x_i$.

Extension of the Randomized Response Models for Proportions to Two Trials and t Trials per Respondent

It seems clear enough that a gain in the efficiency of the randomized response models can be achieved by using the additional information provided by repeated trials with each respondent. Ordinarily, doubling the number of observations will reduce the variance in random samples by about one-half. We shall now illustrate this technique on two trials and t trials per respondent.

Two Trials per Respondent

An extension of the randomized response model for proportions requires each respondent to make two independent selections of the two questions using the same randomizing device.

It is evident that for an individual from whom two responses are required there are four possible responses: (Yes, Yes), (Yes, No), (No, Yes), and (No, No).

If we let n_{11}, n_{10}, n_{01} and n_{00} be the number of individuals answering (Yes, Yes), (Yes, No), (No, Yes) and (No, No) respectively in the sample and n be the sample size, then the estimate for the probability that a respondent will answer "Yes", p, in each model will be

$$\hat{p}^* = (2n_{11} + n_{10} + n_{01})/2n. \tag{2.54}$$

$\hat{P}_S$ can then be found by substituting $\hat{p}^*$ for p in each model. Similarly, the estimated variance of $\hat{P}_S$ can be found by substituting $\hat{p}^*$ for p in $Var(\hat{P}_S)$ but instead of dividing by n we have to divide by $2n$. The results for each model are summarized in Table IV.

t Trials per Respondent

In this case, each respondent is required to make t independent selections of the t questions using the same randomizing device. As before, let $n_{11 \ldots 1}$, $n_{11 \ldots 0}$, ..., and $n_{00 \ldots 0}$ be the numbers of individual answering (Yes, Yes, ..., Yes), (Yes, Yes, ..., No), ..., and (No, No, ..., No) respectively.

TABLE IV

ESTIMATORS OF P_S AND $\mathrm{Var}(\hat{P}_S)$ IN RANDOMIZED RESPONSE MODELS FOR PROPORTION WITH TWO TRIALS PER RESPONDENT

Model	$\hat{P}_S$	$\widehat{\mathrm{Var}}(\hat{P}_S)$
I	$\frac{m_1 + m_2}{m_1}\hat{p}^* - \frac{m_2}{m_1}$	$\frac{1}{2n}\left(\frac{m_1 + m_2}{m_1}\right)^2 \hat{p}^*(1 - \hat{p}^*)$
II (Without Replacement)	$\hat{p}^* - \frac{\binom{m - m_1}{2}\binom{m}{2}}{\binom{m}{2}k_1}$	$\frac{\binom{m}{2}}{2k_1 n}\hat{p}^*(1 - \hat{p}^*)$

Let n be the sample size, then the estimate for p can be written in the form

$$\hat{p}^{**} = \frac{1}{tn}[tn_{11\ldots1} + (t-1)(n_{11\ldots10} + n_{11\ldots01} + \cdots + n_{01\ldots1})$$
$$+ (t-2)(n_{11\ldots00} + n_{11\ldots010} + \cdots + n_{001\ldots1})$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$+ 2\ (n_{110\ldots0} + n_{1010\ldots0} + \cdots + n_{00\ldots11})$$
$$+ (n_{10\ldots0} + n_{010\ldots0} + \cdots + n_{00\ldots1})]. \tag{2.55}$$

Proceeding exactly as in the two trials case we obtain the same formulae for $\hat{P}_S$ and $\widehat{Var}(\hat{P}_S)$ as in the two trials case except we put $\hat{p}^{**}$ for $\hat{p}^{*}$ and tn for 2n.

The alternative way of applying repeated trials with each respondent is to use t different sets of randomizing devices for each respondent. The estimator for P_S is the weighted average of $\hat{P}_S$ obtained from the t samples. The weights used in this case are the inverse of the variances of the t estimators. We shall now formulate the problem in mathematical terms.

$$\hat{P}_S(\text{Repeated}) = \sum_{i=1}^{t} [\hat{P}_{Si}/Var(\hat{P}_{Si})] \Big/ \left[\sum_{i=1}^{t} \left\{ 1/Var(\hat{P}_{Si}) \right\} \right] \tag{2.56}$$

$$Var[\hat{P}_S(\text{Repeated})] = \sum_{i=1}^{t} \left[1 \Big/ \left\{ Var(\hat{P}_{Si}) \quad \left(\sum_{i=1}^{t} (1/Var(\hat{P}_{Si}) \right)^2 \right\} \right] \tag{2.57}$$

where $\hat{P}_{Si}$ and n_i represent the estimators of the true proportion and the number of respondents answering "Yes" in the i^{th} sample respectively. We note that both (2.56) and (2.57) involve the quantities $Var(\hat{P}_{Si})$ which are unknown. It is usual in a practical application to estimate $Var(\hat{P}_{Si})$ by $\widehat{Var}(\hat{P}_{Si})$ and make this substitution in the appropriate places in (2.56) and (2.57).

If we compare the efficiency of these two methods of applying repeated trials, we will see that the gain in efficiency is somewhat less with the alternative method, however, because of the correlation among the t responses.

CHAPTER III

RANDOMIZED RESPONSE MODELS FOR QUANTITATIVE DATA

Extensions of the Multi-Proportions Randomized Response Models to Estimate the Mean for Quantitative Data

Model III

The multi-proportions randomized response model presented earlier in Model II can be extended to estimate the mean for quantitative data. For the purpose of illustration, let us suppose that the midpoint of the i^{th} interval or class of the quantitative measures belonging to "Group i" is represented by X_i. Then the estimated population mean of the quantitative measures can be written in the form

$$\hat{\mu} = \sum_{i=1}^{t} \hat{P}_i X_i \qquad (3.1)$$

with

$$\widehat{Var}(\hat{\mu}) = \sum_{i=1}^{t} X_i^2 \widehat{Var}(\hat{P}_i) + 2 \sum_{i<j=1}^{t} X_i X_j \widehat{Cov}(\hat{P}_i, \hat{P}_j)$$

$$= \frac{\sum_{i=1}^{t} n_i(n - n_i)X_i^2 - 2 \sum_{i<j=1}^{t} n_i n_j X_i X_j}{n^3 \left[1 - \sum_{i=1}^{t} Pr(r_i)\right]^2} , \qquad (3.2)$$

where $\hat{P}_i$ is the estimator of the true proportion of respondents belonging to the i^{th} interval of the quantitative measures.

Before proceeding to the next model, we shall give an example of this application. Suppose we want to estimate the average income of people in a certain area and we divided the expected income of these people into three classes or intervals ($0-7,499, $7,500-14,999 and $15,000-22,499). The randomizing device used in this model is the same as the randomizing device used in Model II (for multi-proportions). The instructions tell the respondent to answer "Group i" if a ball is drawn with "Group i" on it. If a "Give the true answer" ball is selected, then the respondent should tell the interviewer the group in which he actually belongs. An example of group definitions is given in Table V.

The midpoints, as we have seen, are 3,749.5, 11,249.5 and 18,749.5 respectively.

In order to get a simpler procedure for administering the model, the randomizing device used in this extension should be a box of (t + 1) kinds of balls. Each kind of ball says either "I belong to Group i" (i = 1, 2, ..., t) or "Give the true answer". The required statements corresponding to each group are marked on the surface of the balls (e.g., range of income). At the interview, the respondent is asked to shake the box and draw a ball from that box. The respondent is required to answer in accordance with the instructions or statements on the surface of the balls. If it is a ball which says "Give the true answer" then the respondent is expected to answer truthfully with the number of the group to which he belongs. If the ball shows "Group i", then the respondent answers with the number i. As before, the respondent only answers the specified question without telling the interviewer which question is being answered.

TABLE V

THE CORRESPONDING QUESTIONS TO EACH GROUP

Model II (Multi-Proportions)	Model III
Group 1	Did you earn less than \$7,500 last year?
Group 2	Did you earn more than \$7,500 but less than \$15,000 last year?
Group 3	Did you earn more than \$15,000 but less than \$22,500 last year?

Now, consider the method of estimation of our parameters. Let m_1 represent the number of balls saying "Give the true answer" and let the number of balls for each group be one, then the total number of balls in the box is $m_1 + t$.

Let P_i represent the true proportion of respondents who belong to "Group i", then the probability that a respondent will answer "Group i" is

$$p_i = P_i\left(\frac{m_1}{m_1 + t}\right) + \frac{1}{m_1 + t} \ . \tag{3.3}$$

If a random sample of size n is taken, let n_i denote the number of respondents answering "Group i", then the maximum likelihood estimates of P_i, its estimated variance and covariance are

$$\hat{P}_i = \left(\frac{m_1 + t}{m_1}\right)\frac{n_i}{n} - \frac{1}{m_1} \ , \tag{3.4}$$

$$\widehat{\mathrm{Var}}(\hat{P}_i) = \frac{1}{n}\left(\frac{m_1 + t}{m_1}\right)^2\left(\frac{n_i}{n}\right)\left(1 - \frac{n_i}{n}\right), \tag{3.5}$$

$$\widehat{\mathrm{Cov}}(\hat{P}_i, \hat{P}_j) = -\frac{1}{n}\left(\frac{m_1 + t}{m_1}\right)^2\left(\frac{n_i}{n}\right)\left(\frac{n_j}{n}\right) \ . \tag{3.6}$$

Referring to (3.4), (3.5) and (3.6) the estimated population mean of the quantitative measures and its estimated variance are obtained as follows:

$$\hat{\mu} = \sum_{i=1}^{t}\left[\left(\frac{m_1 + t}{m_1}\right)\left(\frac{n_i}{n}\right) - \frac{1}{m_1}\right]X_i \tag{3.7}$$

$$\widehat{\mathrm{Var}}(\hat{\mu}) = \frac{1}{n^3}\left(\frac{m_1 + t}{m_1}\right)^2\left[\sum_{i=1}^{t} n_i(n - n_i)X_i^2 - 2\sum_{i<j=1}^{t} n_i n_j X_i X_j\right]. \tag{3.8}$$

<u>Model IV</u>

In the case where every person in a population belongs to one of t mutually exclusive groups, and it is required to estimate the proportion in each group, the Warner model can be extended as in Abul-Ela et al. (2). For this purpose $(t - 1)$ simple random samples with replacement of sizes $n_1, n_2, \ldots, n_{t-1}$ are required to estimate the $(t - 1)$ proportions and $\hat{P}_t = 1 - \sum_{k=1}^{t-1} \hat{P}_k$. It is necessary to use $(t - 1)$ sets of different combinations of the probability of getting the j^{th} group from the i^{th} sample and satisfying the condition $\sum_{j=1}^{t} p_{ij} = 1$ for $i = 1, 2, \ldots, t - 1$.

For purposes of illustration let us suppose that n_{i1} is the number of "Yes" answers reported in the i^{th} sample. Then the maximum likelihood estimate of P_S can be written in the form

$$\hat{\underline{P}}_S = \underline{p}^{-1}\underline{n} \tag{3.9}$$

with

$$\widehat{Var}(\hat{\underline{P}}_S) = \underline{p}^{-1}\underline{v}(\underline{p}^{-1})' \tag{3.10}$$

where

$$\hat{\underline{P}}_S = \begin{bmatrix} \hat{P}_1 \\ \hat{P}_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \hat{P}_{t-1} \end{bmatrix} \tag{3.11}$$

$$\underline{p} = \begin{bmatrix} p_{11} - p_{1t} & p_{12} - p_{1t} & \cdots\cdots & p_{1(t-1)} - p_{1t} \\ p_{21} - p_{2t} & p_{22} - p_{2t} & \cdots\cdots & p_{2(t-1)} - p_{2t} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ p_{(t-1)1} - p_{(t-1)t} & p_{(t-1)2} - p_{(t-1)t} & \cdots\cdots & p_{(t-1)(t-1)} - p_{(t-1)t} \end{bmatrix} \tag{3.12}$$

$$\underline{n} = \begin{bmatrix} \frac{n_{11}}{n_1} - p_{1t} \\ \frac{n_{21}}{n_2} - p_{2t} \\ \cdots\cdots\cdots \\ \frac{n_{(t-1)1}}{n_{(t-1)}} - p_{(t-1)t} \end{bmatrix} \tag{3.13}$$

$$\underline{v} = \begin{bmatrix} v_{11} & & & \emptyset \\ & v_{22} & & \\ & & \ddots & \\ \emptyset & & & v_{(t-1)(t-1)} \end{bmatrix} \tag{3.14}$$

$$v_{ii} = \frac{1}{n_i}\left(\frac{n_{i1}}{n_i}\right)\left(1 - \frac{n_{i1}}{n_i}\right), \quad i = 1, 2, \ldots, t - 1. \tag{3.15}$$

Again, if we let X_i represent the midpoint of the i^{th} interval of the quantitative measures belonging to "Group i", then the estimated population mean of the quantitative measures and its estimated variance will be obtained as follows:

$$\hat{\mu} = \left(J_1^{t-1}\right)' \underline{\underline{Xp}}^{-1}\underline{n} + X_t\left(1 - \sum_{i=1}^{t-1} \hat{P}_i\right) \tag{3.16}$$

$$\widehat{Var}(\hat{\mu}) = \left(J_1^{t-1}\right)' \underline{\underline{Xp}}^{-1}\widehat{Var}(\underline{n})(\underline{p}^{-1})' \underline{X}J_1^{t-1} + X_t^2\widehat{Var}\left[\left(J_1^{t-1}\right)' \underline{\hat{P}}_S\right]$$

$$= \left(J_1^{t-1}\right)' \underline{\underline{Xp}}^{-1}\underline{v}(\underline{p}^{-1})' \underline{X}J_1^{t-1} + X_t^2\left(J_1^{t-1}\right)' \underline{p}^{-1}\underline{v}(\underline{p}^{-1})' J_1^{t-1} \tag{3.17}$$

where

$$J_1^{t-1} = \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix}_{(t-1)\times 1} , \quad \text{and} \quad \underline{X} = \begin{bmatrix} X_1 & & & & \emptyset \\ & X_2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ \emptyset & & & & X_{t-1} \end{bmatrix} . \tag{3.18}$$

By applying Model IV to the previous example, two independent non-overlapping simple random samples of size n_1 and n_2 are drawn with replacement from the population. Suppose that the random device consists of two decks of cards. Deck 1 is used in the first sample, deck 2 in the second sample. Each deck contains three different types of cards. Each type of card says:

(i) Did you earn less than \$7,500 last year?

(ii) Did you earn more than \$7,500 but less than \$15,000 last year?

(iii) Did you earn more than \$15,000 but less than \$22,500 last year?

The proportions of these three kinds of cards within any deck are identical, but the proportions of cards in deck 1 are different from those in deck 2 and the proportions within any deck must not be one third for each kind [see (3.9), (3.10) and (3.12)]. As before, the respondent is required to draw a card and answer only "Yes" or "No" according to the question on that card.

If we let n_{11} and n_{21} denote the number of "Yes" answers reported in the first sample and second sample respectively, then the estimated average income of people in that area is

$$\hat{\mu} = (J_1^2)' \underline{X}\underline{p}^{-1}\underline{n} + X_3(1 - \hat{P}_1 - \hat{P}_2) \quad (3.19)$$

with

$$\widehat{\text{Var}}(\hat{\mu}) = (J_1^2)' \underline{X}\underline{p}^{-1}\underline{v}(\underline{p}^{-1})' \underline{X}J_1^2 + X_3^2(J_1^2)' \underline{p}^{-1}\underline{v}(\underline{p}^{-1})' J_1^2 \quad (3.20)$$

where

$$J_1^2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\underline{X} = \begin{bmatrix} 3,749.5 & 0 & 0 \\ 0 & 11,249.5 & 0 \\ 0 & 0 & 18,749.5 \end{bmatrix}$$

$$\underline{p} = \begin{bmatrix} p_{11} - p_{13} & p_{12} - p_{13} \\ p_{21} - p_{23} & p_{22} - p_{23} \end{bmatrix}$$

$$\underline{n} = \begin{bmatrix} \frac{n_{11}}{n_1} - p_{13} \\ \frac{n_{21}}{n_2} - p_{23} \end{bmatrix}$$

$$\underline{v} = \begin{bmatrix} v_{11} & 0 \\ 0 & v_{22} \end{bmatrix}$$

<u>Advantages of the Extensions</u>

Making use of this technique to estimate the mean for quantitative data, the procedure for administering the model is rather simple and causes less embarrassment to the respondents and hence increases the likelihood of truthful answers. It is evident that the efficiency of the estimate depends on the distribution of the quantitative measures and the choice of intervals or classes. Grouping the measurements into classes, it is then assumed that the midpoint of each class will fairly represent all the observations in that class. This assumption involves a slight approximation, but the results will prove quite satisfactory for significance tests provided that the observations are divided into at least 10 intervals. When the data shows a natural tendency to be spread fairly evenly among all the intervals, regardless of their width, the efficiency of the estimate will increase as the number of intervals used increases.

Model V (Multiplicative Model)

Description of the Model and Estimation of the Parameters

The multiplicative model was first developed by Poole (9). The randomizing device used in this model is a box containing a number of balls. A number such as 1, 2, ..., r will be marked on the surface of each ball and the proportion of balls with different numbers will be predetermined. As in the previous models, the respondent is asked to shake the box and draw one of the balls in that box. Suppose a survey is conducted in order to estimate the average income of some population of interest as in the examples given in Model III and Model IV. The respondent multiplies the number on the surface of the ball he gets by his income and the result is recorded.

By using Poole's technique, the distribution of the random multiplier has to be known beforehand and the estimation procedure is rather complicated. In order to get a simpler procedure of estimation we will introduce an alternative method of estimation.

Let X_i and Y_i represent the true response and randomized response of the i^{th} respondent respectively.

Let p_j represent the proportion of the balls marked r_j, $j = 1, 2, \ldots, t$ and $\sum_{j=1}^{t} p_j = 1$, then

$$E(Y_i) = \sum_{j=1}^{t} r_j p_j X_i$$

$$\sum_{i=1}^{n} E(Y_i) = \sum_{j=1}^{t} r_j p_j \sum_{i=1}^{n} X_i$$

$$= n\bar{X} \sum_{j=1}^{t} r_j p_j . \tag{3.21}$$

Solving for $\bar{X}$, we obtain

$$\bar{X} = \sum_{i=1}^{n} E(Y_i) \Bigg/ \left(n \sum_{j=1}^{t} r_j p_j \right) .$$

The estimator for $\bar{X}$ is then obtained by putting Y_i for $E(Y_i)$. Thus we have

$$\hat{\bar{X}} = \sum_{i=1}^{n} Y_i \Bigg/ \left(n \sum_{j=1}^{t} r_j p_j \right)$$

$$= \bar{Y} \Bigg/ \sum_{j=1}^{t} r_j p_j \tag{3.22}$$

with

$$\mathrm{Var}(\hat{\bar{X}}) = \left(\sum_{i=1}^{n} Y_i^2 - n\bar{\bar{Y}}^2\right) \Big/ \left[n(n-1)\left(\sum_{j=1}^{t} r_j p_j\right)^2\right]. \tag{3.23}$$

A non-detailed empirical investigation will be discussed in the next section.

Empirical Investigation

The objective of this study is to investigate the choice of t, p_j, and r_j which we use in this model in order to minimize the variance of the estimator. For questions such as, "What is the best combination of t, p_j and r_j?", the answers are more difficult to obtain. It is the author's opinion that if we know the range of the true responses, we should divide this range into t equal intervals. The choice of r_j should be the midpoints of each interval, and the choice of p_j should be close to the frequencies of X_i in each interval. It is also clear that t should be as large as possible. In order to attack the problem from an investigation viewpoint we shall consider the choices of t, p_j and r_j as follows:

(i) $t = 2, 3, 4$ and 6.

(ii) $r_j = (1, 3), (1, 2, 3), (4, 5, 6), (7, 8, 9)$

and

$$r_j \doteq [\mathrm{Min}(X_i) + (I/2)] + (j - 1)I, \; j = 1,2,\ldots,t \tag{3.24}$$

where

$$I = [\mathrm{Max}(X_i) - \mathrm{Min}(X_i)]/t \; .$$

(iii) $p_j = 1/t$

and

$$p_j = f_k/n$$

where f_k is the frequency of X_i in the k^{th} interval, $k = 1, 2, \ldots, t$ and n is the sample size,

and

$$p_j = \frac{1/f_k}{\sum_{k=1}^{t} (1/f_k)} . \tag{3.25}$$

In this study, the per capita personal income for 25 states from different regions of the United States of America in 1974 (see Table XXXIII, Appendix) will be considered as X_i $(i = 1, 2, \ldots, 25)$. The corresponding r_j for the true response X_i are found by using a table of random numbers. The estimator of $\bar{X}$, its variance and coefficient of variation for five different sets of random numbers and for various combinations of t, r_j and p_j are summarized in Table VI to Table VIII.

From the data in Table VI to Table VIII, it is somewhat surprising to find that in order to minimize the variance of the estimate we should choose t, p_j and r_j such that the variance of r_j is as small as possible. We observe that for the same variance of r_j, the variance of $\hat{\bar{X}}$ will decrease as $\bar{r}_j = \frac{1}{t} \sum_{j=1}^{t} r_j$ increases. Also, the p_j, $j = 1, 2, \ldots, t$, can be chosen in any fashion but we must have $\sum_{j=1}^{t} p_j = 1$.

TABLE VI

THE VALUES OF $\hat{\bar{X}}$ FOR EACH SET OF THE SAMPLE AND THE AVERAGE OBTAINED FROM MODEL V

Combination	r_j	p_j	$Var(r_j)$	$\bar{r}_j$	$\hat{\bar{X}}$ 1	2	3	4	5	Average
a	1:3	1/2	2.0000	2.00	4542.00	5134.44	4899.52	5308.82	5057.80	4988.52
b	1:2:3	1/3	1.0000	2.00	5173.82	4923.60	4985.14	4909.24	5217.06	5041.77
c	4:5:6	1/3	1.0000	5.00	5206.54	5106.48	5131.10	5100.74	5239.07	5153.74
d	4:5:6	.44:.40:.16	1.0000	5.00	5140.78	5091.33	5106.51	5196.75	5140.04	5135.08
e	4:5:6	.20:.23:.57	1.0000	5.00	5132.92	5226.84	5142.47	5307.68	5140.83	5190.15
f	4:5:6	.16:.40:.44	1.0000	5.00	5142.64	5165.93	5034.23	5227.47	5190.17	5152.09
g	4.87:6.33	1/2	1.0658	5.60	5049.43	5203.89	5142.63	5249.39	5183.91	5165.85
h	4.62:5.60:6.57	1/3	0.8946	5.60	5206.15	5150.41	5178.46	5114.45	5221.44	5174.18
i	4.50:5.23:5.96:6.69	1/4	0.8135	5.60	5187.95	5217.27	5151.40	5188.27	5185.67	5186.11
j	4.38:4.87:5.35:5.84:6.33:6.82	1/6	0.8101	5.60	5234.42	5221.18	5224.55	5243.94	5238.28	5232.47
k	7:8:9	1/3	1.0000	8.00	5214.72	5152.20	5167.57	5148.59	5225.54	5181.73
									True Value ($\bar{X}$)	5228.36

TABLE VII

THE VALUES OF VAR($\hat{\hat{X}}$) FOR EACH SET OF THE SAMPLE AND THE AVERAGE OBTAINED FROM MODEL V

Combi-nation	r_j	p_j	Var(r_j)	$\bar{r}_j$	Var($\hat{\hat{X}}$)[1] 1	2	3	4	5	Average
a	1:3	1/2	2.0000	2.00	2.1238	2.2543	2.0985	2.1141	1.8852	2.0952
b	1:2:3	1/3	1.0000	2.00	1.4754	1.0718	1.2132	0.9755	1.1219	1.1716
c	4:5:6	1/3	1.0000	5.00	0.2589	0.1451	0.1267	0.1158	0.1340	0.1561
d	4:5:6	.44:.40:.16	1.0000	5.00	0.2996	0.1659	0.1472	0.1417	0.1274	0.1764
e	4:5:6	.20:.23:.57	1.0000	5.00	0.2338	0.1736	0.1695	0.1558	0.1478	0.1761
f	4:5:6	.16:.40:.44	1.0000	5.00	0.2008	0.1482	0.1227	0.0789	0.0816	0.1264
g	4.87:6.33	1/2	1.0658	5.60	0.2028	0.1997	0.1668	0.1628	0.1051	0.1675
h	4.62:5.60:6.57	1/3	0.8946	5.60	0.2339	0.1174	0.0970	0.0998	0.1072	0.1311
i	4.50:5.23:5.96:6.69	1/4	0.8135	5.60	0.2207	0.1046	0.1013	0.0998	0.0957	0.1244
j	4.38:4.87:5.35:5.84:6.33:6.82	1/6	0.8101	5.60	0.2165	0.1136	0.1146	0.0875	0.0804	0.1225
k	7:8:9	1/3	1.0000	8.00	0.1750	0.1115	0.0883	0.0947	0.0998	0.1139

[1]All variances have been multiplied by 10^5.

TABLE VIII

THE VALUES OF COEFFICIENT OF VARIATION OF $\hat{\bar{X}}$ FOR EACH SET OF THE SAMPLE AND THE AVERAGE OBTAINED FROM MODEL V

Combination	r_j	p_j	$Var(r_j)$	$\bar{r}_j$	c.v. (%)					
					1	2	3	4	5	Average
a	1:3	1/2	2.0000	2.00	10.1463	9.2472	9.3498	8.6609	8.5845	9.1757
b	1:2:3	1/3	1.0000	2.00	7.4241	6.6493	6.9870	6.3621	6.4204	6.7889
c	4:5:6	1/3	1.0000	5.00	3.0904	2.3587	2.1934	2.1095	2.2162	2.4242
d	4:5:6	.44:.40:.16	1.0000	5.00	3.3670	2.5301	2.3762	2.2905	2.1957	2.5862
e	4:5:6	.20:.23:.57	1.0000	5.00	2.9789	2.5212	2.5319	2.3519	2.3650	2.5570
f	4:5:6	.16:.40:.44	1.0000	5.00	2.7555	2.3569	2.2008	1.6989	1.7400	2.1826
g	4.87:6.33	1/2	1.0658	5.60	2.8206	2.7159	2.5114	2.4310	1.9780	2.5052
h	4.62:5.60:6.57	1/3	0.8946	5.60	2.9376	2.1039	1.9016	1.9537	1.9829	2.2120
i	4.50:5.23:5.96:6.69	1/4	0.8135	5.60	2.8635	1.9605	1.9537	1.9253	1.8869	2.1509
j	4.38:4.87:5.35:5.84:6.33:6.82	1/6	0.8101	5.60	2.8110	2.0415	2.0494	1.7839	1.7122	2.1156
k	7:8:9	1/3	1.0000	8.00	2.5366	2.0495	1.8189	1.8900	1.9121	2.0593

Model VI (Additive Model)

Description of the Model and Estimation of the Parameters

According to Model V there is another case one might consider. What will happen if we ask the sampled respondents to add the numbers marked on the surface of the balls to their incomes instead of multiplying by them. This brings up the question of what to do when it is desired to compare this additive model to the multiplicative model. We shall discuss this problem later. Now, let us first find the formulae for $\hat{\bar{X}}$ and $Var(\hat{\bar{X}})$.

As before, let X_i and Y_i represent the true response and randomized response of the i^{th} sampled respondent respectively.

Let p_j represent the proportion of the balls marked r_j^*, $j = 1, 2, \ldots, t$, and $\sum_{j=1}^{t} p_j = 1$, then

$$E(Y_i) = (r_1^* + X_i)p_1 + (r_2^* + X_i)p_2 + \ldots + (r_t^* + X_i)p_t$$

$$= \sum_{j=1}^{t} r_j^* p_j + X_i. \tag{3.26}$$

$$\sum_{i=1}^{n} E(Y_i) = n \sum_{j=1}^{t} r_j^* p_j + \sum_{i=1}^{n} X_i$$

$$= n\left(\sum_{j=1}^{t} r_j^* p_j + \bar{X}\right).$$

Solving for $\bar{X}$, we obtain

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} E(Y_i) - \sum_{j=1}^{t} r_j^* p_j.$$

Again, the estimator for $\bar{X}$ is then obtained by putting Y_i for $E(Y_i)$. Thus we have

$$\hat{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \sum_{i=1}^{t} r_j^* p_j$$

$$= \bar{Y} - \sum_{j=1}^{t} r_j^* p_j \qquad (3.27)$$

$$= \bar{X} + \bar{r}^* - \sum_{j=1}^{t} r_j^* p_j \qquad (3.28)$$

where

$$\bar{r}^* = \frac{1}{n}\sum_{i=1}^{n} r_i^*.$$

The variance of $\hat{\bar{X}}$ is

$$\mathrm{Var}(\hat{\bar{X}}) = \mathrm{Var}(\bar{Y})$$

$$= \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{r}^*). \qquad (3.29)$$

An empirical investigation of this model will be discussed next.

Empirical Investigation

For the same set of data and the same set of random numbers that we used in Model V, we shall calculate $\hat{\bar{X}}$, the variance of $\hat{\bar{X}}$ and the coefficient of variation for Model VI for some combinations of t, r_j^* and p_j. Referring to (3.28) and (3.29), we shall investigate Model VI using the following combinations:

(i) $r_j^* = (1000, 3000)$ and $p_j = 1/2$.

(ii) $r_j^* = (3000, 5000)$ and $p_j = 1/2$.

(iii) $r_j^* = (1000, 5000)$ and $p_j = 1/2$.

(iv) $r_j^* = (1000, 3000, 5000)$ and $p_j = 1/3$.

(v) $r_j^* = (1000, 2000, 3000, 4000, 5000)$ and $p_j = 1/5$.

(vi) $r_j^* = (2000, 2500, 3000, 3500, 4000)$ and $p_j = 1/5$.

(vii) $r_j^* = (2500, 2750, 3000, 3250, 3500)$ and $p_j = 1/5$.

(viii) $r_j^* = (2750, 2875, 3000, 3125, 3250)$ and $p_j = 1/5$.

The values of $\text{Var}(r_j^*)$, $\hat{\bar{X}}$, $\text{Var}(\hat{\bar{X}})$ and the coefficient of variation of $\hat{\bar{X}}$ are summarized in Table IX to Table XI.

In summary, in order to minimize the variance of the estimator, the choices of t, p_j and r_j^* should be such that the variance of r_j^* should be as small as possible. We note that for the same variance of r_j^*, if t increases, then the variance of $\hat{\bar{X}}$ will decrease, and that p_j can take any values so long as $\sum_{j=1}^{t} p_j$ equals one.

Comparison of Model V and Model VI

In order to compare the efficiency of the multiplicative model and the additive model, the values of $r_j\bar{X}$ and $r_j^* + \bar{X}$ where r_j and r_j^* are random numbers used in the two models, should be the same for both models. For example, if the r_j used in the multiplicative model are (1, 3), then the values of $r_j\bar{X}$ are (5228.36, 15685.08) and hence the corresponding values of r_j^* used in the additive model should be (0, 10456.72).

For the purpose of this comparison, the value of $\hat{\bar{X}}$, its variance and the coefficient of variation of the two models for the same sets of

TABLE IX

THE VALUES OF $\hat{X}$ FOR EACH SET OF THE SAMPLE AND THE AVERAGE OBTAINED FROM MODEL VI

Combination	r_j^*	p_j	$\Sigma r_j^* p_j$	$Var(r_j^*)$[1]	$\hat{X}$ 1	2	3	4	5	Average
a	1000:3000	1/2	2,000	20	5028.36	5268.36	5188.36	5348.36	5268.36	5220.36
b	3000:5000	1/2	4,000	20	5028.36	5268.36	5188.36	5348.36	5268.36	5220.36
c	1000:5000	1/2	3,000	80	4828.36	5308.36	5148.36	5068.36	5308.36	5212.36
d	1000:3000:5000	1/3	3,000	80	5228.36	5388.36	5228.36	5148.36	5388.36	5276.36
e	1000:2000:3000:4000:5000	1/5	3,000	100	5148.36	5228.36	5148.36	5348.36	5228.36	5220.36
f	2000:2500:3000:3500:4000	1/5	3,000	30	5188.36	5228.36	5188.36	5288.36	5228.36	5224.36
g	2500:2750:3000:3250:3500	1/5	3,000	6.25	5208.36	5228.36	5208.36	5258.36	5228.36	5228.36
h	2750:2875:3000:3125:3250	1/5	3,000	1.5625	5218.36	5228.36	5218.36	5243.36	5228.36	5227.36
									True Value ($\bar{X}$)	5228.36

[1]All variances have been multiplied by 10^5.

TABLE X

THE VALUES OF VAR($\hat{\bar{X}}$) FOR EACH SET OF THE SAMPLE AND THE AVERAGE OBTAINED FROM MODEL VI

Combi- nation	r_j^*	p_j	$\Sigma r_j^* p_j$	Var(r_j^*)[1]	Var($\hat{\bar{X}}$)[1]					
					1	2	3	4	5	Average
a	1000:3000	1/2	2000	20	0.6318	0.6478	0.6478	0.6425	0.6478	0.6435
b	3000:5000	1/2	4000	20	0.6318	0.6478	0.6478	0.6425	0.6478	0.6435
c	1000:5000	1/2	3000	80	1.8318	1.8958	1.8958	1.8745	1.8958	1.8787
d	1000:3000:5000	1/3	3000	80	1.2985	1.4211	1.4318	1.2291	1.2878	1.3337
e	1000:2000:3000:4000:5000	1/5	3000	100	1.0291	1.1303	0.9958	0.8758	1.0651	1.0192
f	2000:2500:3000:3500:4000	1/5	3000	30	0.4311	0.3985	0.4228	0.3928	0.4401	0.4171
g	2500:2750:3000:3250:3500	1/5	3000	6.25	0.2816	0.2735	0.2795	0.2720	0.2839	0.2781
h	2750:2875:3000:3125:3250	1/5	3000	1.5625	0.2442	0.2422	0.2437	0.2419	0.2448	0.2434

[1]All variances have been multiplied by 10^5.

TABLE XI

THE VALUES OF COEFFICIENT OF VARIATION OF $\hat{\bar{X}}$ FOR EACH SET OF THE SAMPLE AND THE AVERAGE OBTAINED FROM MODEL VI

Combination	r_j^*	p_j	$\Sigma r_j^* p_j$	$Var(r_j^*)$[1]	c.v.(%)					
					1	2	3	4	5	Average
a	1000:3000	1/2	2000	20	4.9988	4.8311	4.9056	4.7393	4.8311	4.8593
b	3000:5000	1/2	4000	20	4.9988	4.8311	4.9056	4.7393	4.8311	4.8593
c	1000:5000	1/2	3000	80	8.8642	8.2023	8.4572	7.9174	8.2023	8.3156
d	1000:3000:5000	1/3	3000	80	6.8922	6.9961	7.2373	6.8096	6.6599	6.9214
e	1000:2000:3000:4000:5000	1/5	3000	100	6.2310	6.4303	6.1294	5.5333	6.2421	6.1155
f	2000:2500:3000:3500:4000	1/5	3000	30	4.0018	3.8181	3.9631	3.7477	4.0124	3.9092
g	2500:2750:3000:3250:3500	1/5	3000	6.25	3.2219	3.1831	3.2099	3.1364	3.2227	3.1896
h	2750:2875:3000:3125:3250	1/5	3000	1.5625	2.9946	2.9766	2.9915	2.9662	2.9925	2.9845

[1]All variances have been multiplied by 10^5.

random numbers as before and for the following sets of r_j, r_j^* and p_j

(i) $r_j = (1, 3)$, $r_j^* = (0, 10456.72)$ and $p_j = 1/2$

(ii) $r_j = (1, 2, 3)$, $r_j^* = (0, 5228.36, 10456.72)$ and $p_j = 1/3$

(iii) $r_j = (4, 5, 6)$, $r_j^* = (15685.08, 20913.44, 26141.80)$ and $p_j = 1/3$

are calculated and they are summarized in Table XII.

It is observed that: (i) the multiplicative model is much more efficient than the additive model; (ii) the variance of $\hat{\bar{X}}$ for the additive model depends only on the values of the selected r_i, $i = 1, 2, \ldots, n$. It does not depend on the values of the corresponding X_i but the variance of $\hat{\bar{X}}$ for the multiplicative model does.

Now, consider the fact that if there is very much difference in the p_j, or if the variance of r_j or r_j^* is too small, this will cause embarrassment to the respondents and hence decrease the likelihood of truthful answers. The solution proposed is to use p_j equal to $1/t$ and the values of $r_j\bar{X}$ or $r_j^* + \bar{X}$ should not be less than double the expected value of $\bar{X}$.

Model VII

Description of the Model and Estimation of the Parameters

The randomizing device used in this model is a deck of 11 kinds of cards. Ten of the cards have the integers 0 to 9 on them, while the eleventh has "Give the true answer" on it. Let m_1 represent the number of cards that say "Give the true answer" and let the number of

TABLE XII

COMPARISON OF MODEL V AND MODEL VI FOR DIFFERENT COMBINATIONS OF r_j, $p_j = 1/t$

r_j, r_j^*	Sample	$\hat{\hat{X}}$		$Var(\hat{\hat{X}})$[1]		c.v.(%)	
		Model V	Model VI	Model V	Model VI	Model V	Model VI
$r_j = 1:3$	1	4542.00	4182.69	2.1238	11.1661	10.1463	25.2636
	2	5134.44	5437.49	2.2543	11.6035	9.2472	19.8105
$r_j^* = 0:10456.72$	3	4899.52	5019.22	2.0985	11.6035	9.3498	21.4614
	4	5308.82	5855.76	2.1141	11.4577	8.6609	18.2795
	5	5057.80	5437.49	1.8852	11.6035	8.5845	19.8105
$(p_j = 1/2)$	Average	4988.52	5186.53	2.0952	11.4869	9.1757	20.6645
$r_j = 1:2:3$	1	5173.82	5228.36	1.4754	7.5213	7.4241	16.5875
	2	4923.60	5646.63	1.0718	8.3596	6.6493	16.1921
$r_j^* = 0:5228.36:10456.72$	3	4985.14	5228.36	1.2132	8.4325	6.9870	17.5636
	4	4909.24	5019.22	0.9755	7.0475	6.3621	16.7256
	5	5217.06	5646.63	1.1219	7.4484	6.4204	15.2842
$(p_j = 1/3)$	Average	5041.77	5353.84	1.1716	7.7619	6.7889	16.4558
$r_j = 4:5:6$	1	5206.54	5228.36	0.2589	7.5213	3.0904	16.5875
	2	5106.48	5646.63	0.1451	8.3596	2.3587	16.1921
$r_j^* = 15685.08:20913.44$	3	5131.10	5228.36	0.1267	8.4325	2.1934	17.5636
$:26141.80$	4	5100.74	5019.22	0.1158	7.0475	2.1095	16.7256
	5	5239.07	5646.63	0.1340	7.4484	2.2162	15.2842
$(p_j = 1/3)$	Average	5153.74	5353.84	0.1561	7.7619	2.4242	16.4558

[1]All variances have been multiplied by 10^5.

cards for each number (0, 1, ..., 9) be one.

At the interview, the respondent is asked to make a random choice of a card from the deck and answers in accordance with the following instructions. If it is a card saying "Give the true answer", the respondent will be asked to answer the sensitive question. If the card shows any numbers, the respondent will be asked to write down this number, say 3, take all the cards saying "Give the true answer" out of the deck and draw $(m_2 - 1)$ more cards, one by one with replacement, say 0 and 5 (m_2 is the number of digits in the number we want to estimate, in our case $m_2 = 3$). The respondent simply tells the number 305 to the interviewer.

It might appear at first glance that this procedure would not be convenient for the respondents who get the card with any numbers for the first time. The cost of interviewing may also be high because the interviewers have to spend more time explaining how to use the randomizing device to the respondents, and some respondents may need more time for taking the cards out. In order to solve these problems, we may use two decks of cards. Both of them are the same except the second one has no "Give the true answer" cards. We may also use a ten sided die instead of using the second deck of cards with no "Give the true answer" cards.

It seems clear enough that this model can be applied to both discrete and continuous data. There is no problem in applying it to discrete data such as the number of automobile accidents during the past year, number of involvements with the courts, number of induced abortions and number of times the respondent has used heroin in a specified time period. But for continuous data such as personal income, the persons who give the rough estimate of their income (e.g., \$15,600) and high income persons are going to be known to be telling the truth when they answer the question truthfully. Alternatively, low income persons will also be easy to identify. In such a situation, this method may not work well.

In an attempt to estimate our parameters, let us define P_i as the true proportion of respondents who possess the quantitative measure "i".

Then the probability that a respondent randomly selected from a population will answer "i" is

$$p_i = P_i\left(\frac{m_1}{m_1 + 10}\right) + \left(\frac{1}{m_1 + 10}\right)\left(\frac{1}{10^{m_2-1}}\right). \tag{3.30}$$

If a random sample of size n is taken, let n_i denote the number of respondents answering that they possess the quantitative measure "i", then the maximum likelihood estimate of P_i, its estimated variance and covariance are

$$\hat{P}_i = \left(\frac{m_1 + 10}{m_1}\right)\left(\frac{n_i}{n} - \frac{1}{(m_1 + 10)10^{m_2-1}}\right) \tag{3.31}$$

$$\widehat{\text{Var}}(\hat{P}_i) = \left(\frac{m_1 + 10}{m_1}\right)^2 \frac{1}{n}\left(\frac{n_i}{n}\right)\left(1 - \frac{n_i}{n}\right) \tag{3.32}$$

$$\widehat{\text{Cov}}(\hat{P}_i, \hat{P}_j) = -\left(\frac{m_1 + 10}{m_1}\right)^2 \frac{1}{n}\left(\frac{n_i}{n}\right)\left(\frac{n_j}{n}\right). \tag{3.33}$$

The same results are obtained by substituting $\frac{n_i}{n}$ as an estimate of p_i in the expression (3.30).

As in Model III and Model IV, let X_i be the value of the quantitative measure "i", then the estimated mean of the quantitative measures is

$$\hat{\mu} = \frac{m_1 + 10}{m_1} \sum_i \left(\frac{n_i}{n} - \frac{1}{(m_1 + 10)10^{m_2 - 1}} \right) X_i \tag{3.34}$$

with

$$\widehat{\text{Var}}(\hat{\mu}) = \frac{1}{n^3} \left(\frac{m_1 + 10}{m_1} \right)^2 \left[\sum_i n_i (n - n_i) X_i^2 - 2 \sum_{i<j} n_i n_j X_i X_j \right]. \tag{3.35}$$

According to the expression (3.35), it is immediately clear that if the number of "Give the true answer" cards increases and the total number of cards is fixed, then the efficiency of the estimate will increase. For the purpose of securing the best cooperation by the respondents with the highest efficiency, the number of "Give the true answer" cards should not be much greater than ten.

Extension of the Randomized Response Models for Quantitative Data to Two Trials per Respondent

There are at least three techniques one might consider in this paper. Some techniques can be applied to every randomized response model for quantitative data, but some techniques can be applied only to some models, or to a specific model. For the purpose of this study, we shall illustrate only those techniques which apply to each of the randomized response models for quantitative data derived in Chapter III.

Model III with Two Trials per Respondent

Suppose that every person in a population belongs to one of t mutually exclusive groups, and it is required to estimate the population mean.

Let each respondent make two independent selections using the same randomizing device, then for an individual from whom two responses are required there are t^2 possible responses:

(Group 1, Group 1), (Group 1, Group 2), ..., (Group 1, Group t)

(Group 2, Group 1), (Group 2, Group 2), ..., (Group 2, Group t)

....

(Group t, Group 1), (Group t, Group 2), ..., (Group t, Group t)

Let n_{ij} represent the number of individuals answering (Group i, Group j) in the sample.

Let n be the sample size, then the estimate for the probability that a respondent will answer "Group i", $\hat{p}_i$, is

$$\hat{p}_i = \frac{1}{2n} \sum_{k=1}^{t} (n_{ik} + n_{ki}) \qquad (3.36)$$

with

$$\sum_{i=1}^{t} \hat{p}_i = 1.$$

Since the probability that a respondent will answer "Group i" is

$$p_i = P_i\left[1 - \sum_{i=1}^{t} Pr(r_i)\right] + Pr(r_i),$$

then

$$\frac{1}{2n}\sum_{k=1}^{t}(n_{ik} + n_{ki}) = \hat{P}_i\left[1 - \sum_{i=1}^{t} Pr(r_i)\right] + Pr(r_i).$$

Solving for $\hat{P}_i$, we find

$$\hat{P}_i = \frac{\frac{1}{2n}\sum_{k=1}^{t}(n_{ik} + n_{ki}) - Pr(r_i)}{1 - \sum_{i=1}^{t} Pr(r_i)} . \quad (3.37)$$

As before, the estimated population mean of the quantitative measures and its variance are

$$\hat{\mu} = \sum_{i=1}^{t} \hat{P}_i X_i$$

and

$$Var(\hat{\mu}) = \sum_{i=1}^{t} X_i^2 Var(\hat{P}_i) + 2 \sum_{i<j=1}^{t} X_i X_j Cov(\hat{P}_i, \hat{P}_j),$$

where X_i denotes the midpoint of the i^{th} interval of the quantitative measure belonging to "Group i".

The estimated variance of $\hat{\mu}$ is

$$\widehat{Var}(\hat{\mu}) = \sum_{i=1}^{t} X_i^2 \widehat{Var}(\hat{P}_i) + 2 \sum_{i<j=1}^{t} X_i X_j \widehat{Cov}(\hat{P}_i, \hat{P}_j) \quad (3.38)$$

where

$$\widehat{Var}(\hat{P}_i) = \frac{1}{\left[1 - \sum_{i=1}^{t} Pr(r_i)\right]^2}\left(\frac{1}{2n}\right)\left[\frac{1}{2n}\sum_{k=1}^{t}(n_{ik} + n_{ki})\right]\left[1 - \frac{1}{2n}\sum_{k=1}^{t}(n_{ik} + n_{ki})\right] \tag{3.39}$$

and

$$\widehat{Cov}(\hat{P}_i,\hat{P}_j) = -\frac{1}{\left[1 - \sum_{i=1}^{t} Pr(r_i)\right]^2}\left(\frac{1}{2n}\right)\left[\frac{1}{2n}\sum_{k=1}^{t}(n_{ik} + n_{ki})\right]\left[\frac{1}{2n}\sum_{k=1}^{t}(n_{jk} + n_{kj})\right] . \tag{3.40}$$

Model IV with Two Trials per Respondent

Let each respondent from each set of $t - 1$ simple random samples of size n_i, $i = 1, 2, \ldots, t - 1$, make two independent selections using the same randomizing device. Then for an individual from whom two responses are required there are four possible responses: (Yes, Yes), (Yes, No), (No, Yes) and (No, No).

Let n_{ijk11}, n_{ijk10}, n_{ijk01} and n_{ijk00} be the numbers of individuals answering (Yes, Yes), (Yes, No), (No, Yes) and (No, No) respectively to (Group j, Group k) of the i^{th} set of sample. Then the estimate for p in the i^{th} sample, $\hat{p}_i^*$, will be

$$\hat{p}_i^* = \sum_{j=1}^{t} \sum_{k\geq j=1}^{t} \frac{2n_{ijk11} + n_{ijk10} + n_{ijk01}}{2n_i} . \tag{3.41}$$

The derivation of the $\hat{\underline{P}}_S$, $\widehat{Var}(\hat{\underline{P}}_S)$, $\hat{\mu}$ and $\widehat{Var}(\hat{\mu})$ proceeds along the same lines as that of those estimators in Model IV. The results of the derivation are

$$\hat{\underline{P}}^*_S = \underline{p}^{-1}\underline{n}^* \quad (3.42)$$

$$\widehat{\mathrm{Var}}(\hat{\underline{P}}^*_S) = \underline{p}^{-1}\underline{v}^*(\underline{p}^{-1})' \quad (3.43)$$

$$\hat{\mu}^* = \left(J_1^{t-1}\right)'\underline{\underline{Xp}}^{-1}\underline{n}^* + X_t\left(1 - \sum_{i=1}^{t-1}\hat{P}_i\right) \quad (3.44)$$

$$\widehat{\mathrm{Var}}(\hat{\mu}^*) = \left(J_1^{t-1}\right)'\underline{\underline{Xp}}^{-1}\underline{v}^*(\underline{p}^{-1})'\underline{X}J_1^{t-1} + X_t^2\left(J_1^{t-1}\right)'\underline{p}^{-1}\underline{v}^*(\underline{p}^{-1})'J_1^{t-1} \quad (3.45)$$

where

$$\underline{n}^* = \begin{bmatrix} \sum_{j=1}^{t}\sum_{k\geq j=1}^{t}\frac{2n_{1jk11}+n_{1jk10}+n_{1jk01}}{2n_1} - p_{1t} \\ \sum_{j=1}^{t}\sum_{k\geq j=1}^{t}\frac{2n_{2jk11}+n_{2jk10}+n_{2jk01}}{2n_2} - p_{2t} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \sum_{j=1}^{t}\sum_{k\geq j=1}^{t}\frac{2n_{(t-1)jk11}+n_{(t-1)jk10}+n_{(t-1)jk01}}{2n_{t-1}} - p_{(t-1)t} \end{bmatrix} \quad (3.46)$$

$$\underline{v}^* = \begin{bmatrix} v_{11}^* & & & \phi \\ & v_{22}^* & & \\ & & \ddots & \\ \phi & & & v_{(t-1)(t-1)}^* \end{bmatrix} \quad (3.47)$$

$$v_{ii}^* = \frac{1}{2n_i}\left(\sum_{j=1}^{t}\sum_{k\geq j=1}^{t}\frac{2n_{ijk11}+n_{ijk10}+n_{ijk01}}{2n_i}\right) \times\left(1 - \sum_{j=1}^{t}\sum_{k\geq j=1}^{t}\frac{2n_{ijk11}+n_{ijk10}+n_{ijk01}}{2n_i}\right) \quad (3.48)$$

Model V, Model VI and Model VII with

Two Trials per Respondent

The estimators using these three models with two trials per respondent can be found by the same method as that of the estimators obtained from the alternative method of applying repeated trials to the randomized response models for proportions in Chapter II.

Referring to (2.56) and (2.57), the variance weighted estimator of μ and its estimated variance for these three models are

$$\hat{\mu} = \frac{\hat{\mu}_1 \widehat{Var}(\hat{\mu}_2) + \hat{\mu}_2 \widehat{Var}(\hat{\mu}_1)}{\widehat{Var}(\hat{\mu}_1) + \widehat{Var}(\hat{\mu}_2)} \tag{3.49}$$

and

$$\widehat{Var}(\hat{\mu}) = \frac{\widehat{Var}(\hat{\mu}_1)\widehat{Var}(\hat{\mu}_2)}{\widehat{Var}(\hat{\mu}_1) + \widehat{Var}(\hat{\mu}_2)} \tag{3.50}$$

$$\leq \text{Minimum}[\widehat{Var}(\hat{\mu}_1), \widehat{Var}(\hat{\mu}_2)] \tag{3.51}$$

where $\hat{\mu}_i$ is the estimator of the population mean in which the sampled respondents used the i^{th} randomizing device, $i = 1, 2$.

CHAPTER IV

DETERMINATION OF THE SAMPLE SIZES

Estimation of the Population Proportion

One of the main advantages of a sample survey is that it is possible to ensure approximately a pre-specified margin of error in the sample results by suitably fixing the sample sizes. Suppose we wish to determine the sample size for each randomized response model for proportions with a given precision $d = |\hat{P}_S - P_S|$ with probability $(1 - \alpha)$, that is, we want

$$\Pr(|\hat{P}_S - P_S| \leq d) \geq 1 - \alpha \tag{4.1}$$

By assuming that the sample proportion, $\hat{P}_S$, is normally distributed, then the formula that connects the sample size n with the desired degree of precision is

$$d = Z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{P}_S)} \tag{4.2}$$

where $Z_{\alpha/2}$ is the ordinate of the normal curve that cuts off an area $\alpha/2$ in one tail. We note that $\mathrm{Var}(\hat{P}_S)$ will depend on n and so we can solve (4.2) for n, but this will be a function of P_S. We also note that if P_S is too close to zero or one, the normality assumption will not be valid.

For practical use, an estimate $\hat{P}_S$ of P_S is substituted in this formula. In practice, there are three ways of estimating population variances for sample size determinations: (1) by the results of a pilot survey; (ii) by previous sampling of the same or a similar population; and (iii) by guess work about the structure of the population, assisted by some mathematical results. These three methods are discussed in (15).

Making use of (4.2) and $Var(\hat{P}_S)$ for each proposed randomized response model for proportions, the formulae for n in each model are summarized in Table XIII.

Estimation of the Population Mean

To determine the sample size for estimating the population mean for each model for a given precision $d = |\hat{\mu} - \mu|$ with probability $(1 - \alpha)$ we wish to have

$$Pr(|\hat{\mu} - \mu| \leq d) \geq 1 - \alpha. \tag{4.3}$$

Again, by assuming that $\hat{\mu}$ is normally distributed, the formula that connects the sample size n with the desired degree of precision is

$$d = Z_{\alpha/2}\sqrt{Var(\hat{\mu})}. \tag{4.4}$$

The sample size n could be more precisely evaluated from

$$d = t(\alpha/2,\ n - 1)\sqrt{Var(\hat{\mu})}. \tag{4.5}$$

TABLE XIII

SAMPLE SIZES FOR SOME RANDOMIZED RESPONSE MODELS FOR PROPORTIONS

Model	n	Note
I with $p_{IS} = 1$	$\frac{Z^2_{\alpha/2}}{d^2}(1 - P_S)\left(P_S + \frac{m_2}{m_1}\right)$	
I with $p_{IS} \neq 1$, $p_{IN} = 1$	$\frac{Z^2_{\alpha/2}}{d^2}\frac{p'(1 - p')}{\left(p_{IS} - \frac{1}{2}\right)^2}$	$p' = P_S(p_{IS} - \frac{1}{2}) + \frac{1}{2}$
I with $n = n_c + n_r$, $k = 1$	$\frac{Z_{\alpha/2}}{d}\left[n_c\left(P_S + \frac{m_2}{m_1}\right)(1 - P_S)\right]^{\frac{1}{2}}$	n_r sampled respondents refuse to answer the questions
I with Two Trials per Respondent	$\frac{Z^2_{\alpha/2}}{2d^2}\left(P_S + \frac{m_2}{m_1}\right)(1 - P_S)$	
II with Sampling Without Replacement	$\frac{Z^2_{\alpha/2}}{kd^2}\left[kP_S\left\{1 - kP_S - 2\binom{m - m_1}{2}\right\} + \binom{m - m_1}{2}\left\{1 - \binom{m - m_1}{2}\right\}\right]$	$k = \binom{m}{2} - \binom{m - m_1}{2}$

TABLE XIII (Continued)

Model	n	Note
II with Sampling with Replacement	$$\frac{Z^2_{\alpha/2}}{k'^2 d^2}\left[k'P_S\left\{1 - k'P_S - 2\left(1 - \frac{m_1}{m}\right)^2\right\} + \left(1 - \frac{m_1}{m}\right)^2\left\{1 - \left(1 - \frac{m_1}{m}\right)^2\right\}\right]$$	$k' = \frac{m_1}{m}\left(2 - \frac{m_1}{m}\right)$
II with Multi-Proportions	$$\frac{Z^2_{\alpha/2}}{d_j^2}\ \frac{p_j(1 - p_j)}{\left[1 - \sum_{i=1}^{t} Pr(r_i)\right]^2}$$	$p_j = P_j\left[1 - \sum_{i=1}^{t} Pr(r_i)\right] + Pr(r_i)$ $d_j = \lvert\hat{P}_j - P_j\rvert$ $Var(\hat{P}_j) \geq Var(\hat{P}_i),\ i \neq j$
Regular	$$\frac{Z^2_{\alpha/2}}{d^2}\ (P_S P_S)(1 - P_S P_S)$$	

TABLE XIII (Continued)

Model	n	Note
Warner	$\frac{Z^2_{\alpha/2}}{d^2}\left[P_S(1 - P_S) + \frac{p(1 - p)}{(2p - 1)^2}\right]$	$p \neq \frac{1}{2}$
Unrelated Questions (P_N Known)	$\frac{Z^2_{\alpha/2}}{d^2} \; \frac{C(1 - C)}{p^2}$	$C = pP_S + (1 - p)P_N$
Multiple Trials	$\frac{P_S(1 - P_S)}{\frac{d^2}{Z^2_{\alpha/2}} - \frac{T - k}{T - 1}\left[\frac{p(1 - p)}{(2p - 1)^2}\right]}$	$p \neq \frac{1}{2}$

However, $t(\alpha/2, n - 1)$ depends on n. As a result n is under-estimated, since $Z_{\alpha/2}$ is less than $t(\alpha/2, n - 1)$. The obvious correction which suggests itself is to increase the value of n in the ratio $t^2(\alpha/2, n - 1)/Z^2_{\alpha/2}$, where n is evaluated from (4.4) but the correction is not likely to be important unless n is small.

The calculation of n for estimating the population mean also assumes knowledge of $Var(\hat{\mu})$ when the error, d, permissible in the estimate of the population value of the mean is given. Proceeding exactly as in the case of estimating the population proportion, we obtain the formulae for sample sizes for Model III to Model VII. The formulae are given in Table XIV for each of the models.

The Extra Cost in Terms of Sample Size for Model I as Compared with the Regular Model

The objective of this study is to investigate the extra cost in terms of sample size for the randomized response models if the mean square error of the regular estimate is the same as the variance of the randomized response estimate. In the investigation given in this paper we shall use Model I as our randomized response model.

Suppose that members of the sensitive group tell the truth with probability one in Model I and tell the truth with probability p_S in the regular model.

If it is desired that the mean square error of the regular estimator be the same as the variance of the Model I estimator, that is

TABLE XIV

SAMPLE SIZES FOR MODEL III TO MODEL VII

Model	n	Note
III	$\frac{Z^2_{\alpha/2}}{d^2}\left[\frac{\sum_{i=1}^{t} X_i^2 p_i(1 - p_i) - 2\sum_{i<j=1}^{t} X_i X_j p_i p_j}{\left(1 - Pr(r_i)\right)^2}\right]$	$p_i = P_i\left[1 - \sum_{i=1}^{t} Pr(r_i)\right] + Pr(r_i)$
IV	$\frac{Z^2_{\alpha/2}}{d^2}\left(\frac{m_1 + t}{m_1}\right)^2\left[\sum_{i=1}^{t} X_i^2 p_i(1 - p_i) - 2\sum_{i<j=1}^{t} X_i X_j p_i p_j\right]$	$p_i = P_i\left(\frac{m_1}{m_1 + t}\right) + \frac{1}{m_1 + t}$
V	$\frac{Z^2_{\alpha/2} S^2_Y}{d^2\left(\sum_{j=1}^{t} r_j p_j\right)^2}$	S^2_Y is the estimate of the variance of Y.
VI	$\frac{(Z^2_{\alpha/2} S^2_X)/d^2}{1 - (Z^2_{\alpha/2}/d^2)\,Var(\bar{r}^*)}$	$\bar{r}^* = \frac{1}{n}\sum_{i=1}^{n} r_i^*$ S^2_X is the estimate of the variance of X.

TABLE XIV (Continued)

Model	n	Note
VII	$\frac{Z^2_{\alpha/2}}{d^2}\left[\left(\frac{m_1 + 10}{m_1}\right)^2 \left\{\sum_i X_i^2 p_i(1 - p_i) - 2 \sum_{i<j} X_i X_j p_i p_j\right\}\right]$	$p_i = \frac{1}{m_1 + 10}\left(m_1 P_i + 10^{1-m_2}\right)$

$$\frac{1}{n_I}\left[P_S(1 - P_S) + \frac{m_2}{m_1}(1 - P_S)\right] = \frac{1}{n_R}(P_S p_S)(1 - P_S p_S) + [P_S(p_S - 1)]^2 \quad (4.6)$$

where n_I and n_R are the sample sizes in Model I and the regular model respectively.

In order to find the extra cost in terms of sample size, the ratio of n_I/n_R is not fixed for specified values of m_2/m_1, P_S and p_S. For fixed m_2/m_1, P_S and p_S the value of n_I/n_R will change as n_I (or n_R) is varied. This is because $\hat{P}_S$ is a biased estimate of P_S for the regular model and this bias does not depend on n_R. Based upon the knowledge about the selection of m_2/m_1 in the Model I technique (see Model I), we shall fix m_2/m_1 by choosing for it a value as close to zero as is practicable without arousing suspicion in the respondent. This would probably be no smaller than 0.20. Table XV and Figure 6 then exhibit the ratios of the sample sizes of Model I estimates to the sample sizes of regular estimates for various values of P_S and p_S under the assumption that the respondents tell the truth in the randomized method but only tell the truth in the non-randomized method with probability p_S. The sample size n_R is set at 100, 500, and 1,000 in each case.

Referring to Table XV and Figure 6, it is observed that, by fixing m_2/m_1 to be 0.2, an extra cost in terms of sample size for Model I will be incurred when:

(i) P_S is less than 0.174 provided that $p_S = 0.9$ and $n_R =$ 1,000.

(ii) P_S is less than 0.202 provided that $p_S = 0.7$ and $n_R =$ 100.

TABLE XV

THE VALUES OF n_I/n_R IN THE CASE WHEN THE MEAN SQUARE ERROR OF THE REGULAR ESTIMATOR IS THE SAME AS THE VARIANCE OF THE MODEL I ESTIMATOR, m_2/m_1 = 0.2

P_S	n_R	P_S				
		0.1	0.3	0.5	0.7	0.9
	100	0.3293	0.5201	0.9076	1.7408	2.9380
0.1	500	0.0665	0.1089	0.2081	0.5242	2.0470
	1,000	0.0333	0.0548	0.1060	0.2798	1.4843
	100	0.0478	0.0779	0.1472	0.3586	1.2191
0.3	500	0.0096	0.0158	0.0308	0.0830	0.5409
	1,000	0.0048	0.0079	0.0155	0.0423	0.3190
	100	0.0172	0.0283	0.0543	0.1413	0.7035
0.5	500	0.0034	0.0057	0.0111	0.0305	0.2337
	1,000	0.0017	0.0028	0.0056	0.0154	0.1274
	100	0.0068	0.0112	0.0216	0.0579	0.3734
0.7	500	0.0014	0.0022	0.0044	0.0121	0.1006
	1,000	0.0007	0.0011	0.0022	0.0061	0.0526
	100	0.0017	0.0027	0.0054	0.0146	0.1141
0.9	500	0.0003	0.0005	0.0011	0.0030	0.0262
	1,000	0.0002	0.0003	0.0005	0.0015	0.0133

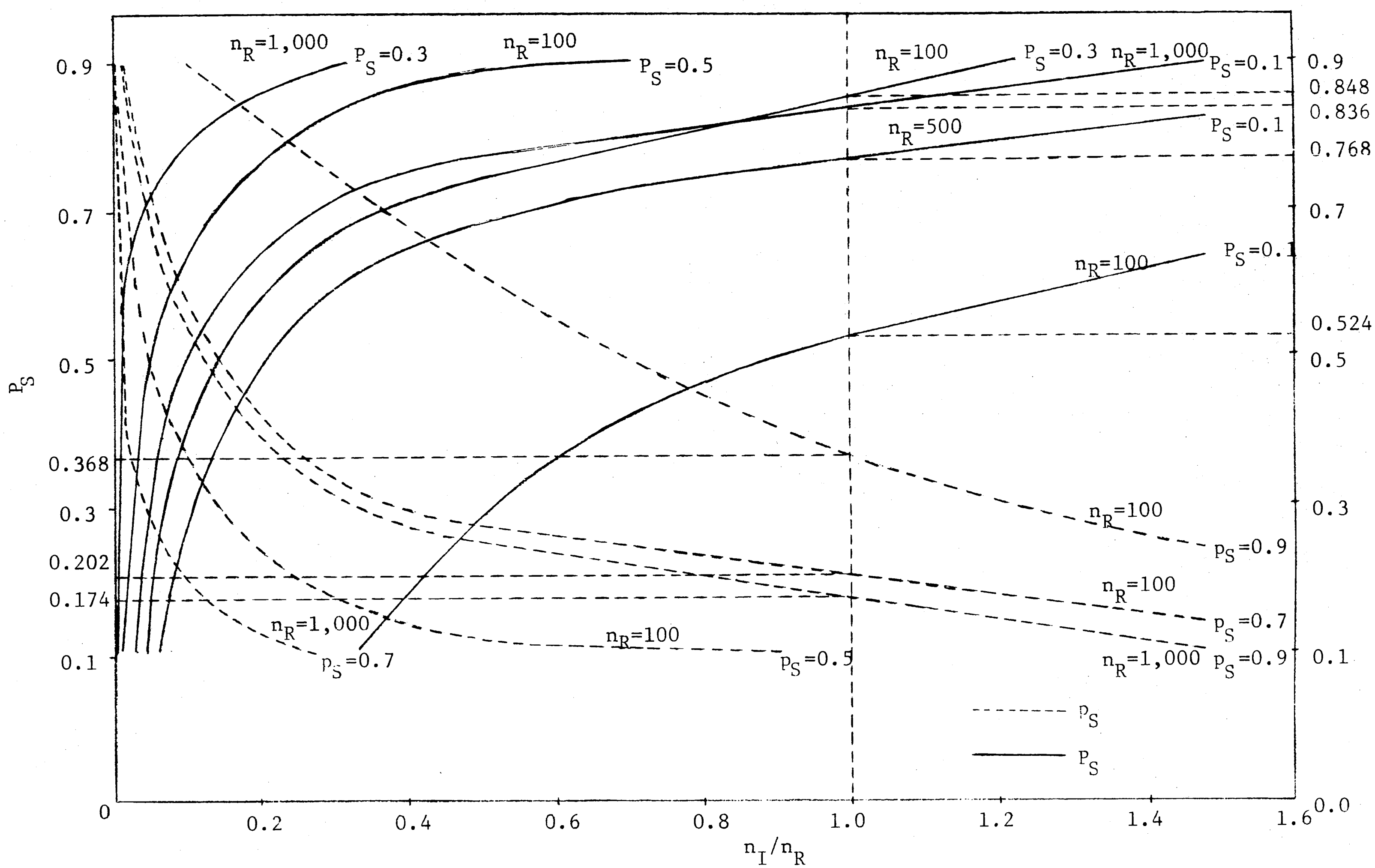

Figure 6. The Values of n_I/n_R in the Case When the Mean Square Error of the Regular Estimator is the Same as the Variance of Model I Estimator, $m_2/m_1 = 0.2$

(iii) P_S is less than 0.368 provided that $p_S = 0.9$ and $n_R = 100$.

(iv) p_S is greater than 0.524 provided that $P_S = 0.1$ and $n_R = 100$.

(v) p_S is greater than 0.768 provided that $P_S = 0.1$ and $n_R = 500$.

(vi) p_S is greater than 0.836 provided that $P_S = 0.1$ and $n_R = 1,000$.

(vii) p_S is greater than 0.848 provided that $P_S = 0.3$ and $n_R = 100$.

Suppose the costs per sampling unit of Model I and the regular model are c_I and c_R respectively. In most circumstances it will be reasonable to expect c_I to be greater than c_R, since in any randomized response method the interviewers will require more training and will spend more time on each interview. Thus, Model I will be cheaper to run whenever $c_I n_I$ is less than $c_R n_R$. We have computed the values of n_I/n_R for various values of n_R, P_S, p_S and for $m_2/m_1 = 0.2$ in Table XV . From this table, one can determine the situations in which Model I is cheaper to use than the regular model for the same accuracy. These situations are characterized by those parameter values for which n_I/n_R is less than c_R/c_I.

CHAPTER V

SUMMARY AND CONCLUSIONS

On the basis of the empirical investigations of Model I, the Warner model, the unrelated questions model and the multiple trials model, it is concluded that the three competitors are Model I, the unrelated questions model and the multiple trials model. If forced to a decision at this time, Model I would probably be chosen as the preferred design. This is because the procedure for administering Model I is simpler than the other two models. If the comparisons are made among Model I and Model II, the efficiency of Model II would be expected to be very similar to the efficiency of Model I but the interviewing costs are expected to be higher. For randomized response models for discrete quantitative data, it seems clear enough that Model VII would be preferable to the other four models. Model III would also be preferable to Model IV but it is somewhat surprising to find that Model V is much more efficient than Model VI. In the development given in Model V, although exact distribution cannot be obtained in general, reasonable approximations can be made yielding a procedure adequate to the situation encountered in practice.

This study has shown that in the case when some sampled respondents do not report truthfully or refuse to answer the questions, or the repeated trials per respondent are used, the randomized response models can be extended. It is also shown that in the case when each sampling unit is composed of a group of elements and it is the elements that are classified, the formula for estimating the true proportion of the respondents who belong to the sensitive group is the same as the formula used in the randomized response model for proportion with one respondent as a sampling unit. But the formula for the variance of the estimator has to be changed by using the approximated variance of the ratio estimator.

As in the general survey designs, the sample size for each randomized response design for a given precision can be determined. However, if it is desired that the mean square error of the regular estimator be the same as the variance of randomized response estimator, the extra cost in terms of sample size is not fixed but varies as the sample size of the regular model or of the randomized response model is changed.

A SELECTED BIBLIOGRAPHY

(1) Warner, S. L. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." Journal of the American Statistical Association, Vol. 60 (1965), 63-69.

(2) Abul-Ela, A. A., Greenberg, B. G., and Horvitz, D. G. "A Multi-Proportions Randomized Response Model." Journal of the American Statistical Association, Vol. 62 (1967), 990-1008.

(3) Horvitz, D. G., Shah, B. V., and Simmons, W. R. "The Unrelated Questions Randomized Response Model." Social Statistics Section Proceedings of the American Statistical Association (1967), 65-72.

(4) Moors, J. J. A. "Optimization of the Unrelated Questions Randomized Response Model." Journal of the American Statistical Association, Vol. 66 (1971), 627-629.

(5) Folsom, R. E., Greenberg, B. G., Horvitz, D. G., and Abernathy, J. R. "The Two Alternative Questions Randomized Response Model for Human Surveys." Journal of the American Statistical Association, Vol. 68 (1973), 525-530.

(6) Greenberg, B. G., Horvitz, D. G., and Abernathy, J. R. "A Comparison of Randomized Response Designs." S. I. A. M. Review, Vol. 16 (1974), 787-815.

(7) Greenberg, B. G., Kuebler, R. R., Jr., Abernathy, J. R., and Horvitz, D. G. "Application of the Randomized Response Technique in Obtaining Quantitative Data." Journal of the American Statistical Association, Vol. 66 (1971), 243-250.

(8) Eriksson, S. A. "A New Model for Randomized Response." International Statistical Review, Vol. 41 (1973), 101-113.

(9) Poole, K. W. "Estimation of the Distribution Function of a Continuous Type Random Variable Through Randomized Response." Journal of the American Statistical Association, Vol. 69 (1974), 1002-1005.

(10) Liu, P. T., and Chow, L. P. "A New Discrete Quantitative Randomized Response Model." Journal of the American Statistical Association, Vol. 71 (1976), 72-73.

(11) Warner, S. L. "The Linear Randomized Response Model." Journal of the American Statistical Association, Vol. 66 (1971), 884-888.

(12) Campbell, C., and Joiner, B. L. "How to Get the Answer Without Being Sure You've Asked." American Statistician, Vol. 27 (1973), 229-237.

(13) Greenberg, B. G., Abul-Ela, A. A., Simmon, W. R., and Horvitz, D. G. "The Unrelated Questions Randomized Response Model: Theoretical Framework." Journal of the American Statistical Association, Vol. 64 (1969), 520-539.

(14) Goodstadt, M. S., and Gruson, V. "The Randomized Response Technique: A Test on Drug Use." Journal of the American Statistical Association, Vol. 70 (1975), 814-818.

(15) Cochran, W. G. Sampling Techniques, 2nd ed. New York: John Wiley and Sons, 1963.

(16) Finkner, A. L. "Methods of Sampling for Estimating Commercial Peach Production in North Carolina." North Carolina Agricultural Experiment Station Technical Bulletin, Vol. 91 (1950), 52-59.

(17) Hendrick, W. A. "Adjustment for Bias by Non-Response in Mailed Survey." Agricultural Economics Research, Vol. 1 (1949), 52.

APPENDIXES

TABLE XVI

COMPARISON OF MODEL I AND THE REGULAR MODEL WITH P_S = 0.1 AND n = 100

Regular Estimate		MSE (Model I)/MSE (Regular) m_2/m_1									
P_S	Bias	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
1.0	0.00	11.0000	10.0000	9.0000	8.0000	7.0000	6.0000	5.0000	4.0000	3.0000	2.0000
0.9	-0.01	10.7726	9.7932	8.8139	7.8346	6.8553	5.8759	4.8966	3.9173	2.9380	1.9586
0.8	-0.02	8.7148	7.9225	7.1303	6.3380	5.5458	4.7535	3.9613	3.1690	2.3768	1.5845
0.7	-0.03	6.3830	5.8027	5.2224	4.6422	4.0619	3.4816	2.9013	2.3211	1.7408	1.1605
0.6	-0.04	4.5749	4.1590	3.7431	3.3272	2.9113	2.4954	2.0795	1.6636	1.2477	0.8318
0.5	-0.05	3.3277	3.0252	2.7227	2.4202	2.1176	1.8151	1.5126	1.2101	0.9076	0.6050
0.4	-0.06	2.4849	2.2590	2.0331	1.8072	1.5813	1.3554	1.1295	0.9036	0.6777	0.4518
0.3	-0.07	1.9071	1.7338	1.5604	1.3870	1.2136	1.0403	0.8669	0.6935	0.5201	0.3467
0.2	-0.08	1.5009	1.3645	1.2280	1.0916	0.9551	0.8187	0.6822	0.5458	0.4093	0.2729
0.1	-0.09	1.2075	1.0976	0.9879	0.8781	0.7684	0.6586	0.5488	0.4391	0.3293	0.2195
0.0	-0.10	0.9900	0.9000	0.8100	0.7200	0.6300	0.5400	0.4500	0.3600	0.2700	0.1800

TABLE XVII

COMPARISON OF MODEL I AND THE REGULAR MODEL WITH P_S = 0.3 AND n = 100

Regular Estimate		MSE (Model I)/MSE (Regular) m_2/m_1									
P_S	Bias	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
1.0	0.00	4.3333	4.0000	3.6667	3.3333	3.0000	2.6667	2.3333	2.0000	1.6667	1.3333
0.9	-0.09	3.1696	2.9258	2.6820	2.4382	2.1943	1.9505	1.7067	1.4629	1.2191	0.9753
0.8	-0.18	1.6778	1.5487	1.4196	1.2906	1.1615	1.0324	0.9034	0.7743	0.6453	0.5162
0.7	-0.27	0.9325	0.8607	0.7890	0.7173	0.6455	0.5738	0.5021	0.4304	0.3586	0.2869
0.6	-0.36	0.5732	0.5291	0.4850	0.4409	0.3968	0.3527	0.3086	0.2645	0.2204	0.1764
0.5	-0.45	0.3827	0.3533	0.3239	0.2944	0.2650	0.2355	0.2061	0.1766	0.1472	0.1178
0.4	-0.54	0.2720	0.2511	0.2301	0.2052	0.1883	0.1674	0.1465	0.1255	0.1046	0.0837
0.3	-0.63	0.2026	0.1870	0.1714	0.1558	0.1402	0.1247	0.1091	0.0935	0.0779	0.0623
0.2	-0.72	0.1564	0.1444	0.1324	0.1203	0.1083	0.0963	0.0842	0.0722	0.0602	0.0481
0.1	-0.81	0.1243	0.1148	0.1052	0.0956	0.0861	0.0765	0.0669	0.0574	0.0478	0.0382
0.0	-0.90	0.1011	0.0933	0.0855	0.0778	0.0700	0.0622	0.0544	0.0467	0.0389	0.0311

TABLE XVIII

COMPARISON OF MODEL I AND THE REGULAR MODEL WITH P_S = 0.5 AND n = 100

Regular Estimate		MSE (Model I)/MSE (Regular) m_2/m_1									
P_S	Bias	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
1.0	0.00	3.0000	2.8000	2.6000	2.4000	2.2000	2.0000	1.8000	1.6000	1.4000	1.2000
0.9	-0.05	1.5075	1.4070	1.3065	1.2060	1.1055	1.0050	0.9045	0.8040	0.7035	0.6030
0.8	-0.10	0.6048	0.5645	0.5242	0.4839	0.4435	0.4032	0.3629	0.3226	0.2822	0.2419
0.7	-0.15	0.3027	0.2825	0.2624	0.2422	0.2220	0.2018	0.1816	0.1614	0.1413	0.1211
0.6	-0.20	0.1781	0.1663	0.1544	0.1425	0.1306	0.1188	0.1069	0.0950	0.0831	0.0712
0.5	-0.25	0.1164	0.1087	0.1009	0.0932	0.0854	0.0776	0.0699	0.0621	0.0543	0.0466
0.4	-0.30	0.0819	0.0764	0.0710	0.0655	0.0600	0.0546	0.0491	0.0437	0.0382	0.0327
0.3	-0.35	0.0606	0.0565	0.0525	0.0485	0.0444	0.0404	0.0363	0.0323	0.0283	0.0242
0.2	-0.40	0.0466	0.0435	0.0404	0.0373	0.0342	0.0311	0.0280	0.0249	0.0217	0.0186
0.1	-0.45	0.0369	0.0345	0.0320	0.0296	0.0271	0.0246	0.0222	0.0197	0.0172	0.0148
0.0	-0.50	0.0300	0.0280	0.0260	0.0240	0.0220	0.0200	0.0180	0.0160	0.0140	0.0120

TABLE XIX

COMPARISON OF MODEL I AND THE REGULAR MODEL WITH P_S = 0.7 AND n = 100

Regular Estimate		MSE (Model I)/MSE (Regular) m_2/m_1									
P_S	Bias	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
1.0	0.00	2.4286	2.2857	2.1428	2.0000	1.8571	1.7143	1.5714	1.4286	1.2857	1.1428
0.9	-0.07	0.7053	0.6556	0.6223	0.5808	0.5393	0.4978	0.4564	0.4149	0.3734	0.3319
0.8	-0.14	0.2311	0.2175	0.2039	0.1903	0.1767	0.1632	0.1496	0.1360	0.1224	0.1088
0.7	-0.21	0.1094	0.1030	0.0966	0.0901	0.0837	0.0772	0.0708	0.0644	0.0579	0.0515
0.6	-0.28	0.0631	0.0594	0.0557	0.0515	0.0482	0.0445	0.0408	0.0371	0.0334	0.0297
0.5	-0.35	0.0409	0.0385	0.0361	0.0337	0.0312	0.0288	0.0264	0.0240	0.0216	0.0192
0.4	-0.42	0.0286	0.0269	0.0252	0.0235	0.0218	0.0202	0.0185	0.0168	0.0151	0.0134
0.3	-0.49	0.0211	0.0198	0.0186	0.0174	0.0161	0.0149	0.0136	0.0124	0.0112	0.0099
0.2	-0.56	0.0162	0.0152	0.0143	0.0133	0.0124	0.0114	0.0105	0.0095	0.0086	0.0076
0.1	-0.63	0.0128	0.0121	0.0113	0.0106	0.0098	0.0090	0.0083	0.0075	0.0068	0.0060
0.0	-0.70	0.0104	0.0098	0.0092	0.0086	0.0079	0.0073	0.0067	0.0061	0.0055	0.0049

TABLE XX

COMPARISON OF MODEL I AND THE REGULAR MODEL WITH P_S = 0.9 AND n = 100

Regular Estimate		MSE (Model I)/MSE (Regular) m_2/m_1									
P_S	Bias	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
1.0	0.00	2.1111	2.0000	1.8889	1.7778	1.6667	1.5555	1.4444	1.3333	1.2222	1.1111
0.9	-0.09	0.1971	0.1867	0.1764	0.1660	0.1556	0.1452	0.1349	0.1245	0.1141	0.1037
0.8	-0.18	0.0552	0.0523	0.0494	0.0465	0.0436	0.0407	0.0378	0.0349	0.0320	0.0290
0.7	-0.27	0.0252	0.0239	0.0226	0.0213	0.0199	0.0186	0.0173	0.0159	0.0146	0.0133
0.6	-0.36	0.0144	0.0136	0.0129	0.0121	0.0113	0.0105	0.0098	0.0091	0.0083	0.0076
0.5	-0.45	0.0093	0.0088	0.0083	0.0078	0.0073	0.0068	0.0063	0.0058	0.0054	0.0049
0.4	-0.54	0.0065	0.0061	0.0058	0.0054	0.0051	0.0048	0.0044	0.0041	0.0037	0.0034
0.3	-0.63	0.0048	0.0045	0.0043	0.0040	0.0038	0.0035	0.0032	0.0030	0.0027	0.0025
0.2	-0.72	0.0036	0.0035	0.0033	0.0031	0.0029	0.0027	0.0025	0.0023	0.0021	0.0019
0.1	-0.81	0.0029	0.0027	0.0026	0.0024	0.0023	0.0021	0.0020	0.0018	0.0017	0.0015
0.0	-0.90	0.0023	0.0022	0.0021	0.0020	0.0018	0.0017	0.0016	0.0015	0.0013	0.0012

TABLE XXI

COMPARISON OF MODEL I AND THE REGULAR MODEL WITH P_S = 0.1 AND n = 1,000

Regular Estimate		MSE (Model I)/MSE (Regular) m_2/m_1									
P_S	Bias	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
1.0	0.00	11.0000	10.0000	9.0000	8.0000	7.0000	6.0000	5.0000	4.0000	3.0000	2.0000
0.9	-0.01	5.4425	4.9478	4.4530	3.9582	3.4634	2.9687	2.4739	1.9791	1.4843	0.9895
0.8	-0.02	2.0904	1.9003	1.7103	1.5203	1.3302	1.1402	0.9502	0.7601	0.5701	0.3801
0.7	-0.03	1.0258	0.9325	0.8393	0.7460	0.6528	0.5595	0.4663	0.3730	0.2798	0.1865
0.6	-0.04	0.5977	0.5433	0.4890	0.4347	0.3803	0.3260	0.2717	0.2173	0.1630	0.1087
0.5	-0.05	0.3886	0.3533	0.3179	0.2826	0.2473	0.2120	0.1766	0.1413	0.1060	0.0706
0.4	-0.06	0.2721	0.2474	0.2226	0.1979	0.1731	0.1484	0.1237	0.0989	0.0742	0.0495
0.3	-0.07	0.2008	0.1826	0.1643	0.1461	0.1278	0.1095	0.0913	0.0730	0.0548	0.0365
0.2	-0.08	0.1542	0.1402	0.1262	0.1121	0.0981	0.0841	0.0701	0.0561	0.0420	0.0280
0.1	-0.09	0.1221	0.1110	0.0999	0.0888	0.0777	0.0666	0.0555	0.0444	0.0333	0.0222
0.0	-0.10	0.0990	0.0900	0.0810	0.0720	0.0630	0.0540	0.0450	0.0360	0.0270	0.0180

TABLE XXII

COMPARISON OF MODEL I AND THE WARNER MODEL

P_S	p	Var (Model I)/Var (Warner) m_2/m_1 1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
0.1	0.1 or 0.9	4.2928	3.9025	3.5123	3.1220	2.7318	2.3415	1.9513	1.5610	1.1707	0.7805
	0.2 or 0.8	1.8524	1.6840	1.5156	1.3472	1.1788	1.0104	0.8420	0.6736	0.5052	0.3368
	0.3 or 0.7	0.7059	0.6417	0.5775	0.5134	0.4492	0.3850	0.3208	0.2567	0.1925	0.1283
	0.4 or 0.6	0.1626	0.1478	0.1330	0.1182	0.1034	0.0887	0.0739	0.0591	0.0443	0.0295
0.2	0.1 or 0.9	3.1934	2.9273	2.6612	2.3950	2.1289	1.8628	1.5967	1.3306	1.0645	0.7983
	0.2 or 0.8	1.5882	1.4559	1.3235	1.1912	1.0588	0.9265	0.7941	0.6618	0.5294	0.3971
	0.3 or 0.7	0.6519	0.5976	0.5433	0.4890	0.4346	0.3803	0.3260	0.2716	0.2173	0.1630
	0.4 or 0.6	0.1558	0.1428	0.1299	0.1169	0.1039	0.0909	0.0779	0.0649	0.0519	0.0390
0.3	0.1 or 0.9	2.5954	2.3957	2.1961	1.9965	1.7968	1.5972	1.3975	1.1979	0.9982	0.7986
	0.2 or 0.8	1.3905	1.2835	1.1766	1.0696	0.9626	0.8557	0.7487	0.6418	0.5348	0.4278
	0.3 or 0.7	0.5977	0.5517	0.5057	0.4598	0.4138	0.3678	0.3218	0.2759	0.2299	0.1839
	0.4 or 0.6	0.1465	0.1353	0.1240	0.1127	0.1014	0.0902	0.0789	0.0676	0.0564	0.0451

TABLE XXII (Continued)

P_S	p	Var (Model I)/Var (Warner) m_2/m_1									
		1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
0.4	0.1 or 0.9	2.2069	2.0493	1.8916	1.7340	1.5764	1.4187	1.2611	1.1035	0.9458	0.7882
	0.2 or 0.8	1.2273	1.1396	1.0519	0.9643	0.8766	0.7890	0.7013	0.6136	0.5260	0.4383
	0.3 or 0.7	0.5411	0.5024	0.4638	0.4251	0.3865	0.3478	0.3092	0.2705	0.2319	0.1932
	0.4 or 0.6	0.1346	0.1250	0.1154	0.1058	0.0961	0.0865	0.0765	0.0673	0.0577	0.0481
0.5	0.1 or 0.9	1.9200	1.7920	1.6640	1.5360	1.4080	1.2800	1.1520	1.0240	0.8960	0.7680
	0.2 or 0.8	1.0800	1.0080	0.9360	0.8640	0.7920	0.7200	0.6480	0.5760	0.5040	0.4320
	0.3 or 0.7	0.4800	0.4480	0.4160	0.3840	0.3520	0.3200	0.2880	0.2560	0.2240	0.1920
	0.4 or 0.6	0.1200	0.1120	0.1040	0.0960	0.0880	0.0800	0.0720	0.0640	0.0560	0.0480
0.6	0.1 or 0.9	1.6815	1.5764	1.4713	1.3662	1.2611	1.1560	1.0509	0.9458	0.8407	0.7356
	0.2 or 0.8	0.9351	0.8766	0.8182	0.7597	0.7013	0.6429	0.5844	0.5260	0.4675	0.4091
	0.3 or 0.7	0.4122	0.3865	0.3607	0.3349	0.3092	0.2834	0.2576	0.2319	0.2061	0.1803
	0.4 or 0.6	0.1026	0.0961	0.0897	0.0833	0.0769	0.0705	0.0641	0.0577	0.0513	0.0449

TABLE XXII (Continued)

P_S	p	Var (Model I)/Var (Warner) m_2/m_1									
		1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
0.7	0.1 or 0.9	1.4546	1.3690	1.2834	1.1979	1.1123	1.0267	0.9412	0.8556	0.7701	0.6845
	0.2 or 0.8	0.7792	0.7334	0.6876	0.6418	0.5959	0.5501	0.5042	0.4584	0.4126	0.3667
	0.3 or 0.7	0.3350	0.3153	0.2956	0.2759	0.2561	0.2364	0.2167	0.1970	0.1773	0.1576
	0.4 or 0.6	0.0821	0.0773	0.0725	0.0676	0.0628	0.0580	0.0531	0.0483	0.0435	0.0386
0.8	0.1 or 0.9	1.1975	1.1310	1.0645	0.9979	0.9314	0.8649	0.7983	0.7318	0.6653	0.5988
	0.2 or 0.8	0.5956	0.5625	0.5294	0.4963	0.4632	0.4301	0.3971	0.3640	0.3309	0.2978
	0.3 or 0.7	0.2445	0.2309	0.2173	0.2037	0.1901	0.1766	0.1630	0.1494	0.1358	0.1222
	0.4 or 0.6	0.0584	0.0552	0.0519	0.0487	0.0454	0.0422	0.0390	0.0357	0.0325	0.0292
0.9	0.1 or 0.9	0.8239	0.7805	0.7371	0.6938	0.6504	0.6070	0.5637	0.5203	0.4770	0.4336
	0.2 or 0.8	0.3555	0.3368	0.3181	0.2994	0.2807	0.2619	0.2432	0.2245	0.2058	0.1871
	0.3 or 0.7	0.1355	0.1283	0.1212	0.1141	0.1069	0.0998	0.0927	0.0856	0.0784	0.0713
	0.4 or 0.6	0.0312	0.0295	0.0279	0.0263	0.0246	0.0230	0.0213	0.0197	0.0181	0.0164

TABLE XXIII

COMPARISON OF MODEL I AND THE UNRELATED QUESTIONS MODEL WITH $P_S = 0.1$

P_N	P_1	Var (Model I)/Var (Unrelated) m_2/m_1									
		1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
	0.1	0.1100	0.1000	0.0900	0.0800	0.0700	0.0600	0.0500	0.0400	0.0300	0.0200
	0.3	0.9900	0.9000	0.8100	0.7200	0.6300	0.5400	0.4500	0.3600	0.2700	0.1800
0.1	0.5	2.7500	2.5000	2.2500	2.0000	1.7500	1.5000	1.2500	1.0000	0.7500	0.5000
	0.7	5.3900	4.9000	4.4100	3.9200	3.4300	2.9400	2.4500	1.9600	1.4700	0.9800
	0.9	8.9100	8.1000	7.2900	6.4800	5.6700	4.8600	4.0500	3.2400	2.4300	1.6200
	0.1	0.0398	0.0362	0.0326	0.0290	0.0254	0.0217	0.0181	0.0145	0.0109	0.0072
	0.3	0.3782	0.3438	0.3094	0.2750	0.2407	0.2063	0.1719	0.1375	0.1031	0.0688
0.5	0.5	1.1786	1.0714	0.9643	0.8571	0.7500	0.6428	0.5357	0.4286	0.3214	0.2143
	0.7	2.8269	2.5699	2.3129	2.0559	1.7989	1.5419	1.2850	1.0280	0.7710	0.5140
	0.9	6.6603	6.0548	5.4493	4.8438	4.2384	3.6329	3.0274	2.4219	1.8164	1.2110
	0.1	0.0671	0.0610	0.0549	0.0488	0.0427	0.0366	0.0305	0.0244	0.0183	0.0122
	0.3	0.3970	0.3610	0.3249	0.2888	0.2527	0.2166	0.1805	0.1444	0.1083	0.0721
0.9	0.5	0.9900	0.9000	0.8100	0.7200	0.6300	0.5400	0.4500	0.3600	0.2700	0.1800
	0.7	2.1618	1.9652	1.7687	1.5722	1.3757	1.1791	0.9826	0.7861	0.5896	0.3930
	0.9	5.4329	4.9390	4.4451	3.9512	3.4573	2.9634	2.4695	1.9756	1.4817	0.9878

TABLE XXIV

COMPARISON OF MODEL I AND THE UNRELATED QUESTIONS MODEL WITH P_S = 0.3

P_N	p_1	Var (Model I)/Var (Unrelated) m_2/m_1									
		1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
0.1	0.1	0.0862	0.0795	0.0729	0.0663	0.0596	0.0530	0.0464	0.0398	0.0331	0.0265
	0.3	0.6094	0.5625	0.5156	0.4687	0.4219	0.3750	0.3281	0.2812	0.2344	0.1875
	0.5	1.4219	1.3125	1.2031	1.0937	0.9844	0.8750	0.7656	0.6562	0.5469	0.4375
	0.7	2.4446	2.2566	2.0685	1.8805	1.6924	1.5044	1.3163	1.1283	0.9402	0.7522
	0.9	3.6562	3.3750	3.0937	2.8125	2.5312	2.2500	1.9687	1.6875	1.4062	1.1250
0.5	0.1	0.0364	0.0336	0.0308	0.0280	0.0252	0.0224	0.0196	0.0168	0.0140	0.0112
	0.3	0.3324	0.3068	0.2812	0.2557	0.2301	0.2045	0.1790	0.1534	0.1278	0.1023
	0.5	0.9479	0.8750	0.8021	0.7292	0.6562	0.5853	0.5104	0.4375	0.3646	0.2917
	0.7	1.9353	1.7864	1.6376	1.4887	1.3398	1.1910	1.0421	0.8932	0.7443	0.5955
	0.9	3.3874	3.1268	2.8663	2.6057	2.3451	2.0845	1.8240	1.5634	1.3028	1.0423
0.9	0.1	0.0677	0.0625	0.0573	0.0521	0.0469	0.0417	0.0364	0.0312	0.0260	0.0208
	0.3	0.4062	0.3750	0.3437	0.3125	0.2812	0.2500	0.2187	0.1875	0.1562	0.1250
	0.5	0.9479	0.8750	0.8021	0.7292	0.6562	0.5833	0.5104	0.4375	0.3646	0.2917
	0.7	1.7864	1.6490	1.5116	1.3742	1.2368	1.0993	0.9619	0.8245	0.6871	0.5497
	0.9	3.1992	2.9536	2.7074	2.4613	2.2148	1.9690	1.7226	1.4766	1.2305	0.9845

TABLE XXV

COMPARISON OF MODEL I AND THE UNRELATED QUESTIONS MODEL WITH P_S = 0.5

P_N	p_1	Var (Model I)/Var (Unrelated) m_2/m_1									
		1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
0.1	0.1	0.0623	0.0581	0.0540	0.0498	0.0457	0.0415	0.0374	0.0332	0.0291	0.0249
	0.3	0.3933	0.3671	0.3409	0.3147	0.2885	0.2622	0.2360	0.2098	0.1836	0.1573
	0.5	0.8928	0.8333	0.7738	0.7143	0.6548	0.5952	0.5357	0.4762	0.4167	0.3571
	0.7	1.5598	1.4558	1.3519	1.2479	1.1439	1.0399	0.9359	0.8319	0.7279	0.6239
	0.9	2.4456	2.2826	2.1196	1.9565	1.7935	1.6304	1.4674	1.3043	1.1413	0.9783
0.5	0.1	0.0300	0.0280	0.0260	0.0240	0.0220	0.0200	0.0180	0.0160	0.0140	0.0120
	0.3	0.2700	0.2520	0.2340	0.2160	0.1980	0.1800	0.1620	0.1440	0.1260	0.1080
	0.5	0.7500	0.7000	0.6500	0.6000	0.5500	0.5000	0.4500	0.4000	0.3500	0.3000
	0.7	1.4700	1.3720	1.2740	1.1760	1.0780	0.9800	0.8820	0.7840	0.6860	0.5880
	0.9	2.4300	2.2680	2.1060	1.9440	1.7820	1.6200	1.4580	1.2960	1.1340	0.9720
0.9	0.1	0.0623	0.0581	0.0540	0.0498	0.0457	0.0415	0.0374	0.0332	0.0291	0.0249
	0.3	0.3933	0.3671	0.3409	0.3147	0.2885	0.2622	0.2360	0.2098	0.1836	0.1573
	0.5	0.8928	0.8333	0.7738	0.7143	0.6548	0.5952	0.5357	0.4762	0.4167	0.3571
	0.7	1.5598	1.4558	1.3519	1.2479	1.1439	1.0399	0.9359	0.8319	0.7279	0.6239
	0.9	2.4456	2.2826	2.1196	1.9565	1.7935	1.6304	1.4674	1.3043	1.1413	0.9783

TABLE XXVI

COMPARISON OF MODEL I AND THE UNRELATED QUESTIONS MODEL WITH $P_S = 0.7$

P_N	p_1	Var (Model I)/Var (Unrelated) m_2/m_1 1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
0.1	0.1	0.0379	0.0357	0.0335	0.0312	0.0290	0.0268	0.0245	0.0223	0.0201	0.0178
	0.3	0.2277	0.2143	0.2009	0.1875	0.1741	0.1607	0.1473	0.1339	0.1205	0.1071
	0.5	0.5312	0.5000	0.4687	0.4375	0.4062	0.3750	0.3437	0.3125	0.2812	0.2500
	0.7	1.0012	0.9423	0.8834	0.8245	0.7656	0.7067	0.6478	0.5889	0.5300	0.4711
	0.9	1.7930	1.6875	1.5820	1.4766	1.3711	1.2656	1.1601	1.0547	0.9492	0.8437
0.5	0.1	0.0204	0.0192	0.0180	0.0168	0.0156	0.0144	0.0132	0.0120	0.0108	0.0096
	0.3	0.1863	0.1753	0.1644	0.1534	0.1424	0.1315	0.1205	0.1096	0.0986	0.0877
	0.5	0.5312	0.5000	0.4687	0.4375	0.4062	0.3750	0.3437	0.3125	0.2812	0.2500
	0.7	1.0846	1.0208	0.9570	0.8932	0.8294	0.7656	0.7018	0.6380	0.5742	0.5104
	0.9	1.8984	1.7868	1.6751	1.5634	1.4517	1.3401	1.2284	1.1167	1.0050	0.8934
0.9	0.1	0.0483	0.0454	0.0426	0.0398	0.0369	0.0341	0.0312	0.0284	0.0256	0.0227
	0.3	0.3415	0.3214	0.3013	0.2812	0.2612	0.2411	0.2210	0.2009	0.1808	0.1607
	0.5	0.7969	0.7500	0.7031	0.6562	0.6094	0.5625	0.5156	0.4687	0.4219	0.3750
	0.7	1.3701	1.2895	1.2089	1.1283	1.0477	0.9671	0.8865	0.8059	0.7253	0.6447
	0.9	2.0491	1.9286	1.8080	1.6875	1.5670	1.4464	1.3259	1.2053	1.0848	0.9643

TABLE XXVII

COMPARISON OF MODEL I AND THE UNRELATED QUESTIONS MODEL WITH $P_S = 0.9$

P_N	p_1	Var (Model I)/Var (Unrelated) m_2/m_1									
		1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
	0.1	0.0129	0.0122	0.0115	0.0108	0.0102	0.0095	0.0088	0.0081	0.0074	0.0068
	0.3	0.0762	0.0722	0.0682	0.0642	0.0602	0.0561	0.0521	0.0481	0.0441	0.0401
0.1	0.5	0.1900	0.1800	0.1700	0.1600	0.1500	0.1400	0.1300	0.1200	0.1100	0.1000
	0.7	0.4149	0.3930	0.3712	0.3494	0.3275	0.3057	0.2839	0.2620	0.2402	0.2184
	0.9	1.0427	0.9878	0.9329	0.8780	0.8232	0.7683	0.7134	0.6585	0.6036	0.5488
	0.1	0.0076	0.0072	0.0068	0.0064	0.0060	0.0056	0.0052	0.0048	0.0044	0.0040
	0.3	0.0726	0.0688	0.0649	0.0611	0.0573	0.0535	0.0497	0.0458	0.0420	0.0382
0.5	0.5	0.2262	0.2143	0.2042	0.1905	0.1786	0.1667	0.1548	0.1428	0.1309	0.1190
	0.7	0.5425	0.5140	0.4854	0.4569	0.4283	0.3998	0.3712	0.3426	0.3141	0.2855
	0.9	1.2782	1.2110	1.1437	1.0764	1.0091	0.9419	0.8746	0.8073	0.7400	0.6727
	0.1	0.0211	0.0200	0.0189	0.0178	0.0167	0.0155	0.0144	0.0133	0.0122	0.0111
	0.3	0.1900	0.1800	0.1700	0.1600	0.1500	0.1400	0.1300	0.1200	0.1100	0.1000
0.9	0.5	0.5278	0.5000	0.4722	0.4444	0.4167	0.3889	0.3611	0.3333	0.3055	0.2778
	0.7	1.0344	0.9800	0.9255	0.8711	0.8167	0.7622	0.7078	0.6533	0.5989	0.5444
	0.9	1.7100	1.6200	1.5300	1.4400	1.3500	1.2600	1.1700	1.0800	0.9900	0.9000

TABLE XXVIII

COMPARISON OF MODEL I AND THE MULTIPLE TRIALS MODEL WITH T = 100 AND P_S = 0.1

p	k	Var (Model I)/Var (Multiple Trials) m_2/m_1									
		1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
0.1	5	8.3813	7.6194	6.8574	6.0955	5.3335	4.5716	3.8097	3.0477	2.2858	1.5239
	10	9.5137	8.6488	7.7840	6.9191	6.0542	5.1893	4.3244	3.4595	2.5946	1.7298
0.2	5	5.5344	5.0313	4.5282	4.0250	3.5219	3.0188	2.5156	2.0125	1.5094	1.0063
	10	7.3639	6.6944	6.0250	5.3555	4.6861	4.0167	3.3472	2.6778	2.0083	1.3389
0.3	5	2.8085	2.5532	2.2979	2.0425	1.7872	1.5319	1.2766	1.0213	0.7659	0.5106
	10	4.4746	4.0678	3.6610	3.2542	2.8474	2.4407	2.0339	1.6271	1.2203	0.8135
0.4	5	0.7674	0.6977	0.6279	0.5581	0.4884	0.4186	0.3488	0.2791	0.2093	0.1395
	10	1.4348	1.3043	1.1739	1.0435	0.9130	0.7826	0.6522	0.5217	0.3913	0.2609

TABLE XXIX

COMPARISON OF MODEL I AND THE MULTIPLE TRIALS MODEL WITH T = 100 AND P_S = 0.3

p	k	Var (Model I)/Var (Multiple Trials) m_2/m_1 1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
0.1	5	3.8216	3.5276	3.2337	2.9397	2.6457	2.3517	2.0578	1.7638	1.4698	1.1759
	10	4.0614	3.7490	3.4366	3.1242	2.8117	2.4993	2.1869	1.8745	1.5621	1.2497
0.2	5	3.0447	2.8105	2.5763	2.3421	2.1079	1.8737	1.6394	1.4052	1.1710	0.9368
	10	3.5765	3.3014	3.0262	2.7511	2.4760	2.2009	1.9258	1.6507	1.3756	1.1004
0.3	5	1.9259	1.7778	1.6296	1.4815	1.3333	1.1852	1.0370	0.8889	0.7407	0.5926
	10	2.6667	2.4615	2.2564	2.0513	1.8461	1.6410	1.4359	1.2308	1.0256	0.8205
0.4	5	0.6454	0.5957	0.5461	0.4964	0.4468	0.3972	0.3475	0.2979	0.2482	0.1986
	10	1.1234	1.0370	0.9506	0.8642	0.7778	0.6913	0.6049	0.5185	0.4321	0.3457

TABLE XXX

COMPARISON OF MODEL I AND THE MULTIPLE TRIALS MODEL WITH T = 100 AND P_S = 0.5

p	k	Var (Model I)/Var (Multiple Trials) m_2/m_1 1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
0.1	5	2.6967	2.5169	2.3371	2.1573	1.9776	1.7978	1.6180	1.4382	1.2584	1.0788
	10	2.8403	2.6509	2.4616	2.2722	2.0828	1.8935	1.7041	1.5148	1.3254	1.1361
0.2	5	2.2132	2.0656	1.9181	1.7705	1.6230	1.4754	1.3279	1.1803	1.0328	0.8853
	10	2.5472	2.3774	2.2076	2.0378	1.8679	1.6981	1.5283	1.3585	1.1887	1.0189
0.3	5	1.4634	1.3658	1.2683	1.1707	1.0732	0.9756	0.8780	0.7805	0.6829	0.5854
	10	1.9672	1.8361	1.7049	1.5738	1.4426	1.3115	1.1803	1.0492	0.9180	0.7869
0.4	5	0.5172	0.4827	0.4483	0.4138	0.3793	0.3448	0.3103	0.2759	0.2414	0.2069
	10	0.8823	0.8235	0.7647	0.7059	0.6470	0.5882	0.5294	0.4706	0.4118	0.3529

TABLE XXXI

COMPARISON OF MODEL I AND THE MULTIPLE TRIALS MODEL WITH T = 100 AND P_S = 0.7

p	k	Var (Model I)/Var (Multiple Trials) m_2/m_1									
		1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
0.1	5	2.1418	2.0158	1.8898	1.7638	1.6378	1.5118	1.3858	1.2599	1.1339	1.0079
	10	2.2762	2.1423	2.0084	1.8745	1.7406	1.6067	1.4728	1.3389	1.2050	1.0711
0.2	5	1.7064	1.6060	1.5056	1.4052	1.3049	1.2045	1.1041	1.0037	0.9034	0.8030
	10	2.0044	1.8865	1.7686	1.6507	1.5328	1.4149	1.2970	1.1790	1.0611	0.9432
0.3	5	1.0794	1.0159	0.9524	0.8889	0.8254	0.7619	0.6984	0.6349	0.5714	0.5079
	10	1.4945	1.4066	1.3187	1.2308	1.1428	1.0549	0.9670	0.8791	0.7912	0.7033
0.4	5	0.3617	0.3404	0.3191	0.2979	0.2766	0.2553	0.2340	0.2128	0.1915	0.1702
	10	0.6296	0.5926	0.5555	0.5185	0.4815	0.4444	0.4074	0.3704	0.3333	0.2963

TABLE XXXII

COMPARISON OF MODEL I AND THE MULTIPLE TRIALS MODEL WITH T = 100 AND P_S = 0.9

p	k	Var (Model I)/Var (Multiple Trials) m_2/m_1									
		1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
0.1	5	1.6085	1.5239	1.4392	1.3545	1.2699	1.1852	1.0006	1.0159	0.9312	0.8466
	10	1.8259	1.7298	1.6337	1.5376	1.4415	1.3454	1.2493	1.1532	1.0571	0.9610
0.2	5	1.0622	1.0063	0.9503	0.8944	0.8385	0.7826	0.7267	0.6708	0.6149	0.5590
	10	1.4133	1.3389	1.2645	1.1901	1.1157	1.0413	0.9670	0.8926	0.8182	0.7438
0.3	5	0.5390	0.5106	0.4823	0.4539	0.4255	0.3972	0.3688	0.3404	0.3120	0.2837
	10	0.8587	0.8135	0.7684	0.7232	0.6780	0.6328	0.5876	0.5424	0.4972	0.4520
0.4	5	0.1473	0.1395	0.1318	0.1240	0.1163	0.1085	0.1008	0.0930	0.0853	0.0775
	10	0.2754	0.2609	0.2464	0.2319	0.2174	0.2029	0.1884	0.1739	0.1594	0.1449

VITA 2

Sorachai Bhisalbutra

Candidate for the Degree of

Doctor of Philosophy

Thesis: RANDOMIZED RESPONSE DESIGNS

Major Field: Statistics

Biographical:

Personal Data: Born in Bangkok, Thailand, February 24, 1941, the son of Suwat and Madee Bhisalbutra.

Education: Attended elementary school in Bangkok, Thailand; graduated from Pre-University High School, Bangkok, Thailand, in 1959; received the Bachelor of Science degree in Mathematics from Chulalongkorn University, Bangkok, Thailand, in 1964; received the Master of Science degree in Statistics from Iowa State University, Ames, Iowa, in 1968; completed the requirements for the Doctor of Philosophy degree at Oklahoma State University in December, 1976.

Professional Experience: Second grade Statistician at the National Statistical Office, Bangkok, Thailand, 1964-1973; Instructor of Statistics at Chulalongkorn University, Bangkok, Thailand, 1973-1974.