BLOCK TOTAL RESPONSE AS AN ALTERNATIVE TO THE RANDOMIZED RESPONSE METHOD IN SURVEYS

D. Raghavarao[1] and W. T. Federer
Temple University and Cornell University

BU-490-M (revised) February 1978

SUMMARY

The randomized response technique appears to have been an innovative and useful procedure for eliciting reliable responses from individuals on sensitive or embarrassing questions. In this paper a new and alternative method is proposed for the same problem. Through the use of supplemented block, (v,k,r,b,λ) balanced incomplete block, and spring balance weighing designs, the individual is required to give a total of the responses to k questions, sensitive or not. From these block totals it is possible to obtain estimated responses for each of the v questions used in the survey, yet not obtain individual response to single questions. Anonymity of response for a single interviewee is thus maintained. Estimators and their variances for the estimated responses are obtained. The method allows the surveyor to obtain answers to several sensitive questions without being unduly time-consuming.

Keywords: Balanced incomplete block design, supplemented block design, spring balance weighing design, anonymous response, sensitive question.

---

[1] Work on the original version of this paper was done while this author was at the University of Guelph.

Paper No. BU-490-M in the Biometrics Unit Mimeo Series, Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York - 14853.

1. INTRODUCTION

The problem of eliciting information on sensitive, embarrassing, incriminating, and/or delicate questions from respondents is of considerable concern and interest to survey statisticians. In attempting to obtain the information, the form of the questioning may be direct or indirect and the identity of the respondee's answer may be known, unknown, known but the respondee is convinced it is unknown, or unknown but the respondee is unconvinced about the anonymity of his answer. The amount of bias in the last case depends upon the respondees' degree of belief concerning anonymity of their responses and/or upon distracting their attention in such a manner as to obtain correct responses. One early anonymous-direct-question method that was used successfully (e.g., by A. J. King and others at Iowa State University) was to have the respondee complete an unmarked questionnaire in secret and to deposit the questionnaire in a large locked box in which other questionnaires had been deposited; then, the respondee observed that the contents of the box were thoroughly mixed. We shall call this method the "black box" (BB) method.

A second method to obtain correct answers to sensitive questions is known as the randomized response (RR) method and was presented by Warner (1965). Since its introduction there have been several extensions to the theory and use of the RR procedure (see e.g., Folsom et al. (1973), Greenberg et al. (1969), Warner (1971)). Although the RR procedure is regarded as being of considerable value, it must be realized that it is not useful and optimal for all situations. For example, if one were able to convince the respondent of his anonymity equally for both the BB and RR procedures, then the variance of the estimated parameter by the BB procedure would be smaller than for the RR method. Also, the use of the latter may be questionable in surveys with several sensitive questions because it is too time-consuming, costly, and complicated. Also, note that criteria other than variance-optimality may be justifiable in assessing the worth of a procedure. Time, cost, accuracy, and acceptance should be considered.

Because an all-perfect method for all situations has not been devised, it is necessary to consider alternatives. One such alternative is considered in the present paper. It is an anonymous-direct question approach when the answer to one sensitive question is sought or when answers to several of v sensitive questions are desired. Answers to one or more of the v questions may be obtained from every member in the sample or from only a subsample of members of the sample. The key idea in our approach is that scores for a set of k of the v questions, sensitive and/or nonsensitive, are added and only a total score for the k questions is reported by the respondent. Different respondents receive different sets of k questions; there are b different sets of questions made up according to known block experiment designs such as the supplemented block (SB) designs and balanced incomplete block (BIB) designs. The block of k questions is randomly assigned a respondent, with the stipulation that all blocks have an equal or nearly equal number of respondents. Then, from the block totals, the responses, we are able to obtain estimates of population proportions or means for each question; with properly constructed scores for responses, we are unable to determine what an individual's response was to a particular question.

In section two, we first discuss two special cases using a supplemented block design procedure. It is shown how to compare the proposed procedure with the RR procedure used by Greenberg et al. (1969, section 4), involving an unrelated question and a proportion or mean known in advance of sampling. Then, some general results for the SB designs are presented.

In the third section of the paper, a procedure utilizing BIB design theory is presented. This method is suitable when a number of sensitive questions are being asked in a survey. Here only a subsample respond to any particular question.

The precision of the BIB method will be less than for an RR procedure, but the saving in interviewing time for several sensitive questions can be considerable.

We shall denote our method as the block total response (BTR) procedure. When particular designs are used the symbol BTR will be prefixed by the design symbols. For example, the supplemented block total response procedure is denoted as the SBTR procedure (section 2), and the balanced incomplete block total response procedure is denoted as the BIBTR procedure (section 3). Some suggestions for constructing scores to be used in the BIBTR procedure are also presented. The purpose is to assure anonymity of an individual's response.

Since the solutions for treatment effects for BIB designs have been given when treatment totals are available, it was considered necessary to present solutions for treatment effects when only the block totals are available. This is given in the appendix.

2. SUPPLEMENTED BLOCK TOTAL RESPONSE (SBTR) METHOD

To illustrate the procedure, let us first consider two special cases before presenting some general results. Suppose that answers are sought to v questions with one of them being the sensitive question. Further, suppose that the sensitive question will be asked of every person in the sample and that each of the remaining v-1 questions will be asked of 2n individuals, given that the total sample size is vn. The remaining v-1 questions could be sensitive or nonsensitive, but it is desired to ask these questions on a $2/v^{th}$ subsample of the total sample. Suppose that the response is the total for k = 2 questions, that there are v-1 such blocks, that there is one block where k = v questions, and that there are v parameters, p_i, $i=1,2,\cdots,v$, to be estimated. Let the design be:

Block	Questions in the set	Block size = k_j	Response	Expected value of response
1	1 and 2	2	Y_1	$p_1 + p_2$
2	1 and 3	2	Y_2	$p_1 + p_3$
3	1 and 4	2	Y_3	$p_1 + p_4$
4	1 and 5	2	Y_4	$p_1 + p_5$
⋮	⋮	⋮	⋮	⋮
v-1	1 and v	2	Y_{v-1}	$p_1 + p_v$
b = v	all v	v	Y_v	$\Sigma_1^v p_i$

The Y_j are means of n responses. If the Y_j have equal variances σ^2/n, the least squares solutions for the p_i are:

$$\hat{p}_1 = \left[\sum_{1}^{v-1} Y_j - Y_v \right] \Big/ (v-2) \tag{2.1}$$

$$\hat{p}_i = \left[(v-3)Y_{i-1} + Y_v - \sum_{\substack{j=1 \\ j\neq i-1}}^{v-1} Y_j \right] \; (v-2) \tag{2.2}$$

for $i=2,3,4,\cdots,v$. If the variances of the Y_j are not equal, then weighted least squares estimates may be obtained where the true weights will more than likely have to be replaced by the estimated weights $\hat{w}_j = n/\hat{\sigma}_j^2$ for $\hat{\sigma}_j^2 = \Sigma_{h=1}^n (Y_{jh}-Y_j)^2/(n-1)$. The sum of squares to be minimized would be $\Sigma_{j=1}^v \hat{w}_j (Y_j - EY_j)^2$. Note that for the development in (2.1) and (2.2), the observed mean response $Y_j = EY_j + \epsilon_j$ and the ϵ_j are assumed to be $IID(0,\sigma^2/n)$. Under these assumptions the variances of the $\hat{p}_i$ are:

$$V(\hat{p}_1) = \frac{\sigma^2}{n} \left(\frac{v}{(v-2)^2} \right) \tag{2.3}$$

$$V(\hat{p}_i,\ i=2,3,\cdots,v) = \frac{\sigma^2}{n}\left(\frac{v^2-5v+8}{(v-2)^2}\right) . \tag{2.4}$$

Given that the parameter for the sensitive question is p_1 and that the RR procedure used is the unrelated question technique with the true proportion known in advance, then the relative precision of the RR procedure to the SBTR procedure is:

$$\frac{(\sigma_1^2/n)/v\pi^2}{V(\hat{p}_1)} = \frac{\sigma_1^2/nv\pi^2}{v\sigma^2/n(v-2)^2} = \frac{\sigma_1^2}{\sigma^2}\left(\frac{(v-2)^2}{v^2\pi^2}\right) . \tag{2.5}$$

In order to compare the procedures one would need information on the ratio σ_1^2/σ^2. Note that $(v-2)^2/v^2\pi^2$ becomes greater than one for $v \geq 7$ and $\pi^2 = \frac{1}{2}$, and it approaches two as v becomes large; also, if $\pi^2 = \frac{1}{3}$ and $v \geq 3$ this ratio is greater than or equal to one, and it approaches three as v becomes large.

Consider now the following SBTR design, where the Y_j are means of n observations.

Block	Questions in the set	Block size = k_j	Response	Expected value of response
1	1,2,3	3	Y_1	$p_1 + p_2 + p_3$
2	1,2,4	3	Y_2	$p_1 + p_2 + p_4$
3	1,2,5	3	Y_3	$p_1 + p_2 + p_5$
⋮	⋮	⋮	⋮	⋮
v-2	1,2,v	3	Y_{v-1}	$p_1 + p_2 + p_v$
v-1	1,3,4	3	Y_{v-1}	$p_1 + p_3 + p_4$
⋮	⋮	⋮	⋮	⋮
(v-1)(v-2)/2 = a	1,v-1,v	3	Y_a	$p_1 + p_{v-1} + p_v$
a+1	all v	v	Y_{a+1}	$\Sigma_1^v\ p_i$
⋮	⋮	⋮	⋮	⋮
b = a+v-2	all v	v	Y_b	$\Sigma_1^v\ p_i$

Again assuming homoscedasticity and minimizing the sum of squares $\Sigma_1^b(Y_j-EY_j)^2$, the following solution is obtained for p_1:

$$\hat{p}_1 = \frac{2}{(v-2)(v-3)}\left[\sum_{1}^{a} Y_j - \sum_{a+1}^{b} Y_j\right] , \qquad (2.6)$$

with a variance of

$$V(\hat{p}_1) = 2(v+1)\sigma^2/(v-2)(v-3)^2 n \quad . \qquad (2.7)$$

Note that it is possible to obtain SB designs which provide partial sample information on v-1 additional parameters. This could be important in surveys which desire less information on v-1 items and full information on a sensitive item. There is no difficulty in obtaining large v as the sample size is usually quite large, even within strata. Also, note that it is possible to obtain SB designs yielding lower precision than the one above. For example, if one uses the preceding design with only one block containing all v treatments instead of v-2 such blocks, it has a larger variance than the SBTR design given.

We now proceed to present some general results for supplemented block designs. If N^* is the incidence matrix of any design D on v-u treatments in b^* blocks of sizes $k_1^*, k_2^*, \cdots, k_b^*$ where the treatments are replicated $r_1^*, r_2^*, \cdots, r_{v-u}^*$; then the supplemented design D^* with u supplemented treatments has incidence matrix $M = \begin{pmatrix} N^* \\ J_{u,b^*} \end{pmatrix}$. If one wishes to use the full resources of the sample in estimating the parameters corresponding to sensitive questions by the alternative procedure, the supplemented treatments of D^* should be identified with the sensitive questions and we have the following result:

Lemma 2.1. The parameters corresponding to not more than one sensitive question can be estimated by D^* (this implies $u \nmid 1$); also, the parameters corresponding to even one sensitive question cannot be estimated by D^* if $k_1^* = k_2^* = \cdots = k_b^*$.

Proof. Let the unknown parameters for the v questions be $\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_v$, where $\bar{X}_i$ may be the mean of a continuous variable or a proportion from a dichotomous population, let $\underline{w}' = (\bar{X}_1\ \bar{X}_2\ \cdots\ \bar{X}_v)$ be the parametric row vector, let $Y_1, Y_2, \cdots, Y_b$ be means of the n block totals, and let the row vector of block means be $\underline{Y}' = (Y_1\ Y_2\ \cdots\ Y_b)$. Then, $E(\underline{Y}) = M'\underline{w}$, where $E(\underline{Y})$ denotes the expected value of the random vector $\underline{Y}$. By applying the estimability condition, it is easy to verify that $\bar{X}_i$ is non-estimable for $i = v - u + 1,\ v - u + 2, \cdots, v$. When $u = 1$ and $k_1^* = k_2^* = \cdots = k_b^*$, $\bar{X}_v$ can be verified to be non-estimable. Lemma 2.1 implies that D^* with $u = 1$ should be considered from unequal block sized designs D in order to estimate the parameter corresponding to the sensitive question.

The class of unequal block sized designs extensively studied in the literature are the pairwise balanced designs of index λ introduced by Bose and Shrikhande (1960). An arrangement of $v - u$ symbols in b^* sets will be called a pairwise balanced design of index λ^* and type $(v\text{-}u;\ k_1^*, k_2^*, \cdots, k_m^*)$ where $k_i^* \neq k_j^*$ for every $i \neq j = 1, 2, \cdots, m$, if each block contains $k_1^*, k_2^*, \cdots$, or k_m^* treatments that are all distinct ($k_i^* \leq v\text{-}u$) and every pair of distinct treatments occurs in exactly λ blocks of the design. Symmetric unequal block (SUB) arrangements discussed in Chapter 11 of Raghavarao (1971) form a subclass of pairwise balanced designs.

For simplicity we use equi-replicated pairwise balanced designs or SUB arrangements with $4t\text{-}2$ treatments each replicated r^* times, to construct supplemented designs. Let the sensitive question correspond to the supplemented treatment $4t\text{-}1$. Then the variance of the estimated parameter corresponding to the delicate question using the supplemented design is the $(4t-1,4t-1)^{th}$ element of

$$\begin{bmatrix} (r^*-\lambda^*)I_{4t-2} + \lambda^* J_{4t-2,4t-2} & r^* J_{4t-2,1} \\ r^* J_{1,4t-2} & b^* \end{bmatrix}^{-1} \tag{2.8}$$

and is

$$\left\{ b^* - \frac{(4t-2)r^{*2}}{r^*+\lambda^*(4t-3)} \right\}^{-1} \sigma^2 \quad . \tag{2.9}$$

3. THE BIBTR TECHNIQUE

In this section, we discuss the general theory of the BTR technique with respect to balanced incomplete block design theory. Suppose that there are u sensitive questions and $v - u \geq 0$ nonsensitive questions to be used in a survey, and further suppose that $b = 4t-1$ for some integral value of t. Let $Q_1, Q_2, \cdots, Q_v$ denote the v questions and let $T_1, T_2, \cdots, T_b$ denote the $b = v$ blocks of a BIB design with treatments $Q_1, Q_2, \cdots, Q_v$ and with parameters

$$v = 4t-1 = b, \quad r = 2t = k, \quad \lambda = t \; . \tag{3.1}$$

We quantify the answers to all v questions in such a way that the total response for k questions in each T_j could arise from more than one possible assignment of answers to the individual questions. Such coding is not possible when all questions have dichotomous answers. So we choose our questions such that each set T_j contains at least one question which has quantitative answers. Further, we number the questions in such a way that each block T_j contains a mixture of sensitive and nonsensitive questions. Let the total sample size n be a multiple of v, say $n = bm = vm$. Now form b sets of m questionnaires each where the m respondents in the j^{th} set of questionnaires give their answer as a total of the k questions in block T_j of the design. Since the respondee gives only a total for k questions and since a single total response for the k questions in a block can arise in a variety of ways, the respondent's answers to individual questions are unknown. If the questions in the survey are such that the total response to T_j reveal the identity of individual responses, the interviewer can carry a table of values normally distributed with mean zero and variance one and suggest that the interviewee choose one number from the tables randomly and add it to his block total response and give the final value to him. Alternatively, the interviewee could add the last digit of his social security number to the block total before reporting it to the interviewer. This preserves the anonymity of answers to sensitive questions. In mail surveys, the different sets of questions can be mailed conveniently. For v large, a given respondee need only answer k questions, thereby shortening the time for completing the questionnaire. This would be useful also in censuses when partial information is desired on some questions.

Let $\bar{X}_i$ be the population mean or proportion for the i^{th} question $(i=1,2,\cdots,v)$. The h^{th} individual's response to the i^{th} question may be considered to be:

$$X_{ij} = \bar{X}_i + e_{ih}, \quad h=1,2,\cdots,m, \quad i=1,2,\cdots,v \quad , \tag{3.2}$$

where the e_{ih} are deviations of the individual response from the population mean and where the e_{ih} have mean zero and variance σ_i^2.

If Y_{jh} is the h^{th} response receiving the j^{th} set of questionnaires T_j, then

$$Y_{jh} = \sum_{\alpha} X_{\alpha h} = \sum_{i=1}^{v} n_{ij} X_{ih} \quad , \tag{3.3}$$

where the summation α is over the k questions appearing in set T_j of the design. Let $Y_{j.} = \Sigma_{h=1}^{m} Y_{jh}$ and $y_{j.} = Y_{j.}/m$. Irrespective of whether X_{ih} and $X_{i'h}$ $(i \neq i')$ are correlated or not, the Y_{jh} are uncorrelated. Now let

$$\hat{\sigma}_j^{*2} = \sum_{h=1}^{m} (Y_{jh}-y_{j.})^2/(m-1) \qquad j=1,2,\cdots,v \; . \tag{3.4}$$

Let $\underline{y}' = (y_{1.}\ y_{2.}\ \cdots\ y_{v.})$ and $\hat{\underline{\omega}}' = (\bar{X}_1\ \bar{X}_2\ \cdots\ \bar{X}_v)$. Then

$$E(\underline{y}) = N'\underline{\omega}\ ; \quad V(\underline{y}) = m^{-1}\,\mathrm{diag}(\sigma_1^{*2}\ \sigma_2^{*2}\ \cdots\ \sigma_v^{*2}) \quad , \tag{3.5}$$

where N is the incidence matrix of the selected BIB design used in forming the sets of questions T_j, and $\mathrm{diag}(\sigma_1^{*2}\ \sigma_2^{*2}\ \cdots\ \sigma_v^{*2})$ denotes a diagonal matrix with entries σ_j^{*2} on the diagonal. (The setup described is analogous to the weighing design situation described in Chapter 17 of Raghavarao (1971).) The least squares estimator $\hat{\underline{\omega}}$ is given by

$$\hat{\underline{\omega}} = (N^{-1})'\underline{y} \tag{3.6}$$

with

$$V(\hat{\underline{\omega}}) = (N^{-1})\{\mathrm{diag}(\hat{\sigma}_1^{*2},\hat{\sigma}_2^{*2},\cdots,\hat{\sigma}_v^{*2})\}N^{-1} \; . \tag{3.7}$$

Suppose that the responses from SB and BIB designs have the same variance σ^2. Then, for a sensitive question using the SBTR procedure, the variance is given by equation (2.9), while the variance of the estimated parameter from a BIB design is

$$\mathrm{Var}_2 = (4t-1)\sigma^2/4t^2 \; . \tag{3.8}$$

We can define the efficiency of a supplemented design over a BIB design as

$$\begin{aligned} \text{Efficiency} &= \mathrm{Var}_2/\mathrm{Var}_1 \\ &= \left\{ b^* - \frac{(4t-2)r^{*2}}{r^*+\lambda^*(4t-3)} \right\} \left\{ \frac{4t-1}{4t^2} \right\} . \end{aligned} \tag{3.9}$$

It was observed that this efficiency is not necessarily greater than 1 for all supplemented designs. In fact, with t = 8, the supplemented design arising from an SUB arrangement listed as series 28 (Raghavarao (1971), p. 220) has efficiency 2.5. Thus one is not always certain of obtaining higher efficiency with supplemented designs.

4. CONCLUDING REMARKS

An experiment comparing the RR, the BIBTR, and a randomized form of the BIBTR techniques has been reported by Smith et al. (1974). Estimated variances for the three procedures were obtained. These authors used the same sample size rather than the same cost to compare the procedures. Also, when population parameters are known for some of the nonsensitive questions included in the set of v questions, the variances of the remaining questions are reduced. In the experiment it was found that a larger bias for the two most sensitive questions appeared to be present in the RR and BIBTR procedures than for the randomized form of the BIBTR procedure.

REFERENCES

BOSE, R. C. and SHRIKHANDE, S. S. (1960). On the composition of balanced incomplete block designs. Canadian J. Math., 12, 177-188.

FOLSOM, R. E., GREENBERG, B. G., HORVITZ, D. G., and ABERNATHY, J. R. (1973). The two alternative questions randomized response model for human surveys. J. Amer. Statist. Assoc., 68, 525-530.

GREENBERG, B. G., ABERNATHY, J. R., and HORVITZ, D. G. (1969). Application of the randomized response technique in obtaining quantitative data. Proc. Soc. Statist. Sect., Amer. Statist. Assoc., 40-43.

RAGHAVARAO, D. (1971). Constructions and Combinatorial Problems in Design of Experiments. John Wiley and Sons, Inc., New York, London, Sydney, Toronto.

SMITH, L. L., FEDERER, W. T., and RAGHAVARAO, D. (1974). A comparison of three techniques for eliciting truthful answers to sensitive questions. Proc. Soc. Statist. Sect., Amer. Statist. Assoc., 447-452.

WARNER, S. L. (1965). Randomized response. A survey technique for eliminating evasive answer bias. J. Amer. Statist. Assoc., 60, 63-69.

WARNER, S. L. (1971). The linear randomized response method. J. Amer. Statist. Assoc., 66, 884-888.

APPENDIX

Some results for BIB designs when only block totals are available

In balanced incomplete block (BIB) design literature solutions are presented for the treatment parameters when the treatment totals are available and block effects are present. In this section we present solutions for treatment parameters when only block totals are available and when there are no block effects. The BIB designs to be considered herein consist of an arrangement of v treatments in b blocks each of size k ($< v$) such that

(i) every treatment occurs at most once in a block,

(ii) every treatment is replicated r times, and

(iii) every pair of treatments occurs together in exactly λ blocks,

where v, b, r, k, and λ are known as the parameters of the BIB design. They satisfy the relations:

$$vr = kb, \quad r(k-1) = \lambda(v-1), \quad b \geq v\ . \tag{A.1}$$

The $v \times b$ matrix $N = (n_{ij})$, where $n_{ij} = 1$ or 0 accordingly as the i^{th} treatment occurs in the j^{th} block or not, is called the incidence matrix of the BIB design. It can be verified that

$$NN' = (r-\lambda)I_v + \lambda J_{v,v} \tag{A.2}$$

and that

$$(NN')^{-1} = \frac{1}{r-\lambda} I_v - \frac{\lambda}{rk(r-\lambda)} J_{v,v}\ , \tag{A.3}$$

where I_v is the identity matrix of order v, and $J_{v,v}$ is a $v \times v$ matrix whose entries are all ones.

In addition to other (0,1)-matrices, the incidence matrices of BIB designs can be used as spring balance weighing designs (cf. Raghavarao (1971)). The incidence matrix of the BIB design with parameters

$$v = 4t-1 = b, \quad r = 2t = k, \quad \lambda = t \ , \tag{A.4}$$

provides smallest generalized variance of estimated weights when used as a weighing design and when the spring balance scale is an unbiased one. BIB designs with parameters (A.4) are presumed to exist for all integral values of t. If v objects whose true weights are $\omega_1, \omega_2, \cdots, \omega_v$ are to be weighed in b weighings using an unbiased spring balance scale, then one may identify the objects as the treatments of a BIB design and weigh the k objects corresponding to each block of the design as a single lot to obtain observed weights as $y_1, y_2, \cdots, y_b$. Then, we have

$$E(\underline{y}) = N'\underline{\omega} \ , \tag{A.5}$$

where $\underline{y}' = (y_1 \ y_2 \ \cdots \ y_b)$, $\underline{\omega}' = (\omega_1 \ \omega_2 \ \cdots \ \omega_v)$, and $E(\underline{y})$ denotes the expected value of the random vector $\underline{y}$. If the dispersion matrix of $\underline{y}$, denoted by $V(\underline{y})$, is $\sigma^2 I_b$, then $\underline{\omega}$ can be estimated by $\hat{\underline{\omega}}$ given by the following equation:

$$\hat{\underline{\omega}} = (NN')^{-1}N\underline{y} = (N^{-1})\underline{y} \ ; \tag{A.6}$$

the variance of the estimated weights is given by

$$V(\underline{\omega}) = \sigma^2(NN')^{-1} \ . \tag{A.7}$$

It should be noted that when v = b and $V(\underline{y}) = \sigma^2 \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is any positive definite matrix of known constants, $\hat{\underline{\omega}}$ is given by (A.6) but its variance becomes

$$V(\hat{\underline{\omega}}) = \sigma^2(N\boldsymbol{\Lambda}^{-1}N')^{-1} \ . \tag{A.8}$$