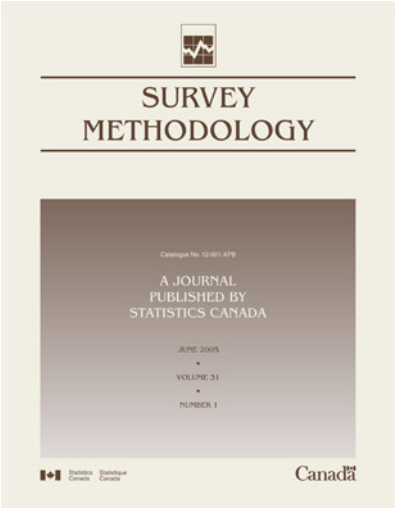
Catalogue no. 12-001-XIE

Survey Methodology

June 2005



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

June 2005

Published by authority of the Minister responsible for Statistics Canada

May 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 31, No. 1, pp. 41-51
Statistics Canada, Catalogue No. 12-001

Domain Estimators for the Item Count Technique

Takahiro Tsuchiya [1]

Abstract

The item count technique, which is an indirect questioning technique, was devised to estimate the proportion of people for whom a sensitive key item holds true. This is achieved by having respondents report the number of descriptive phrases, from a list of several phrases, that they believe apply to themselves. The list for half the sample includes the key item, and the list for the other half does not include the key item. The difference in mean number of selected phrases is an estimator of the proportion. In this article, we propose two new methods, referred to as the cross-based method and the double cross-based method, by which proportions in subgroups or domains are estimated based on the data obtained via the item count technique. In order to assess the precision of the proposed methods, we conducted simulation experiments using data obtained from a survey of the Japanese national character. The results illustrate that the double cross-based method is much more accurate than the traditional stratified method, and is less likely to produce illogical estimates.

Key Words: Indirect questioning techniques; Item count technique; Domain estimators; Survey of Japanese national character.

1. Introduction

1.1 Indirect Questioning Techniques

Suppose that a population U is divided into two sub-populations $U_{(T)}$ and $U^c_{(T)}$, where $U_{(T)}$ is a set of elements having an attribute T, and $U^c_{(T)}$ is a complement of $U_{(T)}$. One purpose of social surveys is to estimate $\pi = \overline{Y} = P(Y=1)$, where

$$Y_k = \begin{cases} 1 & \text{if } k \in U_{(T)} \\ 0 & \text{otherwise} \end{cases}$$

and $P(\cdot)$ denotes the proportion of units having a particular value of the variable. For example, when T is "supporting the present cabinet," π indicates the cabinet support rate, and when T is "using a certain illegal drug," π denotes the prevalence rate of drug use.

In a direct questioning technique, researchers ask respondents "Do you belong to $U_{(T)}$?," and directly obtain the indicator value y_i as "yes" or "no" (Cochran 1977, page 50). When every respondent has an equal inclusion probability, a sample mean $\bar{y}$ serves as one estimator of π.

On the other hand, some indirect questioning techniques, including the randomized response technique (Warner 1965), the nominative technique (Miller 1985), the item count technique (Droitcour, Caspar, Hubbard, Parsley, Visscher and Ezzati 1991), and the three-card technique (Droitcour, Larson and Scheuren 2001), are devised because some respondents tend to evade sensitive questions, such as those concerning highly private matters, socially unaccepted or deviant behaviors or illegal acts. The essential feature of indirect techniques is that instead of a direct observation of Y, another variable $X = g(Y, \boldsymbol{V})$, which is some sort of function of Y and, if necessary, of other random variables $\boldsymbol{V}$, is observed so that respondents feel that their true Y–values are not revealed. While this feature is expected to derive a truthful answer from evasive respondents, both the questioning and the estimation procedures are rather complicated compared to the direct questioning technique partly because the function $g(\cdot)$ sometimes includes some randomization processes. We shall outline two indirect techniques below.

The randomized response is the most popular among the indirect techniques, and various modifications have been proposed (Abul-Ela, Greenberg and Horvitz 1967; Warner 1971; Chaudhuri and Mukerjee 1988; Greenberg, Abul-Ela, Simmons and Horvitz 1969; Takahasi and Sakasegawa 1977). Although the randomized response is not the topic of this article, we shall briefly outline Warner's original procedure here for reference, because this technique will be simulated in a later section.

1. Prepare two types of questionnaires. In questionnaire A, respondents are asked "Do you belong to $U_{(T)}$?," and in questionnaire B, respondents are asked "Do you belong to $U^c_{(T)}$?"

2. Let $p(\neq 0.5)$ be the predetermined probability. Each respondent selects questionnaire A or B with probabilities p or $1-p$ respectively, but no one other than the respondent knows which questionnaire is selected.

1. Takahiro Tsuchiya, The Institute of Statistical Mathematics, 4-6-7, Minami-Azabu, Minato-ku, Tokyo, 106-8569, Japan. E-mail: taka@ism.ac.jp.

3. Suppose X is an indicator variable whose value is 1 if the response is "yes" or 0 if the response is "no." The estimator of π is given by

$$\hat{\pi} = \frac{p - 1 + \bar{x}}{2p - 1}, \qquad (1)$$

where $\bar{x}$ is a sample mean of X.

Since the researchers have no information regarding the type of questionnaire selected by each respondent, more respondents are expected to give truthful answers than they would if asked direct questions.

The item count technique, which is the subject of this article, is not as popular despite its simplicity. The technique is also effective when posing sensitive questions, because respondents are asked not to answer sensitive questions directly but to merely report the number of items that hold true with them. The following are the processes of the item count technique:

1. Prepare the key item T, which is the primary focus of the study, and G other non-key items $E_1, \ldots, E_G$. For example, T is "using a certain illegal drug" as mentioned above, and E_g is some sort of non-sensitive description such as "owning a bicycle."

2. Prepare two types of questionnaires, A and B. In questionnaire A, respondents are asked to answer the number C^A of items that are true with respect to themselves among G non-key items. In questionnaire B, respondents are asked to answer the number C^B of items that are true with respect to themselves out of $G+1$ items, including the key item T.

   Table 1 lists examples of item lists. Our aim is to estimate the proportion of people who use a certain illegal drug. The key item is "using a certain illegal drug" in the questionnaire B and the other four items are non-key items. Except when a response to the questionnaire B is $C^B = 0$ or $C^B = 5$, researchers cannot detect as to which items hold true with the respondent. For example, a respondent will reply that four items in the questionnaire B are true, but we cannot be sure that the respondent uses the drug at all. Hence, it is expected that more respondents using an illegal drug will report truthful answers in such a scenario than when asked a direct question.

3. Divide a total sample into two subgroups, A and B, randomly of size n^A and n^B so that each questionnaire is assigned to a corresponding subgroup.

Table 1
Examples of Item Lists

Questionnaire A	Questionnaire B
How many of the following hold true for you? – owning a bicycle – having travelled abroad – having called an ambulance – owning a summer villa	How many of the following hold true for you? – owning a bicycle – having travelled abroad – having called an ambulance – using a certain illegal drug – owning a summer villa

4. The estimator of π is given by

$$\hat{\pi} = \hat{\bar{C}}^B - \hat{\bar{C}}^A, \qquad (2)$$

where $\hat{\bar{C}}^A$ and $\hat{\bar{C}}^B$ are the estimated means of C^A and C^B respectively. The justification of (2) is explained in section 2.1. When every unit in the sample has an equal inclusion probability, $\hat{\pi}$ can be written as

$$\hat{\pi} = \sum_{c=0}^{G+1} c \frac{n_c^B}{n^B} - \sum_{c=0}^{G} c \frac{n_c^A}{n^A}, \qquad (3)$$

where n_c^A and n_c^B are the number of respondents whose answers are $C^A = c$ and $C^B = c$, respectively. Moreover, when an auxiliary variable Z is available and its distribution $P(Z = z) = m_z$ in the population is known, for example from a census, poststratification is often used to adjust the sample distribution of Z to the population. That is, the poststratified estimator of π is given by

$$\begin{aligned}\hat{\pi}_{\mathrm{PS}} &= \sum_{c=0}^{G+1} c \frac{\sum_z v_z^B n_{cz}^B}{n^B} - \sum_{c=0}^{G} c \frac{\sum_z v_z^A n_{cz}^A}{n^A} \\ &= \sum_{c=0}^{G+1} c \sum_z \frac{m_z}{n_{\cdot z}^B} n_{cz}^B - \sum_{c=0}^{G} c \sum_z \frac{m_z}{n_{\cdot z}^A} n_{cz}^A, \end{aligned} \qquad (4)$$

where n_{cz}^A is the number of respondents for each $C^A = c$ and $Z = z$,

$$n_{\cdot z}^A = \sum_{c=0}^{G} n_{cz}^A, \; n^A = \sum_z n_{\cdot z}^A, \; v_z^A = \frac{m_z n^A}{n_{\cdot z}^A}$$

and $n_{cz}^B, n_{\cdot z}^B, n^B$, and v_z^B are defined in analogous ways.

One practical merit of the item count technique is that it does not demand any randomization devices, which are required for the randomized response technique. It is not the respondent but a researcher who selects the questionnaire to be answered. Hence, the item count technique is easily implemented via any self-administered or telephone surveys. A more elaborate comparison between the randomized response and the item count technique is found in Hubbard, Casper and Lessler (1989).

The questionnaire A is introduced to obtain the distribution of the number of non-key items. That is, respondents to the questionnaire A do not answer the sensitive question. Therefore, it is possible to increase the precision of the estimator using the double-list version of item count (Droitcour *et al*. 1991), which exchanges the roles between the two subgroups. However, we limit our argument in this article to a single-list version, because the extension of estimators to the double-list version is straightforward.

1.2 Purpose of this Article

Thus far, we have focused on the parameter $\pi = \overline{Y} = P(Y=1)$ of a total population. However, estimators in subpopulations or domains (Särndal, Swesson and Wretman 1992 page 5) are often required, *i.e*., either a conditional proportion $P(Y=1|Z=z)$ or a joint proportion $P(Y=1, Z=z)$ must be estimated, where a population is divided into several domains by the Z–value. We refer to the variable Z as the domain variable in this article. The domain variables often used are demographic characteristics such as gender or age. For example, government agencies would like to know the proportion of people who use a certain illegal drug at each age group. Even though the post-stratified estimator $\hat{\pi}_{\mathrm{PS}}$ in (4) uses the domain variable Z, its aim is an estimation of $P(Y=1)$ in the entire population. Our aim in this article is to obtain separate estimations of $P(Y=1|Z=z)$ within each domain.

One simple estimation method is as follows:

1. Post-stratify the sample into strata or domains based on the Z–value.
2. In each stratum or domain, separately determine $p(Y=1|Z=z)$ using (1) or (2), where $p(\cdot)$ is a sample estimate of $P(\cdot)$.
3. If necessary, estimate $p(Y=1, Z=z)$ by multiplying a known domain proportion, $P(Z=z)$, or an estimated domain proportion, $p(Z=z)$.

The above method is referred to throughout this article as a stratified method because estimates are obtained separately in each stratum or domain. Although Rao (2003) refers to the above method as a direct estimate, we have avoided the use of the term "direct" in order to avoid confusion with the term "direct questioning technique."

An advantage of the stratified method is that this method is applicable to any indirect questioning technique, including the randomized response and item count techniques. The U.S. General Accounting Office (1999) adopts the stratified method to estimate domains under the three-card technique. However, one of the serious problems of the stratified method is that it often produces illogical estimates, especially negative estimates, in the case of the randomized response and the item count, as explained later in this article. This is mainly because the reduction of the sample size in each stratum increases the standard errors of the estimators (Lessler and O'Reilly 1997). For example, Droitcour *et al*. (1991, page 206) "calculated estimates separately for the three risk strata" and obtained negative prevalence rate estimates of drug use.

In the case of the randomized response, there is little possibility that domain estimators other than the stratified method are developed because information concerning the type of questionnaire selected by individual respondents is unavailable. In contrast, in the item count technique, the questionnaire answered by each respondent is known. Therefore, the precision of the domain estimators is expected to increase when auxiliary information is used, specifically contingency tables between Z and C^A or C^B.

In this article, we propose new domain estimators for the item count technique, which are referred to as the cross-based method and the double cross-based method. In addition, we will illustrate the fact that the new estimators are more efficient than the traditional stratified method by simulating the item count technique using data obtained from the survey of the Japanese national character concerning the significant attributes of the Japanese character.

2. Domain Estimators for the Item Count Technique

2.1 Stratified Method

Here, we reformulate the stratified method. Let us assume that the following equations hold true for each value of c and z.

Assumption 1.

$$\begin{aligned} P(C^B = c|Z=z) = P(C^A = c, Y=0|Z=z) \\ + P(C^A = c-1, Y=1|Z=z), \\ P(C^A = G+1, Y=0|Z=z) = 0. \end{aligned}$$

These assumptions imply that the difference in the distribution between C^A and C^B depends solely on Y. Question effects, including order effects and context effects (Schuman and Presser 1981) are not considered.

We have the following result based on these assumptions.

Stratified Method.

$$\begin{aligned} P(Y=1|Z=z) &= \sum_{c=0}^{G+1} c\, P(C^B = c|Z=z) \\ &\quad - \sum_{c=0}^{G} c\, P(C^A = c|Z=z) \qquad (5) \\ &= \overline{C}_z^B - \overline{C}_z^A, \qquad (6) \end{aligned}$$

where $\overline{C}_z^A$ and $\overline{C}_z^B$ are the domain means of C^A and C^B.

Derivation.

$$\sum_{c=0}^{G+1} cP(C^B = c|Z = z)$$

$$= \sum_{c=0}^{G+1} cP(C^A = c, Y=0|Z=z) + \sum_{c=0}^{G+1} cP(C^A = c-1, Y=1|Z=z)$$

$$= \sum_{c=0}^{G} cP(C^A = c, Y=0|Z=z) + \sum_{c=0}^{G} (c+1)P(C^A = c, Y=1|Z=z)$$

$$= \sum_{c=0}^{G} c\{P(C^A = c, Y=0|Z=z) + P(C^A = c, Y=1|Z=z)\}$$

$$+ \sum_{c=0}^{G} P(C^A = c, Y=1|Z=z)$$

$$= \sum_{c=0}^{G} cP(C^A = c|Z=z) + P(Y=1|Z=z).$$

Transposing the first term to the left-hand side yields the stratified method (5).

The estimator $p(Y=1|Z=z)$ is obtained by substituting domain means $\overline{C}_z^A$ and $\overline{C}_z^B$ with their estimators, $\hat{\overline{C}}_z^A$ and $\hat{\overline{C}}_z^B$.

$$p(Y=1|Z=z) = \hat{\overline{C}}_z^B - \hat{\overline{C}}_z^A. \tag{7}$$

When the inclusion probabilities are equal for all units in the sample, the estimator of $P(Y=1|Z=z)$ is written as

$$p(Y=1|Z=z) = \sum_{c=0}^{G+1} c\frac{n_{cz}^B}{n_{\cdot z}^B} - \sum_{c=0}^{G} c\frac{n_{cz}^A}{n_{\cdot z}^A}, \tag{8}$$

where $n_{cz}^A, n_{cz}^B, n_{\cdot z}^A$, and $n_{\cdot z}^B$ are defined in the section 1.1. The equations (2) and (3) for the entire population are special cases of (7) and (8).

One merit of the stratified method is that the variance estimator of $p(Y=1|Z=z)$ is easily obtained by

$$\hat{\mathrm{V}}\mathrm{ar}\,(p(Y=1|Z=z)) = \hat{\mathrm{V}}\mathrm{ar}\,(\hat{\overline{C}}_z^B) + \hat{\mathrm{V}}\mathrm{ar}\,(\hat{\overline{C}}_z^A). \tag{9}$$

On the other hand, as noted in the previous section, the reduction of sample size in each stratum increases estimated variances in (9). Further, the marginal estimator $p(Y=1)$ obtained by using (8) does not correspond to that obtained directly by (3), unless $n_{\cdot z}^A = n_{\cdot z}^B$ for all z. That is, when $p(Z=z)$ is not known, its estimator is given by

$$p(Z=z) = (n_{\cdot z}^A + n_{\cdot z}^B)/(n^A + n^B)$$

and

$$\sum_z p(Y=1|Z=z)p(Z=z)$$

$$= \sum_z \frac{n_{\cdot z}^A + n_{\cdot z}^B}{n^A + n^B}\left\{\sum_{c=0}^{G+1} c\frac{n_{cz}^B}{n_{\cdot z}^B} - \sum_{c=0}^{G} c\frac{n_{cz}^A}{n_{\cdot z}^A}\right\}$$

$$\neq \sum_{c=0}^{G+1} c\frac{n_{c\cdot}^B}{n^B} - \sum_{c=0}^{G} c\frac{n_{c\cdot}^A}{n^A} = \hat{\pi}. \tag{10}$$

When the domain proportion $p(Z=z) = m_z$ is available, the marginal estimator corresponds to the poststratified estimator (4).

$$\sum_z p(Y=1|Z=z)P(Z=z)$$

$$= \sum_z m_z\left\{\sum_{c=0}^{G+1} c\frac{n_{cz}^B}{n_{\cdot z}^B} - \sum_{c=0}^{G} c\frac{n_{cz}^A}{n_{\cdot z}^A}\right\}$$

$$= \hat{\pi}_{\mathrm{PS}}.$$

These results indicate that we should use a poststratified estimator $\hat{\pi}_{\mathrm{PS}}$ with the domain estimators if we use the stratified method.

2.2 Cross-based Method

In the stratified method, a total sample is divided into strata for the purpose of direct estimation of $P(Y=1|Z=z)$, which causes sample size reduction. Hence, in the cross-based method proposed in this section, the joint proportion $P(Y=1, Z=z)$ is estimated first in order to use the entire sample, and the conditional proportion is subsequently obtained by

$$p(Y=1|Z=z) = \frac{p(Y=1, Z=z)}{p(Z=z)}$$

$$\text{or} \quad p(Y=1|Z=z) = \frac{p(Y=1, Z=z)}{P(Z=z)}.$$

The term 'cross-based method' is used because this method uses cross tabulations $P(Z=z|C^B=c)$, as shown in (19).

For the cross-based method, we assume that the following equations hold for each value of *c*.

Assumption 2.

$$P(C^B = c+1, Y=1) = P(C^A = c, Y=1), \tag{11}$$

$$P(C^B = 0, Y=1) = P(C^A = -1, Y=1) = 0, \tag{12}$$

$$P(C^B = c, Y=0) = P(C^A = c, Y=0). \tag{13}$$

These assumptions also imply that the difference in the distribution between C^A and C^B depends only on *Y*.

We have the following result based on these assumptions.

Cross-based Method.

$$P(Y=1, Z=z) = \sum_{c=1}^{G+1} P(Z=z|C^B=c)Q_{c-1}, \tag{14}$$

where

$$Q_c = \sum_{d=0}^{c}\{P(C^A = d) - P(C^B = d)\}.$$

In addition, we assume that $P(Z=z|C^B=c,\ Y=1) = P(Z=z|C^B=c)$ for every $c > 0$. This assumption would be valid to some degree when both the key and non-key items describe the same type of stigmatizing behavior.

Derivation.

Based on the assumptions, we have

$$P(C^B = c) = P(C^B = c, Y = 1) + P(C^B = c, Y = 0)$$
$$= P(C^A = c - 1, Y = 1) + P(C^A = c, Y = 0). \quad (15)$$

The following equation holds for any c.

$$P(C^A = c, Y = 0) = P(C^A = c) - P(C^A = c, Y = 1). \quad (16)$$

Hence, substituting (16) in (15) gives

$$P(C^B = c) = P(C^A = c - 1,\ Y = 1) + \{P(C^A = c) - P(C^A = c, Y = 1)\}. \quad (17)$$

Summing (17) over c up to some g, we obtain

$$\sum_{c=0}^{g} P(C^B = c) = \sum_{c=0}^{g} P(C^A = c - 1,\ Y = 1) + \sum_{c=0}^{g} \{P(C^A = c) - P(C^A = c, Y = 1)\}$$
$$= \sum_{c=0}^{g} P(C^A = c) - P(C^A = g, Y = 1).$$

By transposing the terms, we define Q_c.

$$Q_c = \sum_{d=0}^{c} \{P(C^A = d) - P(C^B = d)\}$$
$$= P(C^A = c, Y = 1)$$
$$= P(C^B = c + 1, Y = 1). \quad (18)$$

Here, the joint proportion $P(Y = 1, Z = z)$ is decomposed as

$$P(Y = 1, Z = z) = \sum_{c=0}^{G+1} P(Z = z \mid C^B = c) P(C^B = c, Y = 1). \quad (19)$$

Substituting the equation (18) and the assumption (12) in (19) yields the cross-based method.

The joint estimator $P(Y = 1, Z = z)$ is obtained by substituting each term of (14) for its estimators. When the sample is self-weighting, the estimator is given by

$$P(Y = 1, Z = z) = \sum_{c=1}^{G+1} \frac{n^B_{cz}}{n^B_{c\cdot}} \sum_{d=0}^{c-1} \left(\frac{n^A_{d\cdot}}{n^A} - \frac{n^B_{d\cdot}}{n^B} \right), \quad (20)$$

where

$$n^A_{c\cdot} = \sum_z n^A_{cz} \quad \text{and} \quad n^B_{c\cdot} = \sum_z n^B_{cz}.$$

The conditional estimator $p(Y = 1 \mid Z = z)$ is obtained by dividing $p(Y = 1 \mid Z = z)$ by the domain proportions $P(Z = z)$ or their estimators $p(Z = z)$.

As noted above, the main feature of the cross-based method is that $p(Y = 1,\ Z = z)$ is first estimated using the entire sample. Hence, the variance of $p(Y = 1 \mid Z = z)$ for the cross-based method is expected to be smaller than that of $p(Y = 1 \mid Z = z)$ for the stratified method. Moreover, negative values will seldom be obtained in the case of the cross-based method, while the negative values will be often obtained in the case of the stratified method. Furthermore, the marginal estimator $p(Y = 1)$ obtained by summing (20) is equal to the estimator (3), unless $n^B_{c\cdot} = 0$ for some c:

$$\sum_z p(Y = 1, Z = z) = \sum_z \sum_{c=1}^{G+1} \frac{n^B_{cz}}{n^B_{c\cdot}} \sum_{d=0}^{c-1} \left(\frac{n^A_{d\cdot}}{n^A} - \frac{n^B_{d\cdot}}{n^B} \right)$$
$$= \sum_{c=1}^{G+1} \sum_{d=0}^{c-1} \left(\frac{n^A_{d\cdot}}{n^A} - \frac{n^B_{d\cdot}}{n^B} \right)$$
$$= \sum_{c=1}^{G+1} \left\{ \left(1 - \sum_{d=c}^{G} \frac{n^A_{d\cdot}}{n^A} \right) - \left(1 - \sum_{d=c}^{G+1} \frac{n^B_{d\cdot}}{n^B} \right) \right\}$$
$$= \sum_{c=0}^{G+1} c \frac{n^B_{c\cdot}}{n^B} - \sum_{c=0}^{G} c \frac{n^A_{c\cdot}}{n^A} = \hat{\pi}. \quad (21)$$

Of course, when the domain proportions $P(Z = z) = m_z$ are known, we can use them to obtain a poststratified estimator $p(C^A = d)$ of $P(C^A = d)$ in Q_{c-1} of (14),

$$p(C^A = d) = \sum_z \frac{m_z}{n^B_{\cdot z}} n^B_{dz}.$$

In this case, $\sum_z p(Y = 1, Z = z)$ coincides with the post-stratified estimator $\hat{\pi}_{\mathrm{PS}}$.

One drawback of the cross-based method is that the variance of $p(Y = 1 \mid Z = z)$ is almost impossible to estimate algebraically. Hence, some resampling methods such as the jacknife or bootstrap would be necessary. Additionally, since it is impossible to determine the more efficient method between the stratified method and the cross-based method, simulation studies shall be conducted in a later section.

2.3 Double Cross-based Method

Before proceeding to the simulation study, we suggest a modified version of the cross-based method. In equation (19) of the cross-based method, we use $P(Z = z \mid C^B = c)$. In the same way, when $P(Z = z \mid C^A = c)$ is used, we obtain

$$P(Y = 1, Z = z) = \sum_{c=0}^{G} P(Z = z \mid C^A = c) P(C^A = c, Y = 1)$$
$$= \sum_{c=0}^{G} P(Z = z \mid C^A = c) Q_c. \quad (22)$$

Hence, a double cross-based method is obtained by combining (14) and (22) as follows:

$$P(Y = 1, Z = z) = \sum_{c=0}^{G} \left\{ \begin{matrix} w^A P(Z = z \mid C^A = c) \\ + w^B P(Z = z \mid C^B = c + 1) \end{matrix} \right\} Q_c, \quad (23)$$

where w^A and w^B are the non-negative weights for each subgroup, the sum of which is equal to one.

The following equation also holds for the double cross-based method of any w^A and w^B, unless $n_{c\cdot}^A = 0$ or $n_{c\cdot}^B = 0$ for some c.

$$\sum_z p(Y=1, Z=z) = \hat{\pi}. \qquad (24)$$

3. Numerical Experiments

3.1 Data Set

In order to compare the precision of the estimators, we conducted simulation experiments using data obtained from the survey of the Japanese national character (Sakamoto, Tsuchiya, Nakamura, Maeda and Fouse 2000). Although the respondents were selected via a stratified two-stage sampling from Japanese aged 20 and over, we neglect the sampling design because the collected sample of $N = 1{,}339$ is treated as the "true" population in this experiment. Table 2 lists the results of a question concerning the significant attributes of the Japanese character. Respondents were asked in a face-to-face interview to choose as many adjectives from among ten alternatives as they thought described the Japanese character.

Table 2
Significant Attributes of Japanese character

					$N = 1{,}339$
(Hand card) Which of the following adjectives do you think describes the character of the Japanese people? Choose as many as you like.					
1	Rational	18%	6	Kind	42%
2	Diligent	71%	7	Original	7%
3	Free	13%	8	Polite	50%
4	Open, frank	14%	9	Cheerful	8%
5	Persistent	51%	10	Idealistic	23%

The form of this question is different from that of the item count technique. In the item count technique, the respondent is asked to "answer the number of adjectives." In contrast, in this survey the respondent is asked to "circle as many adjectives you feel are appropriate." In addition, the ten items are not very sensitive, hence the respondents should not hesitate during the selection. However, since the real contingency table between each of the ten items and another variable Z is obtained, we can evaluate the performance of estimators through a pseudo item count procedure.

We took each of the following three items as the key item Y, where $Y = 1$ implies that the item was selected.

– 7 Original (π is the least among the ten items)

– 8 Polite (π is just 50%)

– 2 Diligent (π is the largest among the ten items)

Three combinations of non-key items are used, as listed in Table 3. Combination 1 comprises two items with low proportions, while combination 2 comprises two items with high proportions. Combination 3 is the case with the maximum number of non-key items.

Table 3
Three Combinations of Non-key Items

	Non-key items	
Combination 1 ($G = 2$):	9 Cheerful	(8%)
	3 Free	(13%)
Combination 2 ($G = 2$):	5 Persistent	(51%)
	6 Kind	(42%)
Combination 3 ($G = 9$):	Nine items other than the key item	

We used either gender or age as the domain variable Z. Gender is either male or female, and the age categories are "20 – 29," "30 – 39," "40 – 49," "50 – 59," "60 – 69", and "70 and over."

3.2 Direct Questioning Versus Item Count Technique

3.2.1 Simulation Methods

First, we compare the standard errors between the direct questioning and the item count techniques. In this experiment, we attempted one combination of "7 Original" (key item), combination 3 (non-key items), and gender (domain variable). The contingency table based on the entire sample of $N = 1{,}339$ is listed in Table 4.

Table 4
A Contingency Table Between "7 Original" and Gender

	7 Original					
	$Y = 1$		$Y = 0$		Total	
Male	46	(7.5)	569	(92.5)	615	(100.0)
Female	51	(7.0)	673	(93.0)	724	(100.0)
Total	97	(7.2)	1,242	(92.8)	1,339	(100.0)

The simulation was conducted through the following procedures:

Step 1. Suppose the total sample of $N = 1{,}339$ to be a population.

Step 2. Draw a subsample S of size Nf where f is a sampling fraction with a simple random sampling without replacement.

Step 3. As the simulated result of the direct questioning method, compute the proportion directly, $p(Y = 1 | Z = \text{male})$ and $p(Y = 1 | Z = \text{female})$.

Step 4. Divide the subsample S into two groups S^A and S^B of size n^A and n^B that are not necessarily of equal size. Count the number C^A of selected non-key items for each respondent in S^A. Also, count the number C^B of selected items including both the key item and the non-key items in S^B.

Step 5. As the simulated result of the item count technique, compute $p(Y=1|Z=\text{male})$ $p(Y=1|Z=\text{female})$ and via the three estimation methods; stratified method, cross-based method, and double cross-based method. In the double cross-based method, we let $w^A = n^A/(n^A+n^B)$ and $w^B = n^B/(n^A+n^B)$.

Step 6. We let $f=0.1$ in step 2 and perform steps 2 to 5 for 2,000 iterations. Calculate the means $E_D, E_S, E_C,$ and E_W and the standard deviations $SE_D, SE_S, SE_C,$ and SE_W of each estimation method to approximate the expectations and the standard errors of the estimators, where the subscripts *D*, *S*, *C*, and *W*, indicate the direct questioning method, the stratified method, the cross-based method, and the double cross-based method, respectively. In the same way, we let $f=0.2$ and perform steps 2 to 5 for 2,000 iterations, and so on up to and including $f=0.9$.

3.2.2 Simulation Results

Figure 1 shows the approximated expectations and standard errors of the estimators. The horizontal axes indicate sampling fraction *f*. In both the cases, male and female, the approximated expectations of E_D are stable at every *f*–value while $E_S, E_C,$ and E_W of the item count technique fluctuate irregularly. This is because randomness is introduced twice under the item count, *i.e.*, in the sampling phase and in the division phase, whereas randomness is introduced only in the sampling phase under the direct questioning scenario. Even if $f=1,$ the estimator under the item count technique has a certain amount of variance due to the randomness at the division phase. As the range of fluctuation was negligible compared to the magnitude of the standard errors, which are referred to below, we concluded that the number of repetition was sufficient.

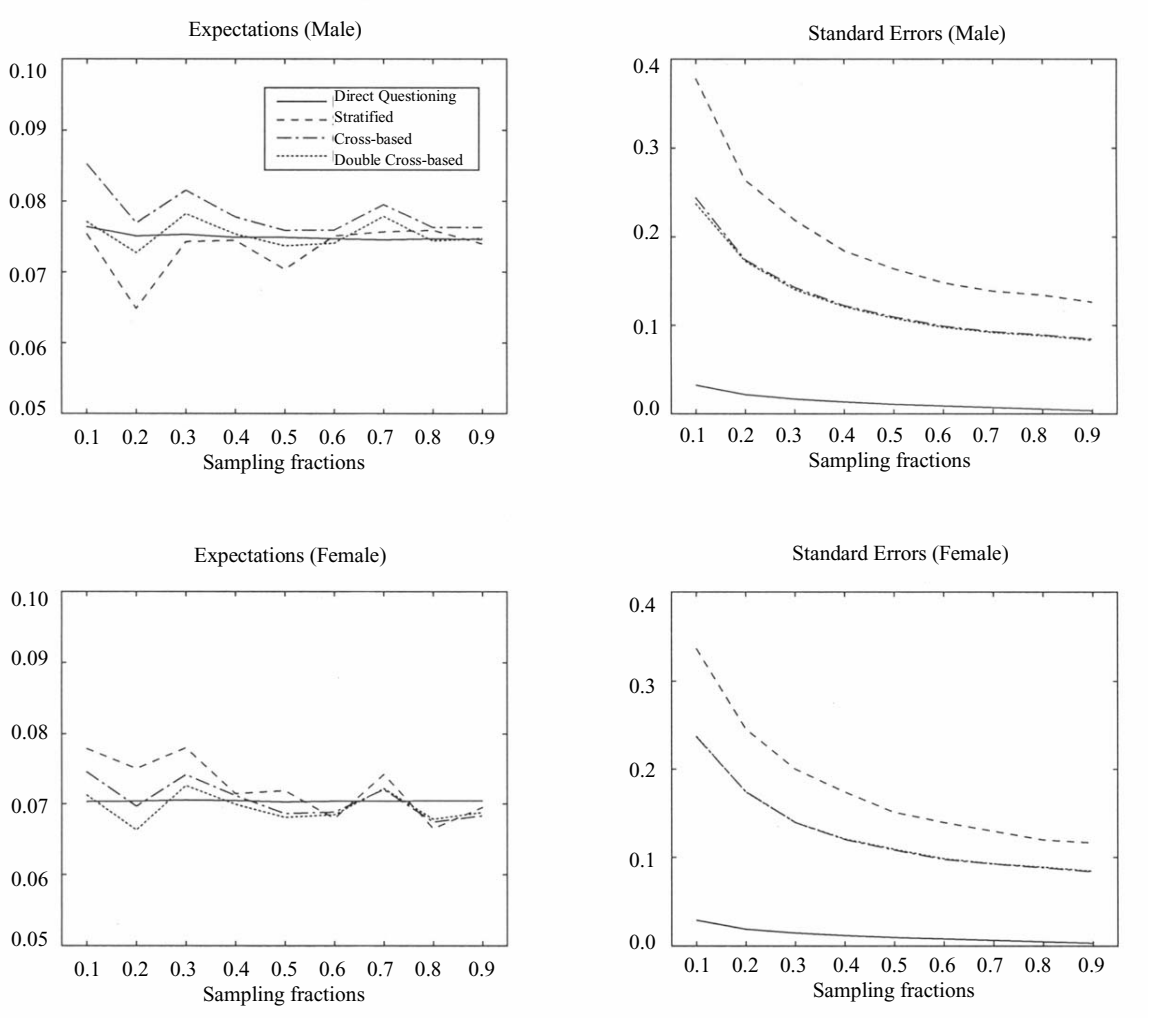


Figure 1. Approximated Expectations and Standard Errors of Estimators.

The standard errors, SE_D, of the direct questioning method is considerably small compared to those of the item count. In the case of the item count, standard errors do not converge to zero even if $f = 1$. As noted above, this is because the randomness is also introduced in the division phase. The standard errors of the stratified method are obviously larger than those of the two cross-based methods. The lines indicating the results for the cross-based method and the double cross-based method almost overlap, and appear to have no outstanding differences.

In order to evaluate the amount of variances or standard errors of estimators, let us consider the following indices that are analogous to the design effect (Kish 1965),

$$\mathrm{Def}_{M_1, M_2} = \frac{SE^2_{M_1}}{SE^2_{M_2}},$$

where M_1 and M_2 indicate one of the four methods *D*, *S*, *C*, and *W*. Although we have omitted the detailed results, roughly summarized, $\mathrm{Def}_{C,D}$ ranges from 50 (when $f = 0.1$) to 700 (when $f = 0.9$). That is, even if we use the cross-based method, the standard errors of the item count inflate nearly seven- to twenty-six-fold as compared to the direct questioning. However, the variance reduction attained by using the double cross-based method instead of the stratified method ranges from $\mathrm{Def}_{W,S} = 0.39$ (male) to 0.55 (female). In other words, the standard errors of the double cross-based method are reduced to about 62 percent of the stratified estimate at the minimum, and 74 percent at the maximum.

3.3 Stratified Versus Cross-based Method

3.3.1 Simulation Methods

In the previous section, the precision of the cross-based and the double cross-based method appeared to be larger than those of the stratified method. We shall check the precision of these methods for other combinations of the key item, the combination of non-key items, and the domain variable *Z* by simulation experiments.

In this section, we used all samples as follows:

Step 1. Compute $P(Y = 1 | Z = z)$ for each z based on all data of size $N = 1{,}339$.

Step 2. Divide the total sample ($N = 1{,}339$) randomly into group *A* and group *B* of size n^A and n^B where $N = n^A + n^B$.

Step 3. Count the number C^A of selected non-key items for each respondent of group *A*, and count the number C^B of selected items, including both the key item and non-key items, in group *B*.

Step 4. Estimate $p(Y = 1 | Z = z)$ by the stratified method, the cross-based method, and the double cross-based method, respectively.

Step 5. Compute the chi-squared distance e^2 between $P(Y = 1 | Z = z)$ and $p(Y = 1 | Z = z)$ for each method.

$$e^2 = \sum_z \frac{\{p(Y = 1 | Z = z) - P(Y = 1 | Z = z)\}^2}{P(Y = 1 | Z = z)}$$

Step 6. Repeat the above procedure from step 2 through step 5 for 1,000 iterations. Calculate the means and the standard deviations of e^2 for each method.

In addition, we simulated the stratified method under the randomized response for references via the following procedure:

Step 1. Let *p* be a proportion as described below. Divide the total sample ($N = 1{,}339$) randomly into two groups. Group *A* is composed of Np respondents, and group *B* is composed of $N(1 - p)$ respondents.

Step 2. Let n_z^A be the number of respondents who selected the key item and $Z = z$ in group *A*. Let n_z^B be the number of respondents who did not select the key item and $Z = z$ in group *B*. Let n_z be the number of respondents with $Z = z$. Compute

$$p(Y = 1 | Z = z) = \frac{n_z}{1{,}339}\left(\frac{p - 1 + (n_z^A + n_z^B)/n_z}{2p - 1}\right).$$

Step 3. Calculate e^2 employing the same equation as used in the item count technique.

Step 4. Repeat the above procedure from step 1 through step 3 for 1,000 iterations. Calculate the means and the standard deviations of e^2 for each method.

We used three *p* values; $p = 0.2$, $p = 0.3$, and $p = 0.4$.

3.3.2 Simulation Results

Table 5 and Table 6 list the means and the standard deviations of 1,000 e^2s for the domain variable *Z* of gender and age, respectively. A smaller mean of "e^2–value" indicates that the domain estimators are more precise. In some repetitions, illogical estimates $p(Y = 1 | Z = z)$, which deviate from the range [0, 1], were obtained. The columns of the tables denoted by "under" indicate the number of repetitions when at least one of the estimates $p(Y = 1 | Z = z)$ was under 0, and "over" indicates that the estimates were over 1. Ideally, the figures of the columns of "illogical *p*" should be 0.

Table 5
Means and Standard Deviations of e^2s and Number of Times Illogical Estimates were Obtained (Domain Variable Z is Gender)

	7 Original (7%)				8 Polite (50%)				2 Diligent (71%)			
	e^2 –value		illogical p		e^2 –value		illogical p		e^2 –value		illogical p	
	mean	(s.d.)	under	over	mean	(s.d.)	under	over	mean	(s.d.)	under	over
Stratified method												
Combination 1	38	(36)	39	0	6	(6)	0	0	4	(4)	0	0
Combination 2	89	(92)	179	0	16	(17)	0	0	10	(11)	0	0
Combination 3	341	(330)	457	0	44	(43)	0	0	33	(32)	0	7
Cross-based method												
Combination 1	18	(24)	1	0	4	(5)	0	0	3	(3)	0	0
Combination 2	45	(65)	41	0	10	(12)	0	0	7	(8)	0	0
Combination 3	163	(239)	186	0	22	(31)	0	0	17	(23)	0	1
Double cross-based method												
Combination 1	18	(24)	1	0	3	(4)	0	0	2	(3)	0	0
Combination 2	45	(65)	31	0	9	(12)	0	0	6	(8)	0	0
Combination 3	163	(240)	177	0	21	(31)	0	0	16	(23)	0	0
Randomized response												
$p = 0.2$	12	(14)	0	0	3	(3)	0	0	2	(2)	0	0
$p = 0.3$	35	(43)	41	0	8	(7)	0	0	5	(5)	0	0
$p = 0.4$	158	(181)	305	0	35	(34)	0	0	23	(23)	0	3

Note: e^2 –value is multiplied by 10^3.

Table 6
Means and Standard Deviations of e^2s and Number of Times Illogical Estimates were Obtained (Domain Variable Z is age)

	7 Original (7%)				8 Polite (50%)				2 Diligent (71%)			
	e^2 –value		illogical p		e^2 –value		illogical p		e^2 –value		illogical p	
	mean	(s.d.)	under	over	mean	(s.d.)	under	over	mean	(s.d.)	under	over
Stratified method												
Combination 1	375	(226)	609	0	60	(39)	0	0	39	(26)	0	0
Combination 2	859	(507)	799	0	152	(91)	0	0	97	(58)	0	18
Combination 3	3,410	(2,108)	926	1	446	(290)	48	41	333	(217)	9	353
Cross-based method												
Combination 1	93	(82)	8	0	32	(20)	0	0	28	(16)	0	0
Combination 2	175	(195)	138	0	80	(42)	0	0	59	(33)	0	0
Combination 3	536	(733)	273	0	89	(95)	0	0	70	(71)	0	10
Double cross-based method												
Combination 1	70	(75)	8	0	13	(13)	0	0	9	(8)	0	0
Combination 2	153	(202)	93	0	45	(35)	0	0	31	(23)	0	0
Combination 3	526	(745)	246	0	72	(94)	0	0	52	(70)	0	1
Randomized response												
$p = 0.2$	158	(101)	284	0	25	(14)	0	0	17	(11)	0	0
$p = 0.3$	476	(294)	720	0	74	(42)	0	0	51	(31)	0	2
$p = 0.4$	2,181	(1,348)	945	0	335	(193)	9	9	232	(136)	0	217

Note: e^2 –value is multiplied by 10^3.

For every combination of the key item, the non-key items, and the domain variable Z, the means of e^2 of the double cross-based method are the smallest, and the cross-based method is the second smallest by a narrow margin. When π of the key item is low ("7 Original"), the number of non-key items is large (combination 3), and the number of alternatives of the domain variable Z is large (age), the accuracy of the stratified method decreases greatly compared to other combinations.

Moreover, when π of the key item is low, negative estimates are often observed when the stratified method is used. For example, when combining "7 Original," combination 3 and age, the frequency of observed negative estimates is 926 out of 1,000 iterations. When the double cross-based method is used, the negative estimates are less likely to be observed.

For randomized response, when the number of alternatives of the domain variable Z is small (gender), the accuracy of the estimates seems to be the same as the cross-based and the double cross-based methods. However, the mean e^2 is somewhat larger than that of the cross-based method when the domain variable Z has many options (age).

The randomized response, for which only the stratified method is available, also suffers from negative estimates, particularly when π is small ("7 Original").

4. Conclusion

The following results were obtained through simulation experiments:

- The cross-based method or the double cross-based method, which is proposed in this article, should be used to estimate domain parameters when the data is obtained via the item count technique. In the first simulation, the variances of cross-based estimators were reduced to 39 percent of the variance of the stratified estimate at the minimum to 55 percent at the maximum. In the simulation studies, the double cross-based method made no drastic improvement in precision as compared to the cross-based method.

- Even when the double cross-based method is used, the standard errors of the domain estimators are much larger than those of the direct questioning technique.

The true $\pi = \overline{Y} = P(Y=1)$ of a question, to which respondents evade giving a truthful answer, would be often small. In addition, an indirect questioning technique is used in order to ensure protection of privacy. The respondents feel that their privacy is secured when many non-key items are included (Hubbard *et al.* 1989). The simulation studies show that in such situations, the cross-based method or double cross-based method is more efficient than the traditional stratified method.

The domain estimators obtained by the traditional stratified method are generally inconsistent with the estimator $\hat{\pi}$ as shown in (10). Poststratified estimator $\hat{\pi}_{PS}$ by the domain variable addressed is essential in order to ensure consistency. Alternatively, we have to divide the total sample into two subgroups so that the distributions of their domain variable match in advance. On the contrary, the domain estimators obtained by the cross-based and the double cross-based methods are consistent with $\hat{\pi}$ as shown in (21). However, it does not mean that the cross-based method automatically adjusts the two subgroups so that the sample distributions of the domain variable match between the two subgroups. For the cross-based method, post-stratification by the domain variables or other demographic variables is also admissible, but not indispensable.

Even when the double cross-based method is used, negative domain estimates are sometimes observed. It is possible to avoid negative estimates by letting a negative estimate q_c of Q_c in (23) be zero. However, such an adjustment produces a positive bias in $p(Y=1|Z=z)$.

The data of the survey of the Japanese national character, which were used in the simulation experiments, are neither sensitive nor were they obtained via the item count technique. In the future, the performance of the proposed method should be assessed by applying it to data obtained via the item count technique.

Acknowledgements

The author is grateful to two anonymous reviewers and an assistant editor for their helpful comments on a previous version of this paper.

References

Abul-Ela, Abdel-Latif, A., Greenberg, B.G. and Horvitz, D.G. (1967). A multiproportions RR model. *Journal of the American Statistical Association*, 62, 990-1008.

Chaudhuri, A., and Mukerjee, R.M. (1988). *Randomized Response: Theory and Techniques.* New York: Marcel Dekker.

Cochran, W.G. (1977). *Sampling Techniques, 3rd ed.* New York: John Wiley & Sons, Inc.

Droitcour, J., Caspar, R.A., Hubbard, M.L., Parsley, T.L., Visscher, W. and Ezzati, T.M. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. In *Measurement Errors in Surveys* (Eds. P.P. Biemer, *et al*.), New York: John Wiley & Sons, Inc.

Droitcour, J.A., Larson, E.M. and Scheuren, F.J. (2001). The three card method: Estimating sensitive survey items-with permanent anonymity of response. *Proceedings of the Social Statistics Section of the American Statistical Association.* Alexandria, V.A.: American Statistical Association.

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

Hubbard, M.L., Casper, R.A. and Lessler, J.T. (1989). Respondent reactions to item count lists and randomized response. *Proceedings of the Survey Research Section of the American Statistical Association.* Washington, D.C.: American Statistical Association. 544-548.

Kish, L. (1965). *Survey Sampling.* New York: John Wiley & Sons, Inc.

Lessler, J.T., and O'Reilly J.M. (1997). Mode of interview and reporting sensitive issues: Design and implementation of audio computer-assisted self-interviewing. *NIDA Research Monograph*, 167, 366-382.

Miller, J.D. (1985). The nominative technique: A new method of estimating heroin prevalence. *NIDA Research Monograph*, 57, 104-124.

Rao, J.N.K. (2003). *Small Area Estimation.* New Jersey: John Wiley & Sons, Inc.

Sakamoto, Y., Tsuchiya, T., Nakamura, T., Maeda, T. and Fouse, D.B. (2000). *A Study of the Japanese National Character: The Tenth Nationwide Survey (1998).* Tokyo: The Institute of Statistical Mathematics Research Report General Series 85.

Särndal, C.-E., Swesson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.

Schuman, H., and Presser, S. (1981). *Questions & Answers in Attitude Surveys.* New York: Academic Press.

Takahasi, K., and Sakasegawa, H. (1977). A randomized response technique without making use of any randomizing device. *Annals of the Institute of Statistical Mathematics*, 29, 1-8.

U.S. General Accounting Office (1999). *Survey Methodology. An Innovative Technique for Estimating Sensitive Items.* Washington D.C.: General Accounting Office.

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

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