Journal of Informatics and Mathematical Sciences
Volume 2 (2010), Numbers 2 & 3, pp. 161–169

http://www.rgnpublications.com

On Item Count Technique in Survey Sampling

Zawar Hussain and Javid Shabbir

Abstract. In this paper, an improved *Item Count Technique* (ICT) has been proposed. The major advantage of this technique is that it does not require two subsamples (as is the case in usual ICT) and there is no need of finding optimum subsample sizes. The proposed ICT has been observed performing well, as compared to the usual ICT, in terms of relative efficiency. The novel method of *Randomized Response* (RR) technique proposed by Warner (1965) has also been compared with the proposed ICT and it is found that the proposed technique uniformly performs better when the number of innocuous items is greater than 3.

1. Introduction

In estimating the population proportion of a sensitive characteristic (induced abortion, shoplifting, tax evasion) through direct questioning, truthfulness of the answers may be suspected due to various reasons, namely, social stigma, embarrassment, monetary penalty, etc. An ingenious alternative to direct questioning introduced by Warner (1965), known as randomized response technique, has been developed rapidly. For a good review of developments on randomized response techniques we would refer the reader to Tracy and Mangat (1996) and Chaudhuri and Mukerjee (1988). The RR technique has been used in many studies including Liu and Chow (1976), Reinmuth and Geurts (1975), Geurts (1980) and Larkins *et al.* (1997) etc. Geurts (1980) reported that RR technique had financial limitations since it requires larger sample sizes to obtain the confidence intervals comparable to the direct questioning technique. More time is needed to administer and explain the procedure to the survey respondents. In addition, tabulation and calculation of the results is comparatively laborious. Larkins *et al.* (1997) found that RR technique was not good for estimation the proportion of tax payers/non-payers. Dalton and Metzger (1992) were of the view that RR technique might not be effective through a mailed or telephonic survey. Hubbard *et al.* (1989) stated that the main technical problem for RR

2010 *Mathematics Subject Classification.* 62D05.
Key words and phrases. Item count technique; Survey sampling.

techniques is making the decision what kind of the randomization device would be the best in a given situation and the most crucial aspect of the RR technique is about the respondent's acceptance of the technique. Chaudhuri and Christofides (2007) also gave a criticism on the RR technique in the sense that it demands the respondent's skill of handling the device and also it asks respondents to report the information which may be useless or tricky. A clever respondent may also think that his/her reported response can be traced back to his/her actual status if he/she does not understand the mathematical logic behind the randomization device. Some of the alternatives to the RR technique include the Item Count Technique (Droitcour *et al.* 1991), the Three card method (Droitcour *et al.* 2001), and the Nominative technique (Miller, 1985). These alternatives are designed because, in general, respondent evade sensitive questions especially regarding personal issues, socially deviant behaviors or illegal acts. Chaudhuri and Christofides (2007) also added that in these three alternatives to RR respondents know that what they are revealing about themselves and they do not need to know about any special estimation technique Also respondents provide answers which make sense to them. Details about item count technique can be found in Droitcour *et al.* (1991, 2002). Dalton *et al.* (1994) named ICT as the unmatched count technique and applied it to study the illicit behaviors of the auctioneers and compared to direct questioning they obtained higher estimates of six stigmatized items. Wimbush and Daltons (1997) applied this technique in estimating the employee theft rate in high-theft-exposure business and found higher theft rates. Tsuchiya (2005) extended the ICT to Domain estimators by stratified method, cross-based method, and double Cross-based method. More recently, Tsuchiya *et al.* (2007) studied the properties of the ICT through an experimental web survey and found that ICT yielded higher estimates of the proportions of the shoplifters by nearly 10 percent as that of yielded by direct questioning. They also found that cross-based method was most appropriate method. Besides its fruitful applications ICT has not been found fruitful in many studies, for example, Droitcour *et al.* (1991), Biemer and Wright (2004) and Ahart and Sackett (2004) failed to get higher estimates in their studies of different stigmatized traits. This article is organized as follows. In Section2 we give a short sketch of usual ICT and then the proposed technique is described in Section 3 followed by efficiency comparison in Section 4.

2. **Item Count Technique**

To estimate the proportion of the people with stigmatizing attribute a promising indirect questioning technique, called item count technique, was introduced by Droitcour *et al.* (1991). It consists of taking two subsamples of sizes n_1 and n_2. The ith respondent in the first subsample are given a list of g innocuous items and asked to report the number, say X_i of items that are applicable to them. Similarly, the jth respondent in the second subsample is provided another list of $(g+1)$ items including the sensitive item and asked to report a number, say Y_j of the items that are applicable to them. The g innocuous items may or may not be same in both the subsamples. An unbiased estimator of proportion of the sensitive item in the population is given by

$$\widehat{\pi}_I = \bar{Y} - \bar{X}. \tag{2.1}$$

Its variance is given by

$$V(\widehat{\pi}_I) = \left[\frac{\pi(1-\pi)}{n_2} + \frac{\sum_{j=1}^{g} \theta_j \left(1 - \sum_{j=1}^{g} \theta_j\right)}{n_2} + \frac{\sum_{j=1}^{g} \theta_j \left(1 - \sum_{j=1}^{g} \theta_j\right) + \sum_{\substack{j,k=1 \\ j \neq k}}^{g} \theta_j \theta_k}{n_1} \right], \tag{2.2}$$

where θ_j is the known proportion of the item j in the population.

3. **Proposed Item Count Technique**

Each respondent in a sample of size n is provided a questionnaire consisting of g questions. The jth question consists of queries about an unrelated item (F_j) and the sensitive characteristic (S). The respondent is requested to count 1 if he/she possesses at least one of the characteristics F_j and S, otherwise count 0, as a response to the jth question. And report the total count based on entire questionnaire. Let Z_i be the total count of ith respondent, it can then be written as

$$Z_i = \sum_{j=1}^{g} \alpha_j, \tag{3.1}$$

where α_j can assume values "1" and "0" with probabilities $(\pi + \theta_j - \pi\theta_j)$ and $(1 - \pi + \theta_j - \pi\theta_j)$, respectively.

The expected value Z_i can be written as

$$\begin{aligned} E(Z_i) &= \sum_{j=1}^{g} E(\alpha_j) \\ &= g\pi + \sum_{j=1}^{g} \theta_j - \pi \sum_{j=1}^{g} \theta_j \\ &= \left(g - \sum_{j=1}^{g} \theta_j\right)\pi + \sum_{j=1}^{g} \theta_j. \end{aligned}$$

This suggests defining an unbiased estimator of π as

$$\widehat{\pi}_P = \frac{\bar{Z} - \sum_{j=1}^{g} \theta_j}{g - \sum_{j=1}^{g} \theta_j}. \tag{3.2}$$

Now we find the variance of the estimator $\widehat{\pi}_P$.

Consider

$$Z_i^2 = \sum_{j=1}^{g} \alpha_j^2 + \sum_{\substack{j,k=1\\ j\neq k}}^{g} \alpha_j \alpha_k.$$

After applying expectation operator, we get

$$\begin{aligned}
E(Z_i^2) &= \sum_{j=1}^{g} E(\alpha_j^2) + \sum_{\substack{j,k=1\\ j\neq k}}^{g} E(\alpha_j \alpha_k \\
&= \sum_{j=1}^{g} (\pi + \theta_j - \pi\theta_j) + \sum_{\substack{j,k=1\\ j\neq k}}^{g} (\pi + \theta_j \theta_k) \\
&= g\pi \sum_{j=1}^{g} \theta_j - \pi \sum_{j=1}^{g} \theta_j + g(g-1)\pi + \sum_{\substack{j,k=1\\ j\neq k}}^{g} \theta_j \theta_k \\
&= \left(g^2 - \sum_{j=1}^{g} \theta_j\right)\pi + \sum_{j=1}^{g} \theta_j + \sum_{\substack{j,k=1\\ j\neq k}}^{g} \theta_j \theta_k.
\end{aligned}$$

By definition, variance of Z_i is

$$\begin{aligned}
V(Z_i) &= E(Z_i^2) - (E(Z_i))^2 \\
&= \left(g^2 - \sum_{j=1}^{g} \theta_j\right)\pi + \sum_{j=1}^{g} \theta_j + \sum_{\substack{j,k=1\\ j\neq k}}^{g} \theta_j \theta_k - \left(g - \sum_{j=1}^{g} \theta_j\right)^2 \pi^2 \\
&\quad - \left(\sum_{j=1}^{g} \theta_j\right)^2 - 2\left(g - \sum_{j=1}^{g} \theta_j\right)\pi\left(\sum_{j=1}^{g} \theta_j\right) \\
&= \left(g - \sum_{j=1}^{g} \theta_j\right)^2 \pi(1-\pi) + \sum_{j=1}^{g} \theta_j \left(1 - \sum_{j=1}^{g} \theta_j\right)(1-\pi) + \sum_{\substack{j,k=1\\ j\neq k}}^{g} \theta_j \theta_k.
\end{aligned}$$

Thus the variance of the estimator $\widehat{\pi}_P$ is given by

$$V(\widehat{\pi}_P) = \frac{\pi(1-\pi)}{n} + \frac{\left(\sum_{j=1}^{g} \theta_j\right)\left(1 - \sum_{j=1}^{g} \theta_j\right)(1-\pi)}{n\left(g - \sum_{j=1}^{g} \theta_j\right)^2} + \frac{\sum_{\substack{j,k=1\\ j\neq k}}^{g} \theta_j \theta_k}{n\left(g - \sum_{j=1}^{g} \theta_j\right)^2}. \tag{3.3}$$

4. **Efficiency Comparison**

The proposed estimator would be efficient than the estimator $\widehat{\pi}_I$ if

$$V(\widehat{\pi}_I) - V(\widehat{\pi}_P) \geq 0$$

or if

$$\frac{\pi(1-\pi)}{n_2}+\frac{\sum_{j=1}^{g}\theta_j\left(1-\sum_{j=1}^{g}\theta_j\right)}{n_2}+\frac{\sum_{j=1}^{g}\theta_j\left(1-\sum_{j=1}^{g}\theta_j\right)+\sum_{\substack{j,k=1\\ j\neq k}}^{g}\theta_j\theta_k}{n_1}$$
$$-\frac{\pi(1-\pi)}{n}-\frac{\left(\sum_{j=1}^{g}\theta_j\right)\left(1-\sum_{j=1}^{g}\theta_j\right)(1-\pi)}{n\left(g-\sum_{j=1}^{g}\theta_j\right)^2}-\frac{\sum_{\substack{j,k=1\\ j\neq k}}^{g}\theta_j\theta_k}{n\left(g-\sum_{j=1}^{g}\theta_j\right)^2}\geq 0,$$

or if,

$$\left[\frac{\pi(1-\pi)n_1}{nn_2}+\frac{\sum_{j=1}^{g}\theta_j\left(1-\sum_{j=1}^{g}\theta_j\right)}{n}\left[\frac{n^2-(1-\pi)n_1n_2}{n_1n_2}\right]\right.$$
$$\left.+\frac{\sum_{\substack{j,k=1\\ j\neq k}}^{g}\theta_j\theta_k\left[n\left\{\left(g-\sum_{j=1}^{g}\theta_j\right)^2-1\right\}+n_2\right]}{nn_1}\right]\geq 0.$$

If we set $\sum_{j=1}^{g}\theta_j=1$, above inequality always holds for $g\geq 2$. Additionally, if it is possible to set $\theta_j=\frac{1}{g}$ then above inequality reduces to

$$\left[\frac{\pi(1-\pi)n_1}{nn_2}+\frac{(g-1)(g-2)}{n}\right]\geq 0, \tag{4.1}$$

which is always true for every value of g, the number of innocuous items.

4.1. *Efficiency Comparison with Warner's RR Technique*

As we have discussed that item count technique has been developed as an alternative to RR technique. We have also compared our technique with RR technique proposed by Warner (1965). To have an efficiency comparison, we first give a short description of Warner (1965) RR technique. Warner (1965) introduced this method to decrease the biased ness in the parameters and to increase the response rate. Warner's technique consists of two complimentary questions A and A^c to be answered on probability basis. Assuming a simple random sampling with replacement (SRSWR), the ith selected respondent is asked to select a question (A or A^c) and report "yes" if his/her actual status matches with selected question and "no" otherwise.

The probability of "yes" for a particular respondent is then:

$$P(\text{yes})=\theta=p\pi+(1-p)(1-\pi), \tag{4.2}$$

where p is the probability of selecting question A, and π is the population proportion of individuals with sensitive group and

$A =$ do you belong to sensitive group,

$A^c =$ do you not belong to sensitive group.

From Equation (4.2), we have

$$\pi = \frac{\theta - (1-p)}{2p-1}. \tag{4.3}$$

By the method of maximum likelihood, an unbiased estimator of π is:

$$\widehat{\pi}_W = \frac{\widehat{\theta} - (1-p)}{2p-1}, \tag{4.4}$$

where $\widehat{\theta} = \frac{n'}{n}$ and n' is the number of "yes" responses in the sample of size n.

The variance of the estimator $\widehat{\pi}_W$ is given by

$$\text{Var}(\widehat{\pi}_W) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2}. \tag{4.5}$$

Using the condition that $\theta_j = \frac{1}{g}$, the variance of the proposed estimator $\widehat{\pi}_P$ reduces to

$$V(\widehat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{1}{ng(g-1)}. \tag{4.6}$$

Comparing (4.5) and (4.6) we can see that the proposed estimator $\widehat{\pi}_P$ will be more precise than $\widehat{\pi}_W$ if

$$\text{Var}(\widehat{\pi}_W) - \text{Var}(\widehat{\pi}_P) \geq 0$$

or if

$$\frac{p(1-p)}{n(2p-1)^2} - \frac{1}{ng(g-1)} \geq 0.$$

Above inequality is true for $g > 3$. Table 1 below consists of the relative efficiency of the proposed item count estimator for $0.1 \leq \pi \leq 0.5$, $0.1 \leq p \leq 0.4$ and $g > 3$.

5. Discussion

An alternative item count technique has been presented in this article. One of the main feature of this technique is that it does not require the selection of two subsamples of sizes n_1 and n_2. Therefore, we do not need to worry about the optimum values of n_1 and n_2. Furthermore, the response from a respondent is bounded to lie between 0 and g which help providing the privacy to the respondent because the response can not be traced back to respondent's actual status about the possession of sensitive item. It has been observed that proposed item count technique estimator performs well than the usual item count technique under the condition that $\theta_j = \frac{1}{g}$. It may be difficult to select the items in such a way that their proportions in the population are same and sum to one. But this would be the case if the number of items is large. Thus in practice, one or two innocuous items with same proportions can be found and included in the item list (e.g. Item 1: were you born in the months from January to June, and Item 2: Is your gender male).

If the condition to satisfy the inequality (4.1) is hard to meet we would suggest to look for a large number of (4, 5 or 6 etc.) innocuous items such that their prevalence in the population is rare and consequently we have smaller $\sum_{j=1}^{g} \theta_j$ so that inequality (4.1) is easily satisfied.

When compared to Warner (1965) RR technique, it has been observed that proposes estimator $\widehat{\pi}_P$ is more efficient than $\widehat{\pi}_W$ for $g > 3$ under the given condition of $\theta_j = \frac{1}{g}$. We have presented the results for relative efficiency of the proposed estimator for $0.1 \leq \pi \leq 0.5$ and $0.1 \leq p \leq 0.4$, since the relative efficiency is symmetric around $p, \pi = 0.5$. It has also been observed that relative efficiency of the proposed estimator $\widehat{\pi}_P$ increases with an increase in p for a given value of g and π. The relative efficiency increases, for a given value of p if g increases. Maximum efficiency can be achieved by setting g larger if π is suspected to be larger ($\pi \geq 0.3$) for any value of p. Based on the discussion above, we recommend the use of proposed ICT in surveys about sensitive items instead of RR technique because of the problems with RR techniques discussed in the Section 1.

Table 1. Relative efficiency of the proposed estimator $\widehat{\pi}_P$ relative to $\widehat{\pi}_W$ for $0.1 \leq \pi \leq 0.5$ and $0.1 \leq p \leq 0.4$.

$\pi = 0.1$					$\pi = 0.2$			
	p				p			
g	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
4	1.68	5.28	15.59	71.24	1.67	5.28	15.74	70.66
5	2.78	8.74	25.80	117.89	2.75	8.64	25.46	116.31
6	4.13	13.00	38.36	175.89	4.07	12.76	37.61	171.80
$\pi = 0.3$					$\pi = 0.4$			
4	1.67	5.22	15.38	70.25	1.66	5.21	15.33	70.01
5	2.73	8.57	25.23	115.20	2.72	8.52	25.09	114.54
6	4.02	12.60	37.10	169.39	4.00	12.50	36.79	167.97
$\pi = 0.5$								
4	1.66	5.20	15.32	69.93				
5	2.72	8.51	25.04	114.33				
6	3.99	12.47	36.69	167.51				

References

[1] A.M. Ahart and P.R. Sackett (2004), A new method for examining relationships between individual difference measures and sensitive behavior criteria: Evaluating the unmatched count technique, *Organizational Research Methods* **7**, 101–114.

[2] P.P. Biemer and D. Wright (2004), Estimating cocaine use using the item count methodology: preliminary results form the national survey on drug use and health, paper presented at the *Annual Meeting of the American Association for Public Opinion Research*, Phoenix, AZ.

[3] A. Chaudhuri and R. Mukerjee (1988), *Randomized Response: Theory and Methods*, Marcel-Decker, New York.

[4] A. Chaudhuri and T.C. Christifides (2007), Item Count Technique in estimating the proportion of people with a sensitive feature, *Journal of Statistical Planning and Inference* **137**, 589–593.

[5] D.R. Dalton and M.B. Metzger (1992), Integrity testing for personal selection: an unsparing perspective, *Journal of Business Ethics* **12**, 147–156.

[6] D. R. Dalton, J.C. Wimbush and C.M. Daily (1994), Using the unmatched count technique (UCT) to estimate the base rates for sensitive behavior, *Personnel Psychology* **47**, 817–828.

[7] J.A. Droitcour, R.A. Caspar, M.L. Hubbard, T.L. Parsley, W. Visscher and T.M. Ezzati (1991), The Item count technique as a method of indirect questioning: a review of its development and a case study application, in *Measeurement Errors in Seuveys*, P.P. Biemer, R.M. Groves, L.E. Lyberg, N. Mathiowetz and S. Sudman (editors), Wiley, New York.

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

[9] J.A. Droitcour and E.M. Larson (2002), An innovative technique for asking sensitive questions: the three card method, *Sociological Methodological Bulletin* **75**, 5–23.

[10] M.L. Hubbard, R.A. Casper and J.T. Lessler (1989), Respondents' reactions to item count lists and randomized response, in *Proceedings of the Survey Research Section of the American Statistical Association*, Washington, D.C., 544–448.

[11] M.D. Geurts (1980), Using a randomized response design to eliminate non-response and response biases in business research, *Journal of the Academy of Marketing Science* **8** (2), 83–91.

[12] P.T. Liu and L.P. Chow (1976), A new discrete quantitative randomized response model, *Journal of the American Statistical Association* **71**, 72–73.

[13] E.R. Larkins, E.C. Hume and B.S. Garcha (1997), The validity of randomized response method in tax ethics research, *Journal of the Applied Business Research* **13** (3), 25–32.

[14] J.D. Miller (1985), The nominative technique: a new method of estimating heroin prevalence, *NIDA Research Monograph* **57**, 104–124.

[15] J.E. Reinmuth and M.D. Geurts (1975), The collection of sensitive information using a two stage randomized response model, *Journal of Marketing Research* **12**, 402–407.

[16] D. Tracy and N. Mangat (1996), Some development in randomized response sampling during the last decade – a follow up of review by Chaudhuri and Mukerjee, *Journal of Applied Statistical Science* **4**, 533–544.

[17] T. Tsuchiya, Y. Hirai and S. Ono (2007), A study of the properties of the item count technique, *Public Opinion Quarterly* **71** (2), 253–272.

[18] T. Tsuchiya (2005), Domain estimators for the item count technique, *Statistics Canada* 31(1), 41–51.

[19] J.C. Wimbush and D.R. Dalton (1997), Base rate for employee theft: Convergence of multiple methods, *Journal of Applied Psychology* **82**, 756–763.

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

Zawar Hussain, *Department of Statistics, Quaid-i-Azam University 45320, Islamabad 44000, Pakistan*.
E-mail: zhlangah@yahoo.com

Javid Shabbir, *Department of Statistics, Quaid-i-Azam University 45320, Islamabad 44000, Pakistan*.
E-mail: jsqau@yahoo.com

Received May 26, 2009
Accepted December 4, 2010