An Estimation of the Design Effect for the Two-stage Stratified Cluster Sampling Design

Tsung-Hau Jen[1], Hak-Ping Tam[2], Margaret Wu[3]

[1]Science Education Center, National Taiwan Normal University

[2]Graduate Institute of Science Education, National Taiwan Normal University

[3]Melbourne Graduate School of Education, University of Melbourne

Introduction

The precision of parameters estimation are determined by the sample size and the sampling design used in a study. Due to such practical constraints as the budget and manpower, most large-scale educational studies would not adopt the simple random sampling design. TIMSS 2007 used a two-stage stratified cluster sampling design. In the first stage, about 150 schools were selected according to some variables of interest, such as school types or locations. In the second stage, which is the clustering sampling stage, one or two classes in the sampled school were selected at random and from which all students were surveyed.

On the one hand, students in the same class are subjected to the same contextual variables at the class and school levels, so the effective sample size could be much less than the same number of students selected under the simple random sampling design. The real standard errors of the statistics being estimated could be much larger than those under the simple random sampling setting.

On the other hand, since the purpose is to measure the population mean as precise as possible, some auxiliary variables, known as implicit variables in TIMSS, will be used to divide all the clusters (schools or classes) into different strata. The probability of selecting a cluster within each stratum is proportional to the cluster size. Hence, the estimation of the population mean will be more precise as the implicit variables are more correlated with the school means. In other words, the sample should be more representative at the cluster level.

In view of the importance of the two-stage stratified cluster sampling design, the purpose of this study is to estimate the error variance by means of generalizing the standard error of the population mean under the one-stage equal size cluster sampling (Equation (1)):

$$\mathrm{SE}_{\mathrm{clustering}}(\bar{\mu}) \approx \sqrt{[1+\rho(b-1)]\frac{\mathrm{s}^2}{\mathrm{n}}}\ , \qquad (1)$$

where the $[1+\rho(b-1)]$ is defined as Kish's (1965) design effect (D_{eff}) and in which the intra-class correlation ρ is defined as

$$\rho = 1 - \frac{\mathrm{S}^2_{\mathrm{within-cluster}}}{\mathrm{S}^2}\frac{\mathrm{N}}{\mathrm{N}-1} \approx 1 - \frac{\mathrm{s}^2_{\mathrm{within-cluster}}}{\mathrm{s}^2}\frac{\mathrm{n}}{\mathrm{n}-1} \qquad (2)$$

and b is the cluster size. For unequal sized clusters, the weighted average cluster size

$$b' = \frac{\sum_{i=1}^{m} b_i^2}{\sum_{j=1}^{m} b_j}\ , \qquad (3)$$

where m is the number of the sampled clusters and b_i is the size of the cluster i, in place of b generally provides "*serviceable approximations*" (Dorofeev & Grant, 2006).

And this formula can be generalized for estimating the variance of the mean of each stratum in the two-stage stratified cluster sampling design as

$$Var(\bar{\mu}_h) \approx [1+\rho_h(b'_h-1)]\frac{s_h^2}{n_h}, \qquad (4)$$

where index h refers to the stratum h.

$$Var(\bar{\mu}) = \sum_h w_h^2\, Var(\bar{\mu}_h) \approx \sum_h w_h^2\,[1+\rho_h(b'_h-1)]\frac{s_h^2}{n_h}. \qquad (5)$$

If the following three assumptions is satisfied in a two-stage stratified cluster sampling, then Equation (4) could be substituted as

$$\begin{aligned} Var(\bar{\mu}) &\approx \frac{\sum_h N_h^2\,[1+\rho_h(b'_h-1)]\frac{s_h^2}{n_h}}{(\sum_h N_h)^2} \\ &= \frac{\sum_h N_h\,[1+\rho_h(b'_h-1)]\frac{s_h^2}{f_h}}{(N)^2} \\ &= [1+\rho'(b'-1)]\frac{s^2_{within-stratum}}{fN} \\ &= [1+\rho'(b'-1)]\frac{s^2_{within-stratum}}{n}. \end{aligned} \qquad (6)$$

Assumption 1. Equal sample fraction. For a proportionate sampling design, the sampling fraction in each stratum is equal to the overall sampling fraction. Then we have

$$f_1 = f_2 = f_3 = \cdots = f_h = \frac{n_h}{N_h} = \frac{n}{N} = f.$$

Assumption 2. Equal weighted average cluster size. If the weighted average cluster size across stratum is homogeneous, then

$$b_1' = b_2' = b_3' = \cdots = b_h' = \cdots = b'.$$

Assumption 3. Homogeneity of between and within cluster variance across strata. This assumption gives the equal intra-cluster correlation across strata. So we have

$$\rho_1 = \rho_2 = \rho_3 = \cdots = \rho_h = \cdots = \rho'.$$

Although all the three assumptions usually were not tenable in practical sampling design, but by accepting these constrains we can still make reasonable estimations.

As we have mentioned above, in practical two-stage stratified cluster sampling, the stratification of all clusters was usually based on some auxiliary information at cluster level ,which correlated to the cluster mean, to reduce the standard error of the population mean at the first stage. For one-stage cluster sampling, the total variance of the mean for population can be divided into the between-cluster component and within-cluster component (Equation (7)). For Two-stage stratified cluster sampling, a portion of the variance of cluster means would be explained by the auxiliary variable, the variance of cluster mean within each stratum would be less than the total variance of cluster mean (Equation (8)).

$$\sigma_{total}^2 = \sigma_{between-cluster}^2 + \sigma_{within-cluster}^2 \quad (7)$$

$$\sigma_{between-cluster}^2 = \sigma_{auxiliary-variable}^2 + \sigma_{within-stratum}^2 . \quad (8)$$

The portion of variance explained by auxiliary information is defined as

$$\varphi = \frac{\sigma_{auxiliary-variable}^2}{\sigma_{between-cluster}^2} . \quad (9)$$

Based on the Assumption 3, we can write down the variance of cluster mean within any one of the stratum as

$$\sigma_h^2 = (1-\varphi)\sigma_{between-cluster}^2 + \sigma_{within-cluster}^2 . \quad (10)$$

If the auxiliary variable is continuous and the values of this variable for all the clusters within each stratum are equal, then $\varphi \approx r^2$ and r is the correlation between the auxiliary variable of the cluster and the cluster mean of the surveyed statistic.

Equation (2) and Equation (10) give

$$s_{stratum\_h}^2 = \frac{n_h}{n_h - 1}\sigma_h^2 = \frac{n_h}{n_h - 1}\left[(1-\varphi)\sigma_{between-cluster}^2 + \sigma_{within-cluster}^2\right]$$

$$\approx \left[(1-\varphi)\sigma^2_{between-cluster} + \sigma^2_{within-cluster}\right]. \quad (11)$$

We can simplify Equation (2) and have

$$\rho \approx 1 - \frac{s^2_{within-cluster}}{s^2}\frac{n}{n-1}$$

$$= 1 - \frac{s^2_{within-cluster}}{\sigma^2}$$

$$\approx 1 - \frac{\sigma^2_{within-cluster}}{\sigma^2} = \frac{\sigma^2_{between-cluster}}{\sigma^2}. \quad (12)$$

In Equation (12), we use $\sigma^2_{\text{within-cluster}}$ to substitute $s^2_{\text{within-cluster}}$. In other words, we consider *(b-1)/b* as 1 and in the situations with small cluster size this approximation may not be justified. In the case of two-stage stratified cluster sampling, we combine Equation (11), (12) and Assumption 2 and have

$$\rho' = \rho_h \approx \frac{(1-\varphi)\sigma^2_{between-cluster}}{(1-\varphi)\sigma^2_{between-cluster} + \sigma^2_{within-cluster}}$$

$$= \frac{(1-\varphi)\sigma^2_{between-cluster}}{(1-\varphi)\sigma^2_{between-cluster} + \sigma^2_{within-cluster}}$$

$$= \frac{(1-\varphi)\rho\sigma^2_{total}}{(1-\varphi)\rho\sigma^2_{total} + (1-\rho)\sigma^2_{total}}$$

$$= \frac{\rho - \varphi\rho}{1-\varphi\rho}, \quad (13)$$

and rewrite Equation (6) as

$$Var(\bar{\mu}) \approx [1 + \rho'(b'-1)]\frac{s^2_{within-stratum}}{n}$$

$$= [1 + \rho'(b'-1)]\frac{\sigma^2_h}{n}$$

$$= [1 + \rho'(b'-1)]\frac{\left[(1-\varphi)\sigma^2_{between-cluster} + \sigma^2_{within-cluster}\right]}{n}$$

$$= [1 + \rho'(b'-1)]\frac{\left[(1-\varphi)\rho\sigma^2_{total} + (1-\rho)\sigma^2_{total}\right]}{n}$$

$$= \left[1 + \left(\frac{\rho - \varphi\rho}{1-\varphi\rho}\right)(b'-1)\right]\frac{\sigma^2_{total}(1-\rho\varphi)}{n}. \quad (14)$$

The standard error of the population mean estimation for two-stage stratified cluster sampling can be estimated by

$$SE_{two-stage} \approx \sqrt{\left[\left(\frac{1-\varphi}{\frac{1}{\rho}-\varphi}\right)\times(b'-1)+1\right]\times\frac{\sigma^2(1-\rho\varphi)}{n-1}} \qquad (15)$$

In both equations (1) and (2), σ^2 is the variance of total sampled students, ρ is the intra-cluster correlation, *b'* is the weighted average cluster size, *n* is the total sample size, and the value of "φ" is the ratio of the variance of cluster mean between strata to the total variance of cluster mean. When the variance of cluster mean explained by the auxiliary variable is close to zero, such as in the case of Taiwan in TIMSS 2007, Equation (15) will reduce to Kish's formula (1965) for the one-stage cluster sampling design. Moreover, if the auxiliary variable for stratifying the clusters is continuous variable we can estimate the lower-boundary for the standard error of population mean estimation by using

$$SE_{LB} \approx \sqrt{\left[\left(\frac{1-r^2}{\frac{1}{\rho}-r^2}\right)\times(b'-1)+1\right]\times\frac{\sigma^2(1-\rho r^2)}{n-1}}\ . \qquad (16)$$

In Equation (16), *r* is the correlation between the cluster means and the continuous auxiliary variable that was used for stratification at the cluster level. Notice that in Equation (16), all the clusters in each stratum have the same value under the auxiliary variable. In real cases, it is impossible except for the number of strata is infinite. The variance of the population mean should depend on the number of strata.

In this study, we used the data of 10 participant countries/regions in TIMSS 2007 to validate the Equation (15), which derived in this study provides a simple way to estimate the efficiency of the sampling framework before surveying for the two-stage stratified cluster sampling design as used in TIMSS and PIRLS.

Methodology and Data Sources

In order to verify the validity of the formula, we used Equation (2) to estimate the error variances of science achievement ($SE_{two\text{-}stage}$) for 8th graders in 10 TIMSS 2007 participating countries and compared the results with those estimated by using the jackknife replication (SE_{JKR}) technique as discussed in the international report (Martin et al., 2008).

To compute $SE_{two\text{-}stage}$, one would need to know the values of n, s^2, b', ρ and φ in Equation (2). This paper used an empirical approach to estimate these variables for the 10 countries/regions based on the data of TIMSS 2007 (IEA, 2007). Accordingly, n is the total sample size for the target population, b' is the average cluster size, ρ is the intra-cluster

correlation which is estimated as the ratio of the between-cluster variance and the total variance of students' science achievement scores, and φ is the correlation between the cluster mean and the implicit variable for stratification. More specifically, ρ can be estimated by using the HLM software or by the ANOVA procedure in SPSS, while φ can be estimated by the correlation between the means of the two clusters in the same stratum. For TIMSS 2007, the meaning of "stratum" is close to the definition of JKZONE and the units in a "cluster" refer to the students with the same JKREP variable in the same JKZONE.

Results

Table 1 presents the two standard errors ($SE_{two\text{-}stage}$ and SE_{JKR}) for eighth grade science achievement scores in seven participating countries/regions. Notice that the variables ρ and ϕ were computed by using the first set of plausible values for science achievement.

Table 1. The standard errors of science achievement estimated by the proposed method ($SE_{two\text{-}stage}$) and by the jackknifed replications (SE_{JKR}) for 10 TIMSS 2007 countries/regions

Country/Region	n	σ^2	b'	ρ	φ	$SE_{two\text{-}stage}$	SE_{JKR}
Australia	4069	6452	25.7	0.48	0.42	3.5	3.6
Cyprus	4399	7280	31.2	0.07	0.40	1.9	2.0
England	4025	7293	31.4	0.45	0.14	4.8	4.4
Hong Kong	3470	6557	29.5	0.54	0.22	4.9	4.9
Japan	4312	5946	30.2	0.19	0.58	2.1	1.9
Korea	4240	5755	30.0	0.07	0.00	2.0	2.0
Sweden	5215	6090	36.2	0.15	0.06	2.6	2.6
Taiwan	4046	7971	27.8	0.22	0.00	3.7	3.7
Ukraine	4424	7055	37.4	0.23	0.21	3.5	3.5
United States	7377	6769	55.8	0.36	0.53	3.0	2.9

n: total sample size

σ^2: the variance of

b': weighted average of cluster size

ρ: intra-cluster correlation

φ: correlation between the cluster means within stratum (JKZONE)

The results demonstrated in Table 1 indicated that the standard errors estimated by using the simple formula provided in this study are very close to those that were estimated by using the jackknife replications technique.

Conclusion

Although the replication techniques, such as jackknife replication method suggested in TIMSS and PIRLS studies and Fay's BBR in PISA, are more accurate ways to estimate the variance in complex survey design studies, they can only be processed after the data have been collecting. The most important thing is that based on this formula we could estimate an approximated standard error of the statistic in previous. The data of this study demonstrated that the estimation seems better than a "*serviceable approximation*".

References

S. Dorofeev, & P. Grant (2006). Statistics for real life sample surveys: Non-simple-random samples and weighted data. New York: Cambridge University Press.

IEA (2009). TIMSS 2007 International Database. Retrieved March 31, 2009, from TIMSS & PIRLS International Study Center, Boston College: http://timss.bc.edu/timss2007/idb_ug.html

Kish, L. (1965). *Survey sampling*. London, England: John Wiley and Sons, Inc.

Martin, M. O., Mullis, I. V. S., Foy, P., Olson, J. F., Erberber, E., Preuschoff, C., et al. (2008). *TIMSS 2007 International Science Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.