STRATIFIED INVERSE CLUSTER SAMPLING WITH UPDATING PROCESS FOR SAMPLES FROM A RARE POPULATION

By

Sewon Kim

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Measurement and Quantitative Methods—Doctor of Philosophy

2020

ABSTRACT

STRATIFIED INVERSE CLUSTER SAMPLING WITH UPDATING PROCESS FOR SAMPLES FROM A RARE POPULATION

By

Sewon Kim

Surveys have been a popular research tool and have been used extensively in many fields including education. In practice, most of surveys are conducted with some part of the population, samples. As more surveys are conducted, the range of survey participants becomes wider than ever before. Groups of people, who did not attract enough educational researchers' attention because they were rare in the general population, are now considered populations of interest. However, they are hard to sample using conventional sample designs. Such situation motivated the development of a new sample design and Reckase, Kim, and Ju (2016) developed stratified inverse cluster sampling with updating process (SICSUP) in order to obtain a representative sample from such rare populations.

The objective of this study is to evaluate the performance of SICSUP with respect to statistical and economic aspects. The statistical aspects are: (1) accuracy in parameter estimation, (2) required sample size to achieve desired precision that results of surveys should have, and (3) accuracy in group differentiation were examined. The economic aspect is the number of contacted schools in order to reach the predetermined sample size of elements in SICSUP as compared to that in stratified cluster sampling (SC) was investigated.

The results suggest that SICSUP works as well as SC and can be a useful sample design for rare populations. Also, the results provide guidelines for the application of SICSUP in educational surveys. In terms of precision in mean, standard deviation, and standard error estimation, in general, SICSUP performs as well as SC except with small sample size (n = 50).

The four replication-based standard error estimators, including the jackknife, bootstrap, BRR, and BRR with Fay's adjustment, do not make a substantial difference in standard error estimation.

In terms of determination of sample size, on average, SICSUP needs a slightly larger sample than SC although the difference in sample size between the two sample designs is not sizable. With sampling weight, SICSUP and SC require a sample size about 2.30 and 2.21 times, respectively, larger than that in simple random sampling (SRS) in order to produce estimates as accurate as those in SRS.

In terms of providing country rankings that are identical with those based on the population means, SICSUP works as well as or, depending on the condition, slightly better than SC. However, the results imply that rankings should be interpreted with caution.

With respect to economic aspect, SICSUP needs to contact fewer schools than SC in order to reach a predetermined sample size of elements and thus, is more economical than SC. However, SICSUP might not have advantages for rare populations with large clusters or small number of strata.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my advisor and committee chair, Dr. Mark D. Reckase. The meetings and conversations every week were vital in inspiring and motivating me to keep working on my dissertation. Without his amazing support and encouragements especially during this challenging time, I would not have been able to complete my dissertation.

I would also like to thank my committee members, Dr. Kimberly Kelly, Dr. Richard Houang, and Dr. Amita Chudgar, not only for their time and patience, but for insightful suggestions that substantially improved my work.

Nobody has been more important to my life and the completion of my dissertation than my family. I dedicate this work to my mom and dad, Hyunsug Hong and Kyungman Kim, and my brother, Hansol. Many thanks to my family for all the love and support you have shown me throughout my work even though we are living on two different continents.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1.

INTRODUCTION

1.1 Background

Surveys have been a popular research tool and have been used extensively in many fields including education, psychology, and sociology. In practice, most of surveys are conducted with some part of the population, samples, rather than with the whole population and make inferences about the population.

Every year, more and more surveys are conducted, and the range of survey participants becomes wider than ever before. Groups of people such as cultural minority, the homeless, and nomads once seemed impossible to survey because they are rare in the general population, but now they are considered target populations for surveys although they are harder to survey than the general population.

In the field of educational research, surveys are also popular and have been widely used to increase knowledge in the field. Like other areas, surveys have been done more frequently in recent years, and rare, thus, hard-to-sample populations, such as students from minority group (De Róiste & Dinneen, 2005) and children experiencing long-term foster care (Daly & Gilligan, 2005),have gained increasing attention from educational researchers.

Since the International Association for the Evaluation of Educational Achievement (IEA) conducted the First International Mathematics Study (FIMS) in the early 1960s, which is one of the earliest modern-day international assessments of student skills (Rutkowski, von Davier, & Rutkowski, 2013), international large-scale surveys and assessments in education also have gained importance and popularity among educational researchers and have become one of the most influential studies in current education (Kirsch et al., 2013). More than 50% of the

countries in the world have taken part in some type of international assessment (Kamens & NcNeely, 2010). The results from international surveys do not only provide international comparison but also impact on education policies at the national level (Smith, 2016).

Countries as a whole have characteristics, and different characteristics across countries could make populations of interest difficult to survey at the country level in addition to at the individual level. For example, developing countries tend to lack the resources to carry out surveys and hence, likely to have more hard-to-survey populations. These countries might not have enough funding for surveys so that they might omit specific subpopulations (e.g., people in rural area and people who speak a minor language) or shorten the period for survey. They might not have statistics collected by government censuses and statistical agencies so that general sample frames that are often required for sampling procedure cannot be constructed readily. In addition to the individual level factors (rare by nature), factors at the country level (rare by operation) could increase difficulty in sampling rare populations. With respect to international surveys in education, different educational systems across countries sometimes raise challenges for obtaining samples.

One of the most widely used sample designs for surveys in the area of educational research is cluster sampling because of the hierarchical structure of education systems. Students, the major target population in educational research, are nested in classes, and classes are nested in schools. Stratification variables are also often used to improve the representativeness of the target population. Therefore, stratified cluster sampling is a frequently used sampling technique for educational surveys. As well as domestic educational surveys, international large-scale studies, such as the OECD's Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the Progress in

International Reading Literacy Study (PIRLS), use stratified multi-stage sample designs, which are a complex form of stratified cluster sampling. In order to apply multi-stage sampling to surveys, known proportions or frequencies of sampling units over strata are required. If sampling units are students in schools, researchers should have a list (a frame) of students in the target population before starting the sampling procedure.

Applying these sample designs is inconvenient when target populations are rare populations at the country level, individual level, or both. For instance, at the individual level, a significant proportion of clusters might not include any sampling unit, which can be considered a rare population by nature, and at the country level, a country might not know the distribution of sampling units due to the limited resource, which can be considered a rare population by operation. If the target population is a rare population at the individual level and also at the country level, obtaining samples for the survey would become much harder.

Such problem motivated the development of a new sample design, and Reckase, Kim, and Ju (2016) developed stratified inverse cluster sampling with updating process (SICSUP)[1] in order to obtain a representative sample under these circumstances.

1.2 Stratified Inverse Cluster Sampling with Updating Process (SICSUP)

Before describing the SICSUP procedure, I need to define a few key terms and describe situations that this dissertation focuses on. In this section, only the terms that are necessary for describing SICSUP were mentioned. More concepts and definitions of sampling are discussed in detail in Chapter 2.

In this dissertation, samples are stratified and clustered. Clusters are taken first and elements in the clusters are taken later. The term “primary sampling unit” (PSU) refers to sampling units that are selected first, which are clusters. The term “secondary sampling unit”

[1] It was initially called stratified sequential adaptive cluster sampling (SSACS).

(SSU) refers to sampling units in PSUs, which are elements. It is assumed that researchers need to select a set of stratified and clustered samples from a rare population and they don't know much about the proportions of SSUs over strata in the population. It is also assumed that researchers already have a list of PSUs although they don't have a list of SSUs.

As shown Figure 1.1, SICSUP can be implemented in a four-step procedure (Reckase, Kim, & Ju, 2016).The first step of SICSUP is to determine initial sample sizes for strata based on available information about the proportions of SSUs over strata in the population.

The second step of SICSUP is contacting PSUs from the list of PSUs which is randomly ordered and identifying SSUs available from each PSU contacted. If there are any SSUs available, researchers recruit all of them and include them to the sample. If there is no SSU available, researchers move on to the next PSU in the list of PSUs. Researchers repeat the second step until one of the strata reaches the initial sample size.

The third step of SICSUP is updating the initial proportions of elements over strata. When one of the strata reaches the initial sample size, it becomes possible to update the initial proportions of SSUs over strata, which might not be accurate, based on the current sample proportions over strata at this point. Then, the updated sample sizes for strata would be obtained.

The fourth step of SICSUP is contacting PSUs and recruiting SSUs from the PSUs contacted until all of the strata reach the updated sample sizes. The fourth step is basically the same as the second step. The difference between the fourth step and the second step is whether updated sample sizes are used or the initial sample sizes are used. Once a stratum satisfies the desired number of SSUs, PSUs in that stratum would be ignored when contacting next PSUs in the list. After the fourth step, a final set of samples would be obtained.

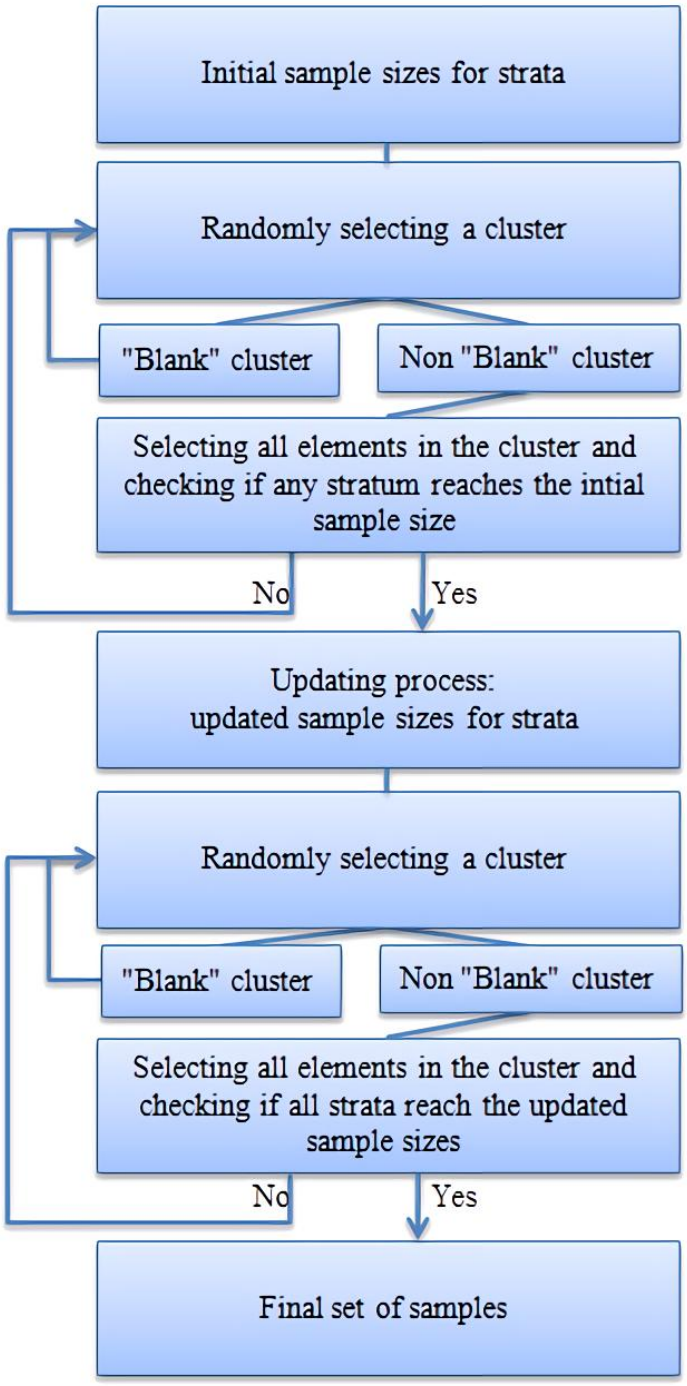


Figure 1.1 Procedure of SICSUP

1.3 Research Questions

Although previous studies evaluated the performance of SICSUP for rare populations, they are done before SICSUP was fully developed (Kim, Ju, & Reckase, 2015) or did not evaluate the entire procedure of SICSUP (Reckase, Kim, & Ju, 2016), excluding the updating process of SICSUP and stratification. Therefore, it is difficult to determine whether SICSUP is a viable sample design for rare populations in education. There is a need for evaluating the full SICSUP procedure. The results would provide guidelines and requirements for the application of SICSUP to educational surveys, enabling researchers to explore rare populations in education.

A good sample design must provide necessary information with maximum precision for fixed allowed resources. The objective of this study is to evaluate the performance of SICSUP with respect to statistical and economic aspects. In terms of statistical aspect, accuracy in parameter estimation, required sample size to achieve desired precision that results of surveys should have, and accuracy in group differentiation are examined. In terms of economic aspect, the number of contacted schools during the sampling procedure in SICSUP as compared to those in stratified cluster sampling is investigated.

The following research questions need to be addressed:

1. Does SICSUP work as well as stratified cluster sampling regarding parameter estimation?
2. How can the appropriate sample size for SICSUP be determined?
3. Can the samples from SICSUP determine whether the means of groups are different from each other?
4. Is SICSUP economically more advantageous than stratified cluster sampling?

The four research questions are answered through simulation studies. Chapter 2 reviews the concept of rare populations, the relationship between SICSUP and existing sampling techniques, and the features of standard error estimators (e.g., replication methods). Chapter 3 presents details about data generation, simulation designs, and evaluation criteria. The last two chapters (Chapter 4 and 5) provide the findings from the simulation studies and discuss the performance of SICSUP and how SICSUP can be used for surveys that aim at rare populations in education. These two chapters also provide guidelines for applying SICSUP to rare populations in education.

CHAPTER 2.

LITERATURE REVIEW

This literature review chapter consists of three main sections. The first section briefly discusses the basic concept and definitions that are used in this dissertation. The second section explores the types of rare populations and sampling techniques useful for these populations. This section also explores the relationship between SICSUP and other sampling techniques. The third section summarizes the features of replication methods for standard error estimation.

2.1 Concepts and Definitions

Different studies on sampling might use different statistical terms to describe the same concept. In order to avoid confusion due to various statistical terms, this section discusses the basic concepts and definitions that are used in this dissertation. The concepts and definitions are based on the three sampling textbooks (Kish, 1965; Murthy, 1967; Thompson, 2002).

A sampling unit, or simply a unit, is an element or a group of elements, on which observations can be made and for which information is sought. A population is a collection of all units in a given region. In element sampling (e.g., simple random sampling), each sampling unit contains only one element; but in cluster sampling, any sampling unit (or primary sampling unit (PSU)) called a cluster may contain several elements.

In general, for using sample designs, a list, or a frame, of all sampling units belong to the population is necessary, and such list or frame is termed the sampling frame. The sampling frame illustrates the distribution of elements over the population.

Surveys aim at estimating population values, which are obtained from all population elements. A population value is called a parameter. A sample value, or statistic, is an estimate computed from elements in a set of samples. In this dissertation, “sample mean” denotes the

sample estimate of the population mean, and "sample standard deviation" denotes the sample estimate of the population standard deviation. The sampling distribution of an estimate is the theoretical distribution of all possible values of the estimate. The standard deviation of the sampling distribution is called the standard error. The squared standard error is called the variance of estimate.

With respect to symbols, capital letters refer to population values, parameters, and lowercase letters denote corresponding sample values, estimates. A bar ($\bar{}$) over a symbol denotes a mean value, and a hat ($\hat{}$) over a symbol denotes an estimate. In general, this dissertation uses n for the number of elements in the sample and N for the number of elements in the population. However, for stratified cluster sampling (SC), I may use different symbols: m for the number of elements in the sample and M for the number of elements in the population; n for the number of clusters in the sample and N for the number of clusters in the population. A value of the variable of interest in a sample is expressed as y. The symbol of θ denotes any parameter such as mean or standard deviation.

For the purpose of comparison, this dissertation uses SC. It is slightly different from the stratified multi-stage sampling, which is widely used for national and international studies. SC can be considered stratified single-stage sampling. In each stratum, clusters are randomly sampled, and all elements in the clusters are selected. In stratified multi-stage sampling, different sampling techniques can be applied to each stage. For example, for two-stage sampling, primary units can be selected with probabilities proportional to size, and secondary units can be selected using simple random sampling.

2.2 SICSUP and Conventional Sampling techniques

2.2.1 Concept of Rare Population

The development of SICSUP was motivated by facing difficulties in obtaining samples from rare populations. One may question what rare populations refer to. A rare population sometimes is defined as a population with a low number of elements. However, there is no universally accepted definition of "rare" population. Terms such as "elusive" and "hard-to-detect" populations are also used for rare populations (Kish, 1991).

McDonald (2004) reviewed definitions of a rare population in the field of biology. Rare populations in biology possess one or more following characteristics: first, the proportion of the elements in the population is small; second, elements practice elusive or secretive behavior; third, elements are sparsely distributed over large ranges; fourth, elements practice differently by time or season; fifth, application of ineffective sampling can make rare populations. Based on the four characteristics, there are two types of rare populations: rare populations by nature and operationally rare populations.

Likewise, Riniolo (1999) discussed rare populations when sampling units are individuals and categorized rare population into five: (1) sparse populations, (2) limited access populations, (3) persons experiencing an infrequent event (e.g., persons with severe allergic reaction), (4) those who newly associated with a rare population (e.g., persons with brain injury), and (5) developmentally uncommon cases (e.g., teenage myocardial infarction patients).

Tourangeau (Tourangeau et al., 2014) discussed hard-to-survey populations mainly in the fields of psychology, sociology, and business. Some populations are hard to survey in different ways. The author distinguished hard-to-survey populations into five categories: populations that are hard to sample, those whose members who are hard to identify, those that are hard to find or

contact, those whose members are hard to persuade to take part, and those whose members are hard to interview. A hard-to-sample population is a population without a sampling frame or with an incomplete sample frame. In the absence of a complete sampling frame, if elements are rare, representing a small fraction of the larger population, the population can be hard to sample. The other factor making rare populations to hard-to-sample populations is the cost of screening. Screening is often used to detect rare elements in the population (for example, a few questions to identify elements in the larger population). If screening is expensive relative to main survey, it affects the final data collection from the main survey. Hard-to-sample populations also contain elusive or mobile populations, such as the homeless and migrant workers. In sum, a rare population is defined as a population with a small proportion of elements in the larger population and is a part of hard-to-sample populations and hard-to-survey populations (see Tourangeau et al., 2014, for the other four categories of hard-to-survey populations).

What kinds of rare populations are there in the field of educational research? Students with special educational needs are an example of rare populations in education, including deaf and hard of hearing students (Scott & Hoffmeister, 2016). The U.S. Census Bureau annually conducts nationwide survey known as the Survey of Income and Program Participation (SIPP) in order to identifying the American population of persons with hearing loss or deafness including children (Mitchell, 2006). Such data provide useful information about this rare population.

As the United States population becomes increasingly diverse, there has been growing interest in immigrant students (Bailey & Weininger, 2002) and bilingual students (Burke, Morita-Mullaney, & Singh, 2016; Lesaux & Kieffer, 2010). This increase in US bilingual populations also led federal authority to conduct empirical research on bilingual students (Greenberg Motamedi, Singh, & Thompson, 2016; Haas et al., 2015).

Drop-out students are another example of rare populations in educational research (Kinnunen & Malmi, 2006; Lassibille & Navarro Gómez, 2008). In general, because they already left schools, researchers experience difficulties in finding them. Information before their drop-out and indications of drop-out are often used for studies.

There are rare populations including teachers as elements. Novice teachers or beginning teachers are a rare population because of the low frequency in the larger population (population of teachers) and their mobility. Various research methods have been applied in order to study such population of novice teachers (Chubbuck et al., 2001; Westerman, 1991).

Schools can also be a rare population. Lee, Ready, and Johnson (2001) investigated "schools-within-schools", which is a type of school reform strategies for U.S. public secondary school, and in their study, such schools are rare elements in the general population (population of schools).

2.2.2 Relationship between SICSUP and Existing Sample Designs

A wide variety of techniques has been suggested for dealing with samples from a rare population such as multipurpose samples, cumulation of a rare population, use of large clusters, controlled selection, batch testing, two-phase sampling, etc. Among these techniques, Kish (1985) and Elliott (National Academies of Sciences, Engineering, and Medicine, 2018) suggested three techniques: (1) creating a list (a frame) of elements, (2) oversampling, and (3) screening.

Rare populations often lack a list of elements. Creating a list of elements could help researchers locate rare elements. Network sample designs can be used to build up a list of elements. In network sampling, a simple random or stratified random sample is selected, and all the elements linked to the previously selected sample are used to create a list of elements

(Thompson, 2002). For example, if researchers want to create a list of novice teachers with less than five years of teaching experience, they take an initial sample of novice teachers. Then, the sampled novice teachers are asked whether they know any novice teachers. If they know any novice teachers, the researchers add them to the list of novice teachers.

If it is hard to create a list of elements, oversampling is another approach for sampling rare populations. In SC, if most of the rare elements are located in a small stratum, the more samples would be selected from that stratum than from other strata. The last suggested technique is screening. Screening may involve a brief interview or short tests to identify rare elements. This technique can associate with different sample designs. Screening to find rare elements is practical if proportion of the elements is about 10 or 20 percent of the population.

Adaptive Sampling. Selection procedures in conventional sample designs, such as SRS, stratified sampling, and cluster sampling, do not depend on observations made during sampling. However, in some sampling situations, making decisions during sampling process may be beneficial in order to obtain a set of samples that provides more precise estimates than conventional sample designs given sample size or cost. For rare populations, researchers often do not have a complete frame of sampling units before starting the sampling procedure. They can take advantage of the knowledge in population characteristics that was obtained during the sampling process and improve accuracy in estimation.

Adaptive sampling refers to sample designs in which the selection procedure may depend on values of the variables of interest observed during the sampling process (Thompson, 2002). Therefore, in general, the sample size tends to vary. Adaptive sampling is a general sampling strategy rather than a specific sample design. Sample designs that employ adaptive strategy can be consider adaptive sampling, such as adaptive cluster sampling and sequential sampling.

The family of adaptive sampling tends to estimate population density or abundance and thus, has been often studied and used in biology (Thompson, 2004). Surveys in educational research pay more attention to estimating variables of interest than estimating population density. SICSUP does not specially focus on estimating abundance of elements and can be used for estimating both of variables of interest and population density. Therefore, SICSUP is more applicable than adaptive sampling for educational research.

Adaptive Cluster Sampling. Adaptive cluster sampling, introduced by Thompson (1990), is a sample design that uses adaptive strategy for selection procedure. Thompson (2002) describes adaptive cluster sampling as follows:

"Adaptive cluster sampling refers to designs in which an initial set of units is selected by some probability sampling procedure, and whenever the variable of interest of a selected unit satisfies a given criterion, additional units in the neighborhood of that unit are added to the sample" (p.319).

Adaptive cluster sampling has a large number of possible designs, and various adaptive cluster sample designs have been developed based on the basic adaptive cluster sampling: systematic and strip adaptive cluster sampling (Thompson, 1991a), stratified adaptive cluster sampling (Thompson, 1991b), two-stage adaptive cluster sampling (Salehi & Seber, 1997), restricted adaptive cluster sampling (Lo, Giffith, & Hunter, 1997), etc. Although they are different in terms of selection process or stopping rules, the basic concept of these designs is selecting additional units in the initial unit's neighborhood. Also, the total number of units in the final sample is adaptive. The collection of units is called "network" in adaptive cluster sampling.

Adaptive cluster sampling is advantageous when elements are highly aggregated or clustered. However, that might not be the case for rare populations in education. For example,

consider one wants to sample students with special educational needs. If one school contains such students, it does not mean that the neighboring schools also tend to include students with special educational needs. Therefore, in general, adaptive cluster sampling may not be very beneficial for rare populations in education.

Sequential Sampling. Sequential analysis or sequential estimation is a method for testing statistical hypotheses in which the number of observations is not fixed in advance but depended on the observations themselves (Wald, 1945). In sequential analysis, every time a sampling unit is added to the set of samples, hypothesis testing is conducted. This sequential selection procedure continues until the hypothesis testing produces a significant result. A merit of the sequential method, as applied to testing statistical hypotheses, is that, on average, the test procedure requires a substantially smaller number of observations than equally reliable test procedures based on a predetermined number of observations (Wald, 1947). Sequential probability ratio test was developed for the purpose of testing statistical hypotheses.

After sequential estimation was introduced, sample designs using the sequential estimation method have been developed. Haldane's inverse sampling (1945) is one of them although it is not under the label "sequential" (Anscombe, 1953). Because there is no sequential sample design that has been widely studied or used in statistics, the name of "sequential sampling" is used for different types of sample designs in different fields.

In general, sequential sampling is a type of adaptive sampling and uses sequential estimation. At each observation in the sampling process, the decision to continue depends on the data recorded to that point. Data collection continues according to the initial design until the stopping rule is satisfied. Sequential sampling can be applied with SRS, stratification, or clusters (Christman, 2004).

Inverse Sampling. Inverse (binomial) sampling uses adaptive strategy where the sample size is adaptive in that it depends on the information that is obtained during the sampling process. Inverse binomial sampling was introduced to select a set of samples from a rare population (Haldane, 1945). Under conventional sample designs with a fixed sample size, one may not be able to observe enough number of rare events to produce precise estimates. Inverse sampling was developed to estimate the frequency of a rare event. Researchers keep selecting sampling units until certain specified conditions are satisfied (Seber & Salehi, 2012).

Because Haldane's inverse sampling uses the sequential estimation method, some scholars use "inverse sampling" and "sequential sampling" interchangeably (Christman, 2004; Pathak, 1976). The major difference between sequential sampling and inverse sampling is that, in general, inverse sampling focuses on estimating parameters such as total and mean while sequential sampling focuses on testing hypotheses.

The similarity among SICSUP, sequential, and inverse sampling is to take samples sequentially and make decisions during the sampling procedure based on the information collected to that point. SICSUP is different from sequential and inverse sampling in terms of what kind of decision to be made. SICSUP makes decisions to adjust sample sizes for strata while sequential and inverse sampling make decisions to determine a stopping point of selection.

SICSUP is a combination of several sampling strategies: stratification, clustering, sequential estimation, and updating process. Without stratification and clustering, SICSUP is similar to Haldane's inverse sampling. The sampling procedure is continued sequentially until certain specified conditions are satisfied. In SICSUP, at each point of selection, a decision, whether any stratum reaches the predetermined sample size (initial sample size or updated sample size), is made and the decision affects the later selection.

Sequential estimation and updating process are required because of stratification in SICSUP. At the beginning of the sampling procedure, the initial sample size for each stratum might not be proportional to the size of the stratum because of lack of information on the proportions of elements over strata. When an additional sample is added to the existing set of samples, researchers check whether there is a stratum that achieved the initial sample size. This process is related to inverse sampling and sequential estimation. If a stratum satisfies the required number of samples, sample sizes for strata are updated using the sampling distribution over strata that was obtained during the sampling process. This updating process is the unique characteristic of SICSUP as compared to different sample designs.

2.2.3 Replication Method for Variance Estimation

Complex sample designs often involve features such as stratification, multiple stage sampling, and unequal selection probabilities (Wolter, 1985). Regarding Wolter's description of complex sample designs, SICSUP as well as SC can be considered a complex sample design. For such a complex sample design, unlike a simple sample design, special procedures are needed to estimate an unbiased or consistent sampling variance of an estimate of a parameter.

There are two procedures to deal with those situations: the Taylor series linearization method and replication (or resampling) methods (Rutkowski, von Davier, & Rutkowski, 2013). In recent large-scale surveys, including educational large-scale surveys and assessments, replication methods have tended to be used more frequently than the Taylor series linearization method for estimating sampling variance. The major reason for the popularity of replication methods is that the Taylor linearization method is, in general, mathematically complicated and, therefore, require significant computation burden as compared to replication methods.

The idea of subsample replication methods was introduced to simplify variance estimation for complex sample surveys (Wolter, 1985). In terms of sample variance of means, the family of replication methods consists in selecting multiple samples from the parent sample; computing a separate estimate of mean from each sample; and computing the sample variance among the several estimates. The jackknife method and balanced repeated replication (BRR) method are commonly used replication methods along with the bootstrap method.

The Jackknife Method. Since 1940's, various kinds of replication methods have been developed. The jackknife method, which was introduced by Quenouille (1949), is one of the most frequently used replication methods. Replicated datasets are typically created by dropping secondary units from one PSU at time to form a replicate until all PSUs have been dropped from each stratum (Skinner et al., 1989).

In general, the procedure of the jackknife method is as follows (Lee, Lee, & Shin, 2016; Wolter, 1985). First, the parent sample is divided into *K* random groups where *K* represents the number of PSUs. Second, all secondary units in the parent sample possess the variable of interest, and the parameter of the variable is θ. An estimate of θ based on the parent sample denotes $\hat{\theta}$. Third, after deleting the K^{th} group, the weights of the remaining secondary units are doubled. With these replicate weights, $\hat{\theta}_{k(i)}$ is calculated using the elements in the remaining groups. Finally, the jackknife estimator of variance is then

$$\sigma_{Jack}\left(\hat{\theta}\right) = \frac{(K-1)}{K}\sum_{i=1}^{K}\left(\hat{\theta}_{k(i)} - \hat{\theta}\right)^2. \quad (2.1)$$

There are some differences in the jackknife procedures when they are applied to a stratified cluster sample design (Chen & Shen, 2019; Smith, Srinath, & Battaglia, 2000). In a stratified cluster sample design, the jackknife procedure is basically identical except that occurs in each stratum. First, in each stratum *h*, there are *K* clusters (or PSUs). After deleting K^{th} cluster,

the weights of the remaining elements in stratum *h* would be doubled to compensate for the deleted cluster and used to compute a variance estimate. Second, $\hat{\theta}_{h(k)}$ is calculated, and there would be K_h estimates of $\hat{\theta}_{h(k_h)}$ for stratum *h*. Third, the jackknife estimator of variance is

$$\sigma_{Jack}\left(\hat{\theta}\right) = \sum_{h=1}^{L} \frac{(K_h - 1)}{K_h} \sum_{i=1}^{K_h} \left(\hat{\theta}_{h(i)} - \hat{\theta}\right)^2. \quad (2.2)$$

A variety of variance estimators based on the jackknife method has been developed. The jackknife repeated replication (JRR) method was developed by Frankel (1971) who first applied jackknife procedure to compute sampling variance in complex surveys. The JRR was developed based on the jackknife estimation procedure and the BRR method. With the BRR method, each of the replications estimates the variance of the entire sample while, with the JRR method, each replication estimates the variance contributed by a single stratum (Kish & Frankel, 1974). The TIMSS and the PIRLS currently use the JRR method to estimate sampling variance.

The major advantages of using the jackknife method are that it is conceptually simple and provides a precise estimate of sampling variance in general. As compared to the bootstrap method, it is less computationally intensive. The jackknife method has a limitation when it is applied to single-stage sample designs. In these sample designs, estimates of sampling variance of non-smooth statistics, such as median or quantiles, are tend to be unstable. Although this problem does not occur when multi-stage sample designs are used, it is advised to avoid using the jackknife method for estimating sampling variance of median or quantiles (Betti, Gagliardi, & Verma, 2018; Rutkowski, von Davier, & Rutkowski, 2013).

The Bootstrap Method. Bootstrapping, which was introduced by Efron (1979), is a technique that relies on random resampling with replacement, and the bootstrap method in statistics is designed to provide information about the population distribution using bootstrapping. The bootstrap is used in practice for a variety of purposes: estimating statistics on

a population (e.g., mean and standard deviation); estimating variance of a statistical estimator; and constructing approximate confidence intervals for parameters of interest (Shalizi, 2016). In this dissertation, the bootstrap method refers to the method to estimate sampling variance of means.

The bootstrap method procedure is as follows: first, a resample is drawn from the parent sample, and a statistic (e.g., mean) is computed; second, after repeating the previous step *B* times, *B* sets of the statistic, $\hat{\theta}_{(B)}$, would be obtained; third, the bootstrap variance is calculated:

$$\sigma_{Bootstrap}\left(\hat{\theta}\right) = \frac{1}{B}\sum_{i=1}^{B}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2, \tag{2.3}$$

where $\hat{\theta}$ is the estimate based on the parent sample.

In SC, $n-1$ sampling units out of the n elements are selected independently with replacement within each stratum. Because the selection is with replacement, a sampling unit may be chosen more than one (Statistics Canada, 2018).

Given sample size of *n*, there are n^n possible sets of samples with replacement. Calculating a statistic (e.g., mean) from all n^n bootstrap samples is basically impossible in practice, thus, researchers choose a number of bootstrap samples that they use to estimate sample variance of the statistic.

The bootstrap variance involves two sources of error: an error due to the fact that the sample size is finite and an error due to the fact that B is less than n^n. The first source of error can be correct by multiplying it by $(n-1)/n$. The second source of error can be reduced by increasing the number of *B*. Previous studies have suggested numbers of replications in order to obtain a reliable estimate using the bootstrap method. Although a minimum number of 200 to 300 for variance estimation was suggested (Efron & Tibshirani, 1993; Hall, 1989), larger *B* would be preferred to obtain a reliable estimate.

As compared to the jackknife method, estimated standard errors using the bootstrap method tend to be slight smaller (Efron, 1982). While the jackknife method provides unstable estimates of sampling variance of non-smooth statistics such as median and quantiles, the bootstrap method is generally work well for these statistics (Ghosh et al., 1984; Riniolo, 1999). The bootstrap method requires less computational burden as compared to the jackknife method (Chen & Shen, 2019). The bootstrap method does not work well for the following situations: correlated data (e.g., time series data), missing data, and data with outliers.

Balanced Repeated Replication and Fay's Adjustment. The balanced repeated replication method (BRR) involves dropping all elements within a PSU in a stratum, but it does so by creating half-samples. One PSU from each stratum is selected and its elements are retained, forming a pseudo-replicate, with the set of remaining PSUs for each stratum forming the complement replicate (Stapleton, 2008). The principle of the BRR is the following: each of the two PSUs can provide an unbiased estimate of the parameter of interest of its stratum.

The BRR design assumes that a population of PSUs is able to be grouped into H strata with two PSUs per stratum. The BRR can thus only be accomplished when the sample design has been undertaken with the selection of two PSUs from each stratum. In practice, it is hard to find such populations. If the sample design did not include the selection of two PSUs from each stratum, similar strata or PSUs can be artificially grouped to obtain such a design (pseudo-strata). This process of allotting each pair of PSUs into pseudo- and complement replicates is repeated many times to create a large set of half-replicates.

There is a complication in creating replicates using half of the PSUs because dependent replicates can produce parameter estimates that are correlated across replicates. In order to obtain a balanced design, a solution is to balance the formation of replicates by using an orthogonal

design matrix. A selection of these matrices, sometimes referred to as Hadamard matrices, are developed and available from Wolter (1985). The BRR provides a way to extract from the complete set of 2^H possible replicates a much smaller subset that gives the very same measure of sampling error as the full set would.

Using these matrices, a minimal set of *K* balanced half-samples are created. In order to obtain a fully balanced design, the number of replicates used needs to be four times greater than the number of strata (Chen et al., 2007).

For each of the retained PSUs as defined by the design matrix, the sampling weight is doubled to create a set of replicate weights from which to calculate replicate estimates. For any given replicate, two times of the sampling weight if the PSU in stratum is retained in the pseudo-replicate, and the weight is equal to zero otherwise (Rust & Rao, 1996).

Once these sets of replicate weights are created, a conventional analysis is run for each set of weights, and the standard errors of the parameter estimates are a measure of the variability across pseudo-replicates

$$\sigma_{BRR}\left(\hat{\theta}\right) = \frac{1}{K}\sum_{i=1}^{K}\left(\hat{\theta}_{(i)} - \bar{\hat{\theta}}\right)^2, \tag{2.4}$$

where $\hat{\theta}$ is the estimated variance using the parent sample, $\hat{\theta}_{(k)}$ is the estimated variance based on the K^{th} replicates, and *K* is the total number of half-sample replicates. .

With larger datasets, the BRR estimates of variance are seen by some as less computationally taxing than JRR because they use only half-samples (Rao, Wu, & Yue, 1992; Rust & Rao, 1996). The replication methods work differently depending on the variable of interest. For ratio estimates, the jackknife is superior to the BRR or bootstrap (Rao & Wu, 1985) while for medians, the BRR works better than the jackknife (Kovar, Rao, & Wu, 1988). This issue in using ratio estimator for the BRR motivated the development of Fay's method (Dippo,

Fay, & Morgansein, 1984). When ratio estimator is used, the BRR might produce extremely large estimates because of zero weighted and double weighted samples. Fay's idea was to use the weights of 0.5 and 1.5 instead of 0 and 2 for the half samples within each stratum. Judkins' study (1990) supports the Fay's method is a reasonable compromise between the BRR and the jackknife for the ratio and the regression coefficient.

CHAPTER 3.

METHODS

This dissertation evaluates stratified inverse cluster sampling with updating process (SICSUP) through four research questions (See Section 1.3). The first to third research questions evaluate SICSUP with respect to statistical aspects and the last research question evaluates SICSUP with respect to economic aspects. This chapter describes details about the research methods for answering to the four research questions.

In general, for each research question, the results from SICSUP are compared to the results obtained from simple random sampling (SRS), stratified cluster sampling (SC), and SICSUP without updating process (SICS). Results of SRS provide a basis for comparison. Results of SC are necessary because SICSUP also uses cluster and stratification. Results of SICS are also necessary in order to examine the effect of updating process. The comparison of SICS and SC describes the effect of sequential process.

SRS selects n distinct units from the N units in the population with the equal selection probability for each unit. SC randomly selects n clusters from the N clusters in each stratum in the population and samples all the units in the n clusters. The procedure of SICS is the same as the procedure of SICSUP except the updating process.

In this dissertation, the population of novice teachers, who are defined as teachers with zero to five years of overall teaching experience, serves as the rare population, which is also a hard-to-survey population. The population of all teachers including novice and non-novice teachers is called here the general population. Novice teachers are rare and hence, hard to sample because of two possible reasons. First, in general, mobility or turn-over rate of novice teachers is higher than veteran teachers (Simon & Johnson, 2015; Smith & Ingersoll, 2004). Even though

researchers have a frame of novice teachers and know where to find them, they may fail to find them in the schools where they are supposed to be due to high mobility. Second, in case of international surveys, some countries may not have data that contain each teacher's years of teaching experience. For example, in the United States, it may be hard to identify years of teaching experience for each teacher whose previous school is in a different state.

Although novice teachers are a rare and hard-to-sample population, at least we know they are in schools, meaning within clusters. Also, previous studies on the general population support the usage of stratification at the school level, such as school type, location of school, and source of funding (OECD, 2017, 2019). Therefore, the rare population of novice teachers with stratification and clustering would be an appropriate population for the application of SICSUP.

Data generation and analysis were done by using MATLAB R2015b (The MathWorks, INC., 1984-2015) and R software (R Core Team, 2019).

3.1 Research Question 1

The first research question is about whether SICSUP works as well as SC in terms of parameter estimation. Simulations were conducted to examine the performance of SICSUP under the various conditions. It was assumed that when sample size is not small, the updating process would be beneficial to estimate parameters including mean, standard deviation, and standard error of the mean, and hence, SICSUP would work at least as well as SC with respect to precision in parameter estimation.

3.1.1 Data Generation

A data set was generated for simulations, and in order to generate a realistic population as possible, the Teaching and Learning International Survey (TALIS) 2018 (OECD, 2019) was used as a basis for generating parameters. The TALIS surveys teachers and school leaders across

countries about working conditions and learning environment at their schools, and the questionnaire for teachers includes a question about their years of teaching experience. It is available to determine whether a participant of the TALIS is a novice teacher or not. The distribution of novice teachers in schools by strata was used to generate data for simulations. Specifically, the distribution of novice teachers in Canada was used because, with respect to the TALIS 2018, Canada is one of the countries that have high proportions of schools with zero novice teacher, indicating a rare population.

Location of school was used for stratification. The TALIS2018 categorizes schools into three locales: rural, town, and city. In the Canada data, about 11% of novice teachers are in rural schools, about 25% of novice teachers are in town schools, and about 64% of novice teachers are in city schools. The three locales serve as stratification.

The variable of teacher's self-efficacy in instruction serves as the variable of interest for simulations. This variable was chosen because, in the TALIS2018, novice teachers in the same stratum (rural, town, or city) behaved similarly to each other. Stratification, which is one of the common sampling techniques used in SICSUP, SICS, and SC, is useful and provides precise estimates of a parameter when values of the variable of interest within each stratum are homogeneous. The selected variable, teacher's self-efficacy in instruction, is an appropriate variable for employing stratification.

A dataset with 2,000 novice teachers in 949 schools was generated. In the data, about 19% of the schools have no novice teacher, and about 17% of the schools have only one novice teacher. In the TALIS2018, about 7% of schools have more than five novice teachers. To avoid too complicated data, an adjustment was applied, so the number of novice teachers in school ranged from zero to five. Additionally, because the TALIS2018 Canada has no school with more

than three novice teachers in rural area, I added schools with four or five novice teachers to the generated data. Their proportions are very small, 2.4% and 1.1%, respectively. The generated data match the TALIS2018 Canada with respect to proportions of novice teachers over strata.

Table 3.1 Number of Novice Teachers per School

Novice teachers in school	Rural		Town		City	
	N	%	N	%	N	%
0	38	23.2	59	20.7	83	16.6
1	52	31.7	55	19.3	59	11.8
2	47	28.7	99	34.7	111	22.2
3	21	12.8	42	14.7	85	17.0
4	4	2.4	19	6.7	91	18.2
5	2	1.2	11	3.9	71	14.2
Total	164	100	285	100	500	100

Table 3.2 Number of Novice Teachers by Location of School

Area	N	%
Rural	235	11.8
Town	510	25.5
City	1,255	62.8
Total	2,000	100.0

Based on these proportions of novice teachers over strata, three sets of the variable of interest were created: first, uncorrelated data, which have zero correlation between school size and the variable of interest, teacher's self-efficacy in instruction; second, mildly correlated data, which have $\rho = .4$; third, highly correlated data, which have $\rho = .7$.

3.1.2 Simulation Design

The simulations were conducted under the three conditions: sample size, type of initial proportions used, and correlation between the variable of interest and school size. School size was measured by the number of novice teachers within school. The simulations focus on examining under which conditions SICSUP would perform well when it is applied to the rare population of novice teachers.

Four levels of sample size were used: 50, 100, 500, and 1,000. The previous research (Reckase, Kim, & Ju, 2016) indicated that the selection probability (or inclusion probability) was changed depending on sample size. As the size of the sample increased, the selection probabilities of novice teachers in different size schools became similar to each other. When a sample was half the population size, selection probabilities became equal. That is, when the target sample size is a half of the population size, there is no selection bias due to clustering. Regarding the total population size of 2,000, the sample size of 1,000 is a half of the population size. Thus, the four levels of sample size can examine whether sample size affects accuracy in parameter estimation.

This dissertation used three types of initial proportions of novice teachers over strata: initial proportions based on data and two types of informal estimate. To apply cluster sampling with stratification for surveys, a frame (list) of sampling units is required. If that is not available, at least one should know the proportions of sampling units over strata and the average number of sampling units per cluster. This dissertation focuses on the situations when the proportions of sampling units over strata are unknown before sampling. The situations can be categorized into three conditions. First, researchers know the true proportions of novice teachers over strata in the population, called “initial proportions based on data” in this dissertation. Second, although

researchers do not know the true proportions, they may have an informal estimate of the proportions in the population. This estimate is based on the proportions of schools over strata, called "informal estimate based on school proportions", which may be different from the proportions of novice teachers over strata. Third, researchers may have another type of informal estimate. This estimate is based on the assumption that the proportions of novice teachers over strata are equal to each other, called "informal estimate based on equal proportions."

Table 3.3 Initial Proportions for Sampling

Area	Proportions of novice teachers over strata		
	Initial proportions based on data	Informal est. based on school proportions	Informal est. based on equal proportions
Rural	.20	.30	.33
Town	.25	.29	.33
City	.55	.42	.33

In practice, the first informal estimate could happen when the proportions of schools over strata are the best information available for researchers before starting the sampling procedure. The second informal estimate assumed that each stratum contains an equal number of novice teachers in the population. If researchers do not know anything about the proportions of novice teachers over strata before sampling, taking samples of equal size from strata could happen.

Correlation between the variable of interest and school size, meaning number of novice teachers within school, was categorized into three: zero, medium, and high (.0, .4, and .7, respectively). Teachers, including novice teachers, in large schools tend to stay longer in teaching than those in small schools (Allensworth, Ponisciak, & Mazzeo, 2009; Shin, 1995). High-quality teachers may find greater opportunities, such as advancement and promotion, in large schools which have more positions. These teachers are also more likely to leave small schools because working conditions in small schools are usually worse than those in large

schools (e.g., heavy teaching and working loads). Additionally, the previous research (Reckase, Kim, & Ju, 2016) showed that sequential cluster sampling (SICS without stratification) worked slightly worse in parameter estimation when school size was correlated with the variable of interest.

To sum, this dissertation considered total 36 conditions (4 sample sizes × 3 types of initial proportions × 3 levels of correlation). SICSUP, SICS, and SC were used to obtain sets of samples, and 10 sets of samples were created for each simulation condition.

3.1.3 Variance Estimator

The first research question focuses on estimating mean, standard deviation, and variance of the mean estimate. In order to estimate variance of the sample mean, $\sigma^2(\hat{\theta})$, four replication methods were used: the jackknife, bootstrap, balanced repeated replication (BRR), and the Fay's methods. The variance of the sampling distribution of $\hat{\theta}$ is defined to be

$$\sigma^2(\hat{\theta}) = E\left[\left(\hat{\theta}_S - E(\hat{\theta})\right)^2\right] = \sum_S P(S)\left[\hat{\theta}_S - E(\hat{\theta})\right]^2, \quad (3.1)$$

where $\hat{\theta}_s$ is the value of $\hat{\theta}$ calculated from sample *S* and *P(S)* is the selection probability. For SICSUP, it is difficult to calculate variance directly due to the complex sampling procedure. In that case, replication methods are used to obtain the variance of $\hat{\theta}$. The PISA uses the BRR with Fay's adjustment (OECD, 2017). The TIMSS uses one variation of the jackknife method, the jackknife repeated replication (JRR) (Martin, Mullis, & Hooperm, 2016). The NAEP also uses the jackknife method[2]. The bootstrap method is also widely used for large-scale surveys (Statistics Canada, 2018, 2019).

The jackknife method is chosen due to its popularity, simplicity, and relative ease of computation (Canty & Davison, 1999). The bootstrap method is relatively easy to implement and

[2] NAEP Assessment Weighting Procedures. https://nces.ed.gov/nationsreportcard/tdw/weighting/

enables researchers to more readily perform design-based analysis (Mach, Dumais, & Robinson, 2005). It also requires less computational burden as compared to other variance estimators (Chen & Shen, 2019). In bootstrap method, 500 replicates were generated for each simulation condition.

For BRR and Fay's methods, pseudo-strata were created. The schools were paired within the original strata, and each pair served as pseudo-stratum. For the jackknife and bootstrap, original and pseudo-strata were used for estimation. The BRR and Fay's methods use Hadamard matrices to create balanced half-samples. R package "survey" (Lumley, 2020) converts a sample to a sample with replicate-weights and estimates the standard error based on either of the jackknife, bootstrap, BRR or Fay's method. For BRR and Fay's methods, "survey" uses Hadamard matrices to create balanced half-samples. In "survey", users can choose whether to provide replicate weights or to create them by program, and I let the program create replicate weights.

3.1.4 Evaluation Criteria

For mean and standard deviation estimates, the mean square errors (MSE) were used to evaluate accuracy in estimation. The MSE is given by

$$MSE = \frac{1}{set}\sum_{i=1}^{set}\left[\hat{\theta}_i - E\left(\hat{\theta}\right)\right]^2 = \text{Variance} + \text{Bias}^2, \quad (3.2)$$

where *set* is the number of sample sets, 10, $\hat{\theta}_i$ is the estimate for each set of samples, and $E\left(\hat{\theta}\right)$ is the population parameter, which is the population mean or standard deviation. The MSE is a sum of the variance of estimates and squared bias of estimates. Smaller MSE indicates a more accurate estimator. Sample means and standard deviations were estimated with sampling weight and without sampling weight.

The purpose of weighting on the data for surveys is to obtain estimates of population parameters that do not suffer from bias due to the use of a complex sample design (Rutkowski, von Davier, & Rutkowski, 2013). Sampling weights are basically an inverse of selection probability.

Sampling weights were applied to each novice teacher with respect to the sample design used. For simple random sampling, all novice teachers have the equal sampling weight, and it is given by

$$w_{SRS} = \frac{N}{n}, \tag{3.3}$$

where N is the total number of novice teachers in the population, and n is the sample size. For SC, all novice teachers in the same school in each stratum have the equal sampling weight, and it is given by

$$w_{SC} = \frac{M_h}{M} \times \frac{N_h}{n_h} \times \frac{m_h^*}{m_h} = \frac{M_h}{M} \times \frac{N_h}{n_h}. \tag{3.4}$$

For each stratum, M_h is the total number of novice teachers in stratum h, M is the total number of novice teachers in the population, N_h is the total number of schools, n_h is the number of schools in the sample, m_h* is the number of novice teachers from n_h schools, and m_h is the number of novice teachers in the sample, which equals sample size. The right most term in the equation is equal to 1 because all teachers in the sampled school were added to samples.

For SICSUP and SICS, the following sampling weights were used:

$$SICSUP\ or\ SICS = \frac{M_h}{M} \times \frac{N_h}{n_h} \times \frac{m_h^*}{m_h}. \tag{3.5}$$

For each stratum, M_h is the total number of novice teachers in stratum h, M is the total number of novice teachers in the population, N_h is the total number of schools, n_h is the number of schools in the sample, m_h is the number of novice teachers in the school that was selected at

the end, called "last" school, and m_h* is the number of novice teachers sampled from the "last" school. All novice teachers in the same school receive the equal weight except the "last" school. Novice teachers in the "last" school might be selected randomly depending on the number of samples obtained before the last school.

In the first research question, standard errors (square root of variance) of each sample mean were estimated. Estimated bias, relative bias, relative MSE, and confidence interval coverage probability were used in order to compare the four replication standard error estimators: the jackknife, bootstrap, BRR, and Fay's estimators.

The estimated bias of a standard error estimator, σ,

$$Estimated\ bias = E\{\sigma\} - \sigma_{EMP} \tag{3.6}$$

is the difference between the average of standard error estimates from the 10 sample means, $E\{\sigma\}$, and the empirical standard error, σ_{EMP}. A positive value indicates that the standard error estimator tends to overestimate the empirical standard error and a negative value indicates that the standard error estimator tends to underestimate the empirical standard error. A value near zero is preferred and represents a good standard error estimator. The empirical standard error is the standard deviation of the 5,000 sample means.

The relative bias of a standard error estimator, σ, is given by

$$Rel.Bias = \frac{Estimated\ \ bias}{\sigma_{EMP}} = \frac{E\{\sigma\}-\sigma_{EMP}}{\sigma_{EMP}}. \tag{3.7}$$

The relative bias is the estimated bias divided by the empirical standard error. Because estimated bias can be a negative or positive value, the relative bias can also be a negative or positive value. The relative bias would be zero when the estimated bias is zero, which is hardly ever the case in real world. The relative bias expresses the estimated bias as a proportion of the

empirical standard error. A small absolute value of relative bias is preferred and indicates a good standard error estimator.

The third criterion is the relative MSE. The relative MSE of the standard error estimator, σ, is given by

$$Rel.MSE = \frac{MSE}{\sigma_{EMP}^2} = \frac{\frac{1}{set}\sum_{i=1}^{set}(\sigma_i - \sigma_{EMP})^2}{\sigma_{EMP}^2}, \quad (3.8)$$

where *set* is the number of the sample sets, 10. The relative MSE expresses the MSE as a proportion of the squared empirical standard error. Like the relative bias, a small value of relative MSE is preferred and indicates a good standard error estimator.

Finally, the confidence interval coverage probability is the probability of the 10 samples for which the estimated 95% confidence interval covers the population mean. The confidence interval for the sample mean, $\hat{y}$, is given by

$$CI = \hat{y} \pm z_{\alpha/2}\sigma, \quad (3.9)$$

where $z_{\alpha/2}$ is approximately 1.96, and σ is the standard error estimator. It is expected that the coverage probability would equal the nominal coverage probability of 95%. However, because 10 sets of samples were generated for each simulation condition, the coverage probability can only be expressed in tenth such as .9 and .8. Therefore, in this study, coverage probability of .9 or higher is preferred and interpreted as a good standard error estimator.

3.2 Research Question 2

The second research question is about how the appropriate sample size for SICSUP can be determined. The first question asked when a survey is being planned is what sample size to be used. The larger the sample size is, the better accuracy in estimation can be achieved although the more likely a hypothesis test will detect a small difference, increasing a probability of rejecting a null hypothesis. In addition, taking a larger sample requires more resource such as

time and cost. A survey should consider the maximum sample error one is willing to accept and the effect of the sample design on estimation precision so that the sample size for the survey can be decided.

3.2.1 Data and Simulation Design

Simulation studies are conducted using the same dataset and simulation conditions that are used in the previous section (the first research question). Each of SICSUP, SICS, and SC takes 10 sets of samples under the 36 conditions (4 sample sizes × 3 correlations × 3 initial proportions), and each set of sample provides the standard error of the sample mean. The results in the previous section (the first research question) indicate that the four replication standard error estimators such as the jackknife, bootstrap, BRR, and Fay's estimators work similarly to each other on average. For the second research question, the standard error for each simulation condition is obtained by averaging standard errors from the four standard error estimators with pseudo-strata.

3.2.2 Evaluation Criteria

Design Effect and Sample Size. Before data collection, sample size should be determined so that the results of the survey could provide a certain degree of precision in estimation. The sample size is determined by the margin of error, design effect, and confidence level. Complex sample designs such as SC usually require larger sample sizes than those for SRS in order to achieve the same level of precision.

The design effect (*Deff*) is the ratio of the variance of a sample that is from a complex sample design to the variance of a SRS sample with the same sample size:

$$Deff = \frac{{\sigma^2}_{Complex}}{{\sigma^2}_{SRS}}. \tag{3.10}$$

The design effect summarizes the effect of various complexities in the sample design such as clustering and stratification (Kish, 1965). The variance of stratified samples could be smaller than the variance of SRS samples due to stratification. Therefore, the design effect could be less than one. For clustered samples, the variance of clustered samples tends to be larger than the variance of SRS samples due to clustering. Thus, the design effect is typically larger than one. For stratified clustered samples, the design effect depends on the effect of stratification and clustering.

The required sample size, n, for the survey is a product of sample size for SRS and design effect and is given by

$$n = n_{SRS} \times Deff, \tag{3.11}$$

where n_{SRS} is the sample size for SRS. The required sample size for a complex sample design, n, and n_{SRS} can produce estimates at the same level of precision. For example, if a design effect is 2 and n_{SRS} is 100, samples of 200 from the complex sample design are required in order to obtain the results as precise as those from 100 SRS samples.

One of the simulation conditions used in this study is the different levels of sample size, ranged from 50 to 1,000. The design effect is computed for each level of sample size and used for calculating sample sizes for SICSUP, SICS and SC. The calculated sample sizes of SICSUP, SICS, and SC provide the same estimation precision as the given sample size of SRS would. It is expected that, under the same simulation condition, as the sample size increases, the design effect decreases. Thus, the difference in sample size between SRS and the three complex sample designs (SICSUP, SICS, and SC) would decrease.

Under the various simulation conditions, required sample sizes for SICSUP are mainly compared to those for SC. Small difference in sample size between SICSUP and SC indicates that SICSUP is as effective as SC.

Margin of Error and Sample Size. A margin of error is another factor for determining sample size. A margin of error refers to a limit of accuracy of a sample estimate (Agresti & Finlay, 2009). In other words, it shows how many points the results can be differ from the population parameter. To determine sample size, researchers should decide on the margin of error desired. The margin of error of an estimate is the maximum likely estimation error expected when the sample statistic is used as an estimator (Peck, 2014). In this study, the sample statistic is mainly the sample mean. The margin of error is

$$\text{ME} = z_{\alpha/2} \times \sqrt{\frac{\sigma^2}{n}} = 1.96 \times \sqrt{\frac{\sigma^2}{n}}, \tag{3.12}$$

where σ^2 is the population variance and n is the sample size. With a conventional 95% confidence level, 1.96 is used for $z_{\alpha/2}$. If the margin of error for the mean is d at a 95% confidence level, 95% sample means fall within the population mean plus or minus d.

Given the margin of error (ME), population variance (σ^2), population size (N), and a 95% confidence level, the necessary sample size for SRS, n_{SRS}, can be obtained by the following formula (Thompson, 2002):

$$n_{SRS} = \frac{1}{\frac{ME^2}{z^2\sigma^2} + \frac{1}{N}} = \frac{1}{\frac{1}{n_0} + \frac{1}{N}} \tag{3.13}$$

where

$$n_0 = \frac{z^2\sigma^2}{ME^2}.$$

In the dataset used for this research question, the population means are 12.37, 12.27, and 12.23 for uncorrelated data (ρ = .0), mildly correlated data (ρ = .4), and highly correlated data (ρ

= .7), respectively; the population standard deviations are 1.94, 2.05, 2.16 for uncorrelated data ($\rho = .0$), mildly correlated data ($\rho = .4$), and highly correlated data ($\rho = .7$), respectively. Considering the standard deviations, the five levels of margin of error were examined, from .1 to .5.

Table 3.4 presents the required sample sizes given the level of margin of error. For the population used in the second research question, the margin of error of .1 might require too many samples considering that the design effects are between 1 and 3. Given the margin of error, .1, if the design effect is 3, the required sample sizes (3 times the last column in Table 3.4) would be 2,520, 2,685, and 2,829 for $\rho = .0$, $\rho = .4$, and $\rho = .7$, respectively. These sample sizes are larger than population size of 2,000, and then, it is impossible to achieve the margin of error of .1 in this population. The lowest margin of error that can be obtained in this population is examined.

Table 3.4 Margin of Error and Required Sample Size for SRS

ρ	Margin of Error				
	0.5	0.4	0.3	0.2	0.1
$\rho = .0$	56	87	149	307	840
$\rho = .4$	63	96	165	337	895
$\rho = .7$	69	106	180	365	943

The sample sizes for SICSUP, SICS, and SC are computed based on the margin of error and sample sizes for SRS in Table 3.4. This study investigates whether SICSUP needs more samples than SC in order to achieve a given margin of error. If the required sample size for SICSUP is similar to that for SC, SICSUP can be considered as effective as SC.

3.3 Research Question 3

The third research question is about whether samples from SICSUP can determine group differences. The overall state rankings or country rankings compared with others are one of the

headline findings form national or international large-scale surveys and assessments in education (OECD, 2016). Education authorities and policy makers have paid great attention to rankings that provide information for their benchmarking tools to help develop educational strategies (Downing & Ganotice Jr., 2016). For the general public, such as parents and students, rankings also provide information about student's relative performance as compared to those in other states or countries. There is strong media interest in rankings because they are clear and easy to understand. Although results should not be interpreted naively and are often abused, state and country rankings are one of the most influential results from national and international surveys and assessments.

In this section, it is assumed that one would like to conduct an international survey in order to compare the rank order position of a country with the positions of other countries. Simulation studies are conducted to examine whether, in such situation, SICSUP would perform as well as SC, which has been frequently used for international surveys.

3.3.1 Data Generation

Like the first research question, the TALIS2018 (OECD, 2019) was used to generate datasets for the simulation studies. The data generation procedure here is basically the same as the one in the previous section.

Five countries were selected in the TALIS2018: Brazil, Canada, New Zealand, Portugal, and Taiwan. These five countries were selected because they have relatively higher proportions of schools with no novice teacher. In other words, novice teachers are rare in their countries as compared to other countries in the TALIS2018. The percentages of schools with no novice teacher are about 60% for Portugal, 31% for Brazil, 24% for Canada, 23% for Taiwan, and 20% for New Zealand.

Table 3.5 provides the summary of the five generated datasets. Although the generated datasets are based on the distributional information of the five countries, they are simulated datasets rather than real datasets. Therefore, in this dissertation, country 1, country 2, country 3, country 4, and country 5 refer to the datasets based on Portugal, Brazil, Canada, Taiwan, and New Zealand, respectively.

For each of the five countries, a population of 10,000 novice teachers was generated. In this research question, the variable of teacher's job satisfaction with profession was considered the variable of interest. The second and third columns in Table 3.5 present the means and standard deviations of the variable of interest in the generated populations. In the simulations, the population means are estimated by samples of SICSUP, SICS, and SC.

Table 3.5 Summary of the Generated Data by Countries

Country	Mean	SD	Total NT	Total School	% Sch. with No NT	Max NT*
Country 1	12.37	2.29	10000	162000	60.6	5
Country 2	11.31	2.06	10000	6222	31.8	9
Country 3	11.46	2.06	10000	5368	25.5	8
Country 4	11.70	1.75	10000	4950	23.8	12
Country 5	11.92	2.03	10000	4299	21.1	11

*Note: maximum number of novice teachers (NT) in a school

In the TALIS2018, the five countries used one or more stratification variables at the school level. The stratification variables were selected based on the stratification variables that were used in the TALIS2018 (OECD, 2019). In the TALIS2018, country 4 (Taiwan) used two types of stratification variables, such as location of school and source of funding. One of the strata had a very small fraction in the population (about 1.4%). This stratum was excluded in the generated population.

Table 3.6 Stratification Variable by Country

Country	Stratification variable	Level	Description
Country 1	School location	4	(1) village, hamlet or rural area, (2) small town, (3) town, and (4) city
Country 2	Source of funding	2	(1) public school and (2) private school
Country 3	School location	3	(1) rural, (2) town, and (3) city
Country 4	School location and source of funding	5	(Rural, town, and city) × (Public school and private school)*
Country 5	School size measured by number of enrolled students	5	(1) under 250, (2) 250-499, (3) 500-749, (4) 750-999, and (5) 1000 and above

*Note: private schools in rural area were excluded from the population due to very small proportion in the population (about 1.4%)

3.3.2 Simulation Design

For each country, the population mean (or country mean) is estimated. Given the margin of error (ME), population variance (σ^2), population size (*N*), and a 95% confidence level, the necessary sample size for SRS, n_{SRS}, can be obtained by the following formula (Thompson, 2002):

$$n_{SRS} = \frac{1}{\frac{ME^2}{z^2\sigma^2} + \frac{1}{N}} = \frac{1}{\frac{1}{n_0} + \frac{1}{N}} \tag{3.14}$$

where

$$n_0 = \frac{z^2\sigma^2}{ME^2}.$$

Based on the formula above, the required sample sizes for SRS given the levels of margin of error were calculated for each country (see Table 3.7). Although design effects of SICSUP, SICS, and SC for the five populations are not known, one may assume that the design effects are less than three considering the results of the second research question. Therefore, the sample sizes for SICSUP, SICS, and SC would be about three times larger than the sample sizes in Table 3.7. In order to achieve the margin of error, .1 for all five countries, around 5,000 samples (1679

× 3 = 5037) seem to be needed, which is a half of the population size of 10,000. About 1,500 samples are desired for the margin of error, .2, about 660 samples for the margin of error, .3, and 380 samples for the margin of error, .4. In this study, the sample size of 600 was chosen and it would give margin of errors slightly larger than .3 for country 1and less than .3 for the other countries, if the design effect is three.

Table 3.7 Required Sample Size for SRS by Margin of Error and Country

Country	Margin of Error				
	0.5	0.4	0.3	0.2	0.1
Country 1	80	125	219	480	1679
Country 2	65	101	178	391	1400
Country 3	65	100	177	390	1397
Country 4	63	98	174	382	1371
Country 5	47	73	129	286	1054

Two types of initial proportions of novice teachers over strata were used for simulations: initial proportions based on data and informal estimate of the population proportions based on school proportions. For each country, 10 sets of samples were taken and the results from the 10 sets of samples were averaged and reported.

3.3.3 Evaluation Criteria

The population means of the five countries are estimated using SICSUP, SICS, and SC samples. For each sample mean, standard error is also estimated using the jackknife estimator with original strata and the BRR estimator with pseudo-strata. The results from the first research question suggest that, on average, they work slightly better than the other estimators in SICSUP. Based on the results of the 10 sets of samples, 95% confidence interval coverage probability is investigated. The preferred values are .9 and 1.0 considering the number of sample sets, which cannot provide probabilities in hundredth.

The five countries are ranked in descending order of the sample means based on the samples from SICSUP, SICS, or SC, and the rankings are compared to those based on the population means. Four types of rank order are examined: country rankings based on each of the sample designs and country rankings based on a combination of SICSUP and SC.

National or international large-scale surveys and assessments often use different sample designs regarding the situations of the participating states or countries. For example, the overall sample design for the TALIS2018 was a stratified two-stage probability sample design (OECD, 2019). Stratification was applied based on the situation of each country. Geography, source of financing, type of educational program, and school size were used as stratification variables. In the case of the PISA, there were countries that used a three-stage design while the overall sample design was a two-stage design (OECD, 2017).

Country 1, 4, and 5 take samples using SICSUP while country 2 and 3 do it using SC. Country 1 has the rarest population in terms of number of schools with no novice teacher. Only 40% of schools contain at least one novice teacher. One of the strata in country 4 and 5 has a very small fraction: 8% in country 4 and 6% in country 5. It is expected that SICSUP would work well under these situations as compared to SC.

Based on the population means, country 1 has the highest mean, followed by country 5, 4, 3, and 2. The rankings estimated by samples of SICSUP, SICS, and SC are compared to the rankings based on the population means.

3.4 Research Question 4

The last research question evaluates the economic aspect of SICSUP as compared with that of SICS and SC. What is a good sample design? What are the optimal characteristics of a sample design? It is often said that a good sample design can achieve a fixed level of precision

with the least amount resources used such as cost and time. This description contains two aspects of good sample designs: statistical and economic aspects. The previous three research questions evaluate SICSUP with respect to statistical aspect. The last research question evaluates SICSUP in terms of economic aspect.

Drawing samples from a rare population often causes difficulties with respect to resource consumption because sampling units are hard to locate. If researchers take samples from a rare population using a conventional sample design, such as cluster sampling or multi-stage sampling, they would see a large proportion of units that do not satisfy the selection criterion. Based on the data used in this dissertation, if one draws schools from the population, there would be a large portion of "blank" schools in the selected schools, meaning schools with no novice teacher. Usually less resources, such as time and cost, are associated with observing a school with no novice teacher than observing a school with at least one novice teacher. Schools with no novice teacher are discarded without administering the survey, so the amount of resources used for such schools is less than for schools with at least one novice teacher. However, drawing schools still spends some resources regardless whether they are added to the final set of samples or not. For example, obtaining approval and cooperation of schools often takes time and cost. One advantage of SICSUP over conventional SC is that it can reduce the frequency of meeting such "blank" schools because of the sequential selection process. Therefore, SICSUP is expected to be more economical than SC.

3.4.1 Data and Simulation Design

To address the last research question, the generated data for the first and third research questions are used. SICSUP, SICS, and SC are used to draw samples from the populations, and the number of schools that are contacted during the sampling procedure and the number of

schools that are included in the final set of samples are examined. For the dataset from the first research question (dataset 1), different levels of sample size are applied including sample sizes of 50, 100, 500, and 1,000. For the dataset from the third research question (dataset 2), five different countries are examined given the sample size of 600.

For both of the datasets, the results are reported by strata in addition to the results based on the whole samples. In some situations, the resources required to conduct a survey may be different between strata. For example, travel cost is proportional to the distance. If location of school such as rural, town, and city is used as stratification, surveying samples in rural schools might be more expensive than those in city schools because the distance between rural schools tends to be greater than the distance between city schools. If SICSUP can achieve a predetermined sample size of novice teachers in rural area, which may require more resources than other strata, with fewer schools contacted as compared to SICS or SC would, SICSUP is more economic than the others.

For each dataset, 500 sets of samples are taken and the averaged results are reported in the result chapter.

3.4.2 Evaluation Criteria

The economic aspect of SICSUP are measured by the number of schools that are contacted during the selection process (n^*). These contacted schools consist of two types of schools: schools without novice teacher and schools with at least one novice teacher. The former is discarded without administering a survey, and the latter is added to the final sample set of novice teachers. This can be expressed by

$$n^* = n_{schools\ without\ novice\ teacher} + n_{schools\ with\ at\ least\ one\ novice\ teacher}. \quad (3.15)$$

Considering that the numbers of schools in the final set of samples are similar regardless of sample design, as the total number of contacted schools increases, the frequency of seeing "blank" schools during the selection process also increases. This can be considered less economical. The ratio of the number of schools in the final sample set to the number of contacted schools is used as the evaluation criterion. The value of 1 indicates that researchers did not meet any "blank" school during the sampling process. All of the contacted schools have at least one novice teacher and are added to the final set of samples. With smaller value, researchers more frequently met "blank" schools and hence, used more resources. The value of 0 indicates that all contacted schools are with no novice teacher, and researchers failed to sample any novice teacher through the sampling procedure. The value close to 1 suggests an economic sample design.

The ratio must be interpreted with the number of contacted schools or the number of schools in the final sample set because the ratio provides only relative information. For example, consider that , given a sample size, in SICSUP, the number of contacted schools and the number of schools in the final sample set are 10 and 9, respectively; in SC, the numbers are 20 and 18, respectively. Both of the cases provide the ratio of .9, but one cannot say that they are equally economical because SICSUP used fewer schools to achieve the given sample size than SC did.

The ratio of two sample designs is also used in order to evaluate the performance of SICSUP; (1) the ratio of the number of contacted schools in SICSUP (n^*_{SICSUP}) to those in SC (n^*_{SC}), $\frac{n^*_{SICSUP}}{n^*_{SC}}$, (2) the ratio of the number of contacted schools in SICSUP to those in SICS (n^*_{SICS}), $\frac{n*_{SICSUP}}{n*_{SICS}}$; and (3) the ratio of the number of contacted schools in SICS to those in SC, $\frac{n*_{SICS}}{n*_{SC}}$. The first ratio shows the effect of the updating process and sequential selection on the number of contacted schools, the second describes the effect of the updating process, and the third reveals the effects of the sequential selection. For each ratio, the smaller the value, the

greater the effect of the updating process, sequential selection, or both upon the number of contacted schools during the sampling procedure.

In addition to the two types of ratios, the probability of using substitute schools in SC is investigated. In this study, it is assumed that all sampled novice teachers participate in the survey, and using substitute schools only occurs when the novice teachers in the selected schools did not reach the predetermined sample size. Although it does not directly give information to evaluate the performance of SICSUP, the probability shows how SC works inappropriately in the rare population of novice teachers.

CHAPTER 4.

RESULTS

This chapter summarizes the results of the analyses organized into four sections corresponding to the four research questions described in Chapter 1. The first two sections report the results of the simulation studies that investigated the level of precision in estimating population parameters under various conditions. The third section also presents the results of the simulation studies that examined another statistical aspect of SICSUP. Unlike the first two sections, that assumed a national survey, the third section focuses on the application of SICSUP to international surveys. The last section focuses on the evaluating SICSUP in terms of economic aspect rather than statistical aspect.

Throughout the chapter, *n* represent a sample size, ρ represents a correlation coefficient between school size and the variable of interest. SICSUP refers to stratified inverse cluster sampling with updating process, SICS to stratified inverse cluster sampling without updating process, and SC to stratified cluster sampling. A symbol of σ represents a type of standard error estimator, and its subscripts denote a simulation condition: σ_J = jackknife standard error estimator, σ_B = bootstrap standard error estimator, σ_R = BRR standard error estimator, σ_F = Fay's standard error estimator, σ_{UJ} = jackknife standard error estimator using SICSUP samples, σ_{IJ} = jackknife standard error estimator using SICS samples, σ_{SJ} = jackknife standard error estimator using SC samples, σ_{UB} = bootstrap standard error estimator using SICSUP samples, σ_{IB} = bootstrap standard error estimator using SICS samples, σ_{SB} = bootstrap standard error estimator using SC samples, σ_{UR} = BRR standard error estimator using SICSUP samples, σ_{IR} = BRR standard error estimator using SICS samples, σ_{SR} = BRR standard error estimator using SC

samples, σ_{UF} = Fay's standard error estimator using SICSUP samples, σ_{IF} = Fay's standard error estimator using SICS samples, and σ_{SF} = Fay's standard error estimator using SC samples.

4.1 Research Question 1

The first research question is about whether SICSUP works at least as well as SC with respect to estimating population mean, standard deviation, and standard error (square root of variance) of the sample mean. For mean and standard deviation estimation, the mean squared errors (MSE) are reported in order to examine the estimation precision. For standard error estimation, estimated bias, relative bias, relative MSE, and 95% confidence interval coverage probability are reported for each simulation condition.

4.1.1 Mean and Standard Deviation

Mean. Table 4.1 shows the MSEs of the sample means and standard deviations in SRS. In both of sample means and standard deviations, as the sample size increases, the MSE decreases. This pattern stays the same regardless of correlation: no correlation between school size and the variable of interest ($\rho = .0$), mild correlation ($\rho = .4$), and high correlation ($\rho = .7$).

Table 4.1 MSE of the Mean and Standard Deviation Using SRS Samples

n	ρ	Mean	SD
50	0.0	0.07	0.02
50	0.4	0.11	0.01
50	0.7	0.04	0.05
100	0.0	0.03	0.03
100	0.4	0.02	0.03
100	0.7	0.04	0.04
500	0.0	0.00	0.00
500	0.4	0.01	0.00
500	0.7	0.00	0.00
1000	0.0	0.00	0.00
1000	0.4	0.00	0.00
1000	0.7	0.00	0.00

Table 4.2 MSE of Mean Using SICSUP, SICS, and SC Samples

		Weighted			Unweighted		
n	ρ	SICSUP	SICS	SC	SICSUP	SICS	SC
Initial Proportions Based on Data							
50	0.0	0.29	0.12	0.17	0.21	0.14	0.14
50	0.4	0.34	0.40	0.29	0.31	0.24	0.21
50	0.7	0.59	0.26	0.20	0.48	0.21	0.11
100	0.0	0.12	0.11	0.10	0.12	0.11	0.10
100	0.4	0.12	0.17	0.13	0.09	0.18	0.12
100	0.7	0.08	0.20	0.16	0.03	0.17	0.13
500	0.0	0.03	0.02	0.01	0.03	0.01	0.01
500	0.4	0.02	0.05	0.02	0.02	0.04	0.01
500	0.7	0.05	0.02	0.02	0.03	0.01	0.01
1000	0.0	0.01	0.00	0.01	0.01	0.00	0.01
1000	0.4	0.01	0.01	0.02	0.01	0.01	0.01
1000	0.7	0.02	0.02	0.02	0.01	0.01	0.01
Informal Estimate Based on School Proportions							
50	0.0	0.07	0.19	0.46	0.09	0.10	0.27
50	0.4	0.42	0.15	0.33	0.31	0.09	0.25
50	0.7	0.42	0.30	0.25	0.25	0.26	0.18
100	0.0	0.18	0.06	0.12	0.15	0.03	0.08
100	0.4	0.18	0.12	0.12	0.12	0.09	0.09
100	0.7	0.14	0.06	0.18	0.13	0.02	0.14
500	0.0	0.01	0.02	0.02	0.01	0.02	0.01
500	0.4	0.02	0.02	0.03	0.01	0.01	0.02
500	0.7	0.03	0.02	0.04	0.02	0.01	0.03
1000	0.0	0.01	0.01	0.00	0.01	0.00	0.00
1000	0.4	0.00	0.01	0.01	0.00	0.01	0.01
1000	0.7	0.02	0.01	0.01	0.01	0.01	0.01
Informal Estimate Based on Equal Proportions							
50	0.0	0.10	0.25	0.11	0.09	0.17	0.14
50	0.4	0.31	0.29	0.22	0.26	0.23	0.14
50	0.7	0.39	0.48	0.23	0.33	0.28	0.36
100	0.0	0.15	0.15	0.07	0.05	0.12	0.08
100	0.4	0.14	0.15	0.11	0.07	0.08	0.06
100	0.7	0.13	0.14	0.12	0.07	0.07	0.05
500	0.0	0.01	0.01	0.04	0.02	0.02	0.01
500	0.4	0.02	0.02	0.03	0.01	0.06	0.03
500	0.7	0.03	0.02	0.05	0.04	0.02	0.01
1000	0.0	0.01	0.01	0.01	0.01	0.00	0.00
1000	0.4	0.01	0.00	0.01	0.00	0.01	0.01
1000	0.7	0.01	0.01	0.01	0.01	0.01	0.01

Table 4.2 shows the MSEs of the sample means using the samples from three different sample designs: SICSUP, SICS, and SC. Weighted means and unweighted means were estimated under each simulation condition. The MSEs in SICSUP, SICS, and SC are greater than those in SRS when sample size is not large ($n \leq 100$). The MSEs in SRS are between .02 and .11 while those in the three sample designs are between .02 and .59. However, with medium to large sample sizes ($n \geq 500$), the MSEs in the three sample designs are very similar to those in SRS. That indicates that the sample means based on the three sample designs are almost as accurate as the sample means based on SRS. That also shows that SICSUP works as well as SC with medium to large sample sizes.

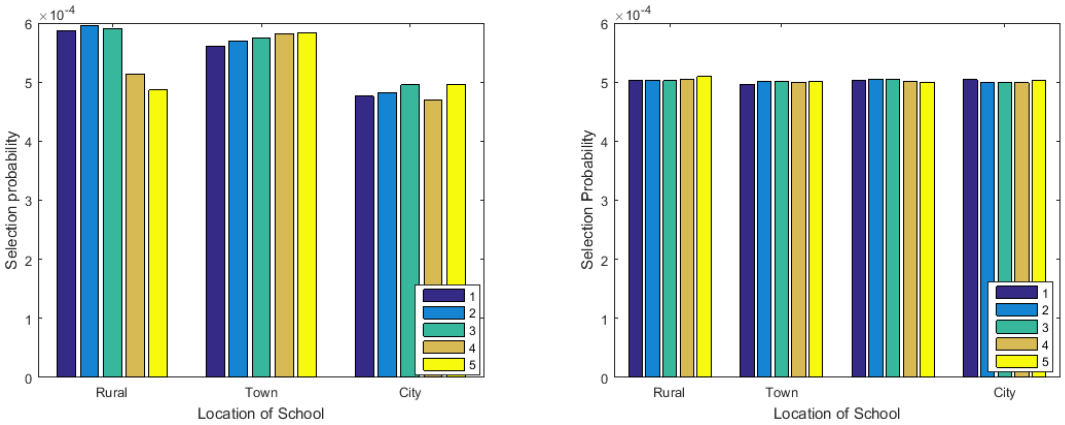
Under the simulation condition of initial proportions based on data and small sample size ($n \leq 100$), the updating process of SICSUP is not helpful to estimate the population mean as compared to the results in SICS and SC. For both of weighted and unweighted sample means, given the sample size of 50, the MSEs in SICSUP are larger than those in SICS and SC in general. For example, under the condition of no correlation ($\rho = .0$), the MSEs in SICSUP are .29 for the weighted mean and .21 for the unweighted mean while the MSEs in SC are .17 for the weighted mean and .14 for the unweighted mean.

When the sample size is small, the updating process of SICSUP might not be able to find the true proportions of novice teachers over strata in the population. The updating process relies on the samples that were collected to this point. Therefore, the number of samples that would be used for the updating process (n_1) is smaller than the predetermined sample size (n), expressed by $n = n_1 + n_2$, where n_2 is the number of novice teachers who are sampled after the updating process. For example, given the sample size of 50, the updating process relies on samples less than 50. Because of the small number of samples that are used for the updating process, the

updated proportions of novice teachers over strata might be different from those in the population. If researchers already know the true proportions (the proportions in the population) before the sampling procedure, the updating process is not necessary and thus, could not be able to increase accuracy in parameter estimation. Therefore, the MSEs in SICSUP are larger than those in SC when sample size is small. As the sample size increases, the updating process can provide proper information about the proportions of novice teachers over strata, and the MSEs in SICSUP and SC become similar to each other.

An interesting finding is that even with small sample size (n = 50), under the condition of $\rho = .0$ and either of informal estimates of the proportions used, SICSUP works better than SC in terms of MSE. Although the updating process with small sample size may not be helpful for parameter estimation, if researchers do not know the true proportions over strata, the updating process at least provides some useful information about the proportions. Therefore, SICSUP could produce better estimates than SC.

As shown in Figure 4.1, empirical selection probabilities (selection probabilities based on 5,000 sets of samples) in SICSUP indicate that, in general, when the sample size is small (n = 50), schools with more than one novice teachers have a slightly higher chance of being sampled than schools with one novice teacher. As the sample size increases, the selection probabilities become almost equal regardless of school size. This could influence accuracy in estimation especially under the simulation condition of $\rho > .0$. Under the condition of $\rho > .0$, large schools tend to have higher means than small schools because school size and school mean are positively correlated. In SICSUP as well as SICS, the MSEs under the condition of $\rho > .0$ and $n = 50$ are greater than those under the condition of $\rho = .0$ and $n = 50$.



*Note: each bar represents the number of novice teachers in school (e.g., dark blue bars refer to schools with one novice teacher).

Figure 4.1 Empirical Selection Probability for n=50 (left) and n=1,000 (right) Using SICSUP

When researchers have initial proportions based on data, with ρ = .0, the MSEs in SICS are similar to those in SC rather than those in SICSUP. This is because of the similarity in sampling procedure between SICS and SC under such condition. Both of them use stratification, clusters, and fixed sample sizes for strata. The only difference is the selection method when the selected schools have more novice teachers than required. In SICS, all novice teachers would be sampled in contacted schools except those in the lastly contacted school. For example, given the sample size of 50, consider that a researcher has collected 49 samples so far. The next contacted school has two novice teachers while the required number of novice teachers from that school is only one. In that case, the researcher would pick one out of the two novice teachers. In SC used in this dissertation, all two novice teachers in that school are once sampled and later, one novice teacher is removed randomly from the whole sample of 51. Except this difference, the sampling procedure between SICS and SC is similar to each other, that may lead the similar MSEs between the two sample designs under the condition mentioned above.

Taking all of results together, there are four main findings. First, as the sample size increases, the MSEs in SICSUP, SICS, and SC decrease and become close to those in SRS. Second, SICSUP works as well as SC when sample size is not small ($n \geq 500$). Three, the updating process of SICSUP may not be beneficial to estimate the population mean accurately with small sample size and correlated data ($\rho > .0$). However, if researchers do not know the true proportions of novice teachers over strata before the sampling procedure, under the condition of $\rho = .0$, SICSUP could be helpful even though the sample size is small. Finally, SICS works similar to SC under the condition of initial proportions based on data and $\rho = .0$.

Standard Deviation. Table 4.3 gives the MSEs of the sample standard deviations. Weighted standard deviations and unweighted standard deviations were estimated using the three sample designs: SICSUP, SICS, and SC.

Like the sample means, the MSEs in SICSUP, SICS, and SC are greater than those in SRS when sample size is not large ($n \leq 100$). The MSEs in SRS are between .01 and .05 while those in the three sample designs are between .03 and .27. However, with medium to large sample sizes ($n \geq 500$), the MSEs in the three sample designs are similar to those in SRS. That indicates that the sample standard deviations based on the three sample designs are as accurate as those based on SRS.

Table 4.3 MSE of Standard Deviation Using SICSUP, SICS, and SC Samples

		Weighted			Unweighted		
n	ρ	SICSUP	SICS	SC	SICSUP	SICS	SC
Initial Proportions Based on Data							
50	0.0	0.27	0.10	0.07	0.19	0.11	0.06
50	0.4	0.15	0.15	0.15	0.10	0.15	0.10
50	0.7	0.07	0.07	0.12	0.05	0.06	0.10
100	0.0	0.04	0.05	0.06	0.05	0.03	0.04
100	0.4	0.07	0.05	0.03	0.07	0.04	0.03
100	0.7	0.07	0.06	0.16	0.06	0.03	0.10
500	0.0	0.01	0.01	0.01	0.00	0.01	0.01
500	0.4	0.01	0.01	0.01	0.01	0.01	0.01
500	0.7	0.01	0.02	0.02	0.00	0.01	0.01
1000	0.0	0.01	0.01	0.00	0.00	0.00	0.00
1000	0.4	0.00	0.00	0.00	0.00	0.00	0.00
1000	0.7	0.01	0.00	0.01	0.00	0.00	0.00
Informal Estimate Based on School Proportions							
50	0.0	0.24	0.10	0.12	0.17	0.05	0.11
50	0.4	0.26	0.18	0.12	0.18	0.14	0.08
50	0.7	0.09	0.10	0.16	0.05	0.09	0.13
100	0.0	0.03	0.11	0.08	0.03	0.06	0.05
100	0.4	0.03	0.06	0.04	0.03	0.05	0.03
100	0.7	0.08	0.06	0.07	0.06	0.04	0.06
500	0.0	0.01	0.02	0.01	0.01	0.02	0.01
500	0.4	0.01	0.01	0.01	0.01	0.01	0.01
500	0.7	0.03	0.03	0.01	0.01	0.02	0.01
1000	0.0	0.00	0.00	0.00	0.00	0.00	0.00
1000	0.4	0.00	0.00	0.00	0.00	0.00	0.00
1000	0.7	0.01	0.01	0.01	0.01	0.01	0.01
Informal Estimate Based on Equal Proportions							
50	0.0	0.03	0.07	0.13	0.04	0.06	0.12
50	0.4	0.12	0.10	0.08	0.12	0.09	0.05
50	0.7	0.17	0.22	0.10	0.13	0.10	0.07
100	0.0	0.10	0.09	0.09	0.07	0.03	0.02
100	0.4	0.07	0.05	0.08	0.09	0.04	0.02
100	0.7	0.12	0.09	0.12	0.12	0.06	0.03
500	0.0	0.01	0.01	0.02	0.02	0.02	0.02
500	0.4	0.02	0.01	0.02	0.01	0.02	0.01
500	0.7	0.01	0.01	0.02	0.02	0.01	0.01
1000	0.0	0.00	0.00	0.00	0.00	0.00	0.00
1000	0.4	0.00	0.00	0.01	0.00	0.01	0.00
1000	0.7	0.00	0.00	0.00	0.01	0.01	0.01

The MSEs of the sample standard deviations show a similar pattern to those of the sample means. SICSUP works as well as SC in terms of MSE except when the sample size is very small (n = 50). Regardless of type of initial proportions and correlation, when the sample size is not very small ($n \geq 100$), the MSEs in SICSUP are not very different from those in SC for both of the weighted and unweighted sample standard deviations.

The greatest difference in MSE between SICSUP and SC occurs under the condition of initial proportions based on data, n = 50, and $\rho = .0$. On the other hand, under such condition, the difference between SICS and SC is not that great, showing the effect of sequential selection. These differences indicate that the updating process does not work well under the condition mentioned above. However, even under the condition of small sample size (n = 50) and $\rho = .0$, when informal estimate based on equal proportions is used, SICSUP works better than SC. The updating process provides at least some useful information about the proportions of novice teachers over strata when researchers don't know the true proportions. These results here agree with those in the previous section (mean estimation).

In SICSUP, the MSEs of the sample means under the condition of $\rho = .0$ tend to be smaller than those under the condition of $\rho = .4$ or .7. However, that is not the case in the sample standard deviations. There are some cases that the MSEs of the sample standard deviations under $\rho = .4$ or .7 are smaller than those under $\rho = .0$. For example, under the condition of initial proportions based on data, n = 50, and weighted samples, the MSE when $\rho = .0$ is .27; the corresponding value when $\rho = .7$ is .07.

In SICSUP, with small sample size of 50, schools with more than one novice teacher tend to have a higher chance of being selected than schools with one novice teacher (see Figure 4.1). When $\rho > .0$, large schools tend to have higher school means than small schools have. These two

factors may make MSEs small when $\rho > .0$ and $n = 50$ as compared to when $\rho = .0$ and $n = 50$. Under the former condition, in each SICSUP sample set, schools are similar in size, and school means are similar to each other (less spread out). Therefore, sample standard deviations in the sets of samples would be also similar to each other. This might reduce the MSEs, which refer to a sum of the variance of the estimates and squared bias of the estimates.

Like the mean estimation, under the condition of initial proportions based on data and $\rho = .0$, the MSEs in SICS are similar to those in SC rather than those in SICSUP because of the same reason as I mentioned in the previous section, which is the similarity in sampling procedure between the SICS and SC under such condition.

The MSEs of the sample standard deviations show a similar pattern to those of the sample means. Taking all of the results together, there are four main findings. First, as the sample size increases, the MSEs in SICSUP, SICS, and SC decrease and become close to those in SRS. Second, SICSUP works as well as SC when sample size is not very small ($n \geq 100$). However, under the condition of small sample size ($n = 50$), $\rho = .0$, and informal estimate based on equal proportions, SICSUP works better than SC. Third, there are some cases that the MSEs of the sample standard deviations when $\rho > .0$ are smaller than those when $\rho = .0$. Finally, SICS and SC works similarly to each other under the condition of initial proportions based on data and $\rho = .0$.

4.1.2 Standard Error of the Sample Mean

Table 4.4 gives the estimated bias, relative bias, relative MSE, and 95% confidence interval coverage probability for different standard error estimators using SRS. The estimates were obtained under the two conditions: without strata and with pseudo-strata. The jackknife and bootstrap estimators can be used for a sample without strata while the BRR and Fay's estimators

require a sample with a certain type of strata. Therefore, pseudo-strata were generated for samples in SRS. The novice teachers in each set of samples were randomly paired, and each pair represented a stratum.

Simple Random Sampling without Strata. The standard errors were computed based on the jackknife and bootstrap methods without using strata. In terms of estimated bias, relative bias, relative MSE, and confidence interval coverage probability, the two standard error estimators performed similarly well regardless of sample size and level of correlation.

The estimated bias is the difference between the empirical standard error and the average of standard error estimates from the 10 sets of samples and can be a negative value. The relative bias is the estimated bias divided by the empirical standard error, and hence, can be also a negative value.

Table 4.4 Estimated Bias, Relative Bias, Relative MSE, and Confidence Interval Coverage Probability (CV) of the Standard Error Estimators Using SRS without Strata

n	ρ	Jackknife				Bootstrap			
		Bias	Rel. Bias	Rel. MSE	CV	Bias	Rel. Bias	Rel. MSE	CV
50	0.0	0.01	0.03	0.01	1.00	0.01	0.04	0.01	1.00
50	0.4	0.00	0.01	0.00	0.90	0.01	0.02	0.01	0.90
50	0.7	-0.02	-0.05	0.01	1.00	-0.02	-0.06	0.01	1.00
100	0.0	-0.01	-0.04	0.01	0.90	-0.01	-0.03	0.01	0.90
100	0.4	0.00	0.02	0.01	1.00	0.01	0.04	0.01	1.00
100	0.7	0.01	0.03	0.01	0.90	0.01	0.04	0.01	0.90
500	0.0	0.01	0.20	0.04	0.90	0.01	0.20	0.04	0.90
500	0.4	0.01	0.16	0.03	0.90	0.01	0.15	0.02	0.90
500	0.7	0.01	0.14	0.02	1.00	0.01	0.15	0.02	1.00
1000	0.0	0.02	0.41	0.17	1.00	0.02	0.40	0.16	1.00
1000	0.4	0.02	0.38	0.14	1.00	0.02	0.35	0.12	1.00
1000	0.7	0.02	0.43	0.19	1.00	0.02	0.43	0.19	1.00

Simple Random Sampling with Pseudo-Strata. The standard errors were computed based on the jackknife, bootstrap, BRR, and Fay's methods with pseudo-strata. Table 4.5 shows the

estimated biases and relative biases. All standard error estimators exhibit similar estimated biases and relative biases regardless of sample size and level of correlation.

Table 4.5 Estimated Bias and Relative Bias of the Standard Error Estimators Using SRS with Pseudo-Strata

n	ρ	Bias				Rel. Bias			
		σ_J	σ_B	σ_{Br}	σ_F	σ_J	σ_B	σ_{Br}	σ_F
50	0.0	0.02	0.01	0.02	0.02	0.06	0.05	0.06	0.06
50	0.4	0.01	0.01	0.01	0.01	0.03	0.04	0.03	0.03
50	0.7	-0.03	-0.02	-0.03	-0.03	-0.09	-0.08	-0.09	-0.09
100	0.0	0.00	0.00	0.00	0.00	-0.02	-0.03	-0.02	-0.02
100	0.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
100	0.7	0.01	0.01	0.01	0.01	0.03	0.03	0.03	0.03
500	0.0	0.01	0.01	0.01	0.01	0.19	0.18	0.19	0.19
500	0.4	0.01	0.01	0.01	0.01	0.15	0.14	0.15	0.15
500	0.7	0.01	0.01	0.01	0.01	0.14	0.15	0.14	0.14
1000	0.0	0.02	0.02	0.02	0.02	0.43	0.40	0.43	0.43
1000	0.4	0.02	0.02	0.02	0.02	0.40	0.43	0.40	0.40
1000	0.7	0.02	0.02	0.02	0.02	0.42	0.42	0.42	0.42

Table 4.6 shows the relative MSE and confidence interval coverage probability. All standard error estimators report similar relative MSEs and confidence interval coverage probabilities regardless of sample size and level of correlation.

Table 4.6 Relative MSE and Confidence Interval Coverage Probability of the Standard Error Estimators Using SRS with Pseudo-Strata

n	ρ	Rel. MSE				CI coverage probability			
		σ_J	σ_B	σ_R	σ_F	σ_J	σ_B	σ_R	σ_F
50	0	0.02	0.03	0.02	0.02	1.00	1.00	1.00	1.00
50	0.4	0.02	0.02	0.02	0.02	0.90	0.90	0.90	0.90
50	0.7	0.03	0.03	0.03	0.03	1.00	1.00	1.00	1.00
100	0	0.00	0.01	0.00	0.00	1.00	1.00	1.00	1.00
100	0.4	0.02	0.02	0.02	0.02	1.00	1.00	1.00	1.00
100	0.7	0.01	0.02	0.01	0.01	0.90	0.90	0.90	0.90
500	0	0.04	0.04	0.04	0.04	0.90	0.90	0.90	0.90
500	0.4	0.03	0.02	0.03	0.03	0.90	0.90	0.90	0.90
500	0.7	0.02	0.03	0.02	0.02	1.00	1.00	1.00	1.00
1000	0	0.18	0.16	0.18	0.18	1.00	1.00	1.00	1.00
1000	0.4	0.16	0.19	0.16	0.16	1.00	1.00	1.00	1.00
1000	0.7	0.18	0.18	0.18	0.18	1.00	1.00	1.00	1.00

As expected, the results of SRS indicate that all of the standard error estimators work similarly to each other regardless of simulation conditions and serve as a basis in order to evaluate the performance of the standard error estimators using samples in SICSUP, SICS, and SC.

Three Sample Designs with Original Strata.

Estimated Bias. Table 4.7 provides the estimated bias for different standard error estimators with the original strata (rural, town, and city) and weights. Only the jackknife and bootstrap estimators were applied because application of the BRR and Fay's estimators require using pseudo-strata. The estimated bias is the difference between the empirical standard error and the average of standard error estimates from the 10 sets of samples and can be a negative value.

In general, the standard errors in SICSUP are similar to those in SC. No substantial difference between SICSUP and SC was found although some simulation conditions showed

larger difference than the other conditions did. When the sample size is small ($n = 50$), SICSUP reported slightly smaller biases than SC did. Considering the smaller bias is the better, SICSUP performed better than SC under such condition.

With non-small sample size, $n \geq 500$, regardless of sample design, type of standard error estimator, and other simulation conditions, the standard error estimators tend to overestimate the empirical standard error, which was obtained using 5,000 sample means. With small sample size, $n < 500$, the estimated bias could be negative or positive.

Table 4.7 Estimated Bias for the Standard Error Estimators with Original Strata and Weight

n	ρ	σ_{UJ}	σ_{IJ}	σ_{SJ}	σ_{UB}	σ_{IB}	σ_{SB}
Initial Proportions Based on Data							
50	0.0	-0.02	0.03	0.04	-0.03	0.02	0.03
50	0.4	0.01	0.09	-0.05	-0.01	0.07	-0.07
50	0.7	-0.02	0.00	-0.04	-0.03	-0.02	-0.05
100	0.0	-0.03	-0.01	-0.04	-0.04	-0.02	-0.04
100	0.4	-0.01	0.02	-0.02	-0.01	0.01	-0.02
100	0.7	0.01	0.00	0.02	0.00	-0.01	0.02
500	0.0	0.03	0.02	0.02	0.02	0.02	0.02
500	0.4	0.02	0.02	0.02	0.02	0.02	0.02
500	0.7	0.03	0.03	0.03	0.03	0.03	0.03
1000	0.0	0.04	0.04	0.04	0.04	0.03	0.04
1000	0.4	0.03	0.04	0.04	0.03	0.03	0.04
1000	0.7	0.05	0.04	0.05	0.04	0.04	0.04
Informal Estimate Based on School Proportions							
50	0.0	-0.02	-0.05	-0.03	-0.04	-0.05	-0.04
50	0.4	0.00	-0.01	-0.03	-0.01	-0.03	-0.05
50	0.7	0.00	0.05	0.07	-0.01	0.03	0.06
100	0.0	0.01	0.00	-0.04	0.01	-0.01	-0.04
100	0.4	0.01	-0.01	-0.02	0.00	-0.01	-0.02
100	0.7	-0.02	0.01	0.01	-0.03	0.01	0.01
500	0.0	0.03	0.01	0.01	0.03	0.01	0.01
500	0.4	0.02	0.03	0.02	0.02	0.02	0.02
500	0.7	0.03	0.02	0.02	0.03	0.02	0.02
1000	0.0	0.03	0.03	0.02	0.04	0.03	0.02
1000	0.4	0.04	0.03	0.04	0.04	0.03	0.03
1000	0.7	0.04	0.03	0.03	0.04	0.03	0.03
Informal Estimate Based on Equal Proportions							
50	0.0	0.03	-0.08	-0.04	0.02	-0.09	-0.05
50	0.4	-0.03	0.03	0.04	-0.04	0.02	0.03
50	0.7	-0.07	0.09	-0.06	-0.07	0.07	-0.07
100	0.0	0.01	0.00	0.01	0.01	-0.03	-0.03
100	0.4	-0.03	-0.04	-0.01	-0.02	0.00	-0.01
100	0.7	-0.08	-0.09	-0.05	-0.05	-0.02	-0.03
500	0.0	0.03	0.02	0.02	0.02	0.01	0.01
500	0.4	0.03	0.03	0.03	0.03	0.01	0.01
500	0.7	0.02	0.02	0.02	0.02	0.02	0.02
1000	0.0	0.04	0.04	0.03	0.03	0.02	0.03
1000	0.4	0.04	0.04	0.03	0.03	0.03	0.03
1000	0.7	0.04	0.04	0.04	0.04	0.03	0.03

On average, in SICSUP, the jackknife and bootstrap estimator worked similarly to each other. The difference in standard error between the two standard error estimators is very small, with maximum difference of .03. Figure 4.2 illustrates the absolute values of the estimated biases. With very small sample size (n = 50), the jackknife estimator might be slightly better than the bootstrap estimator in SICSUP.

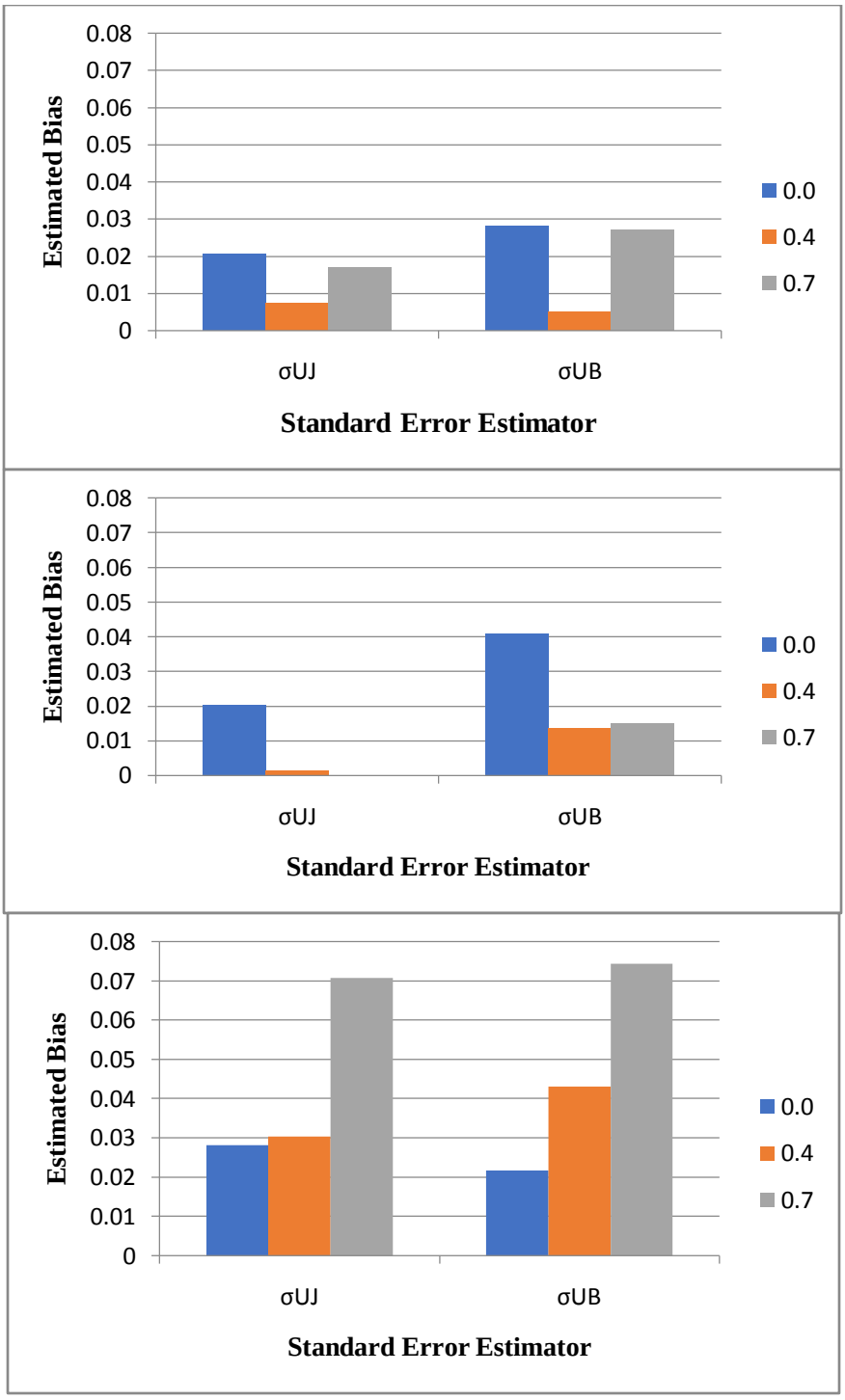


Figure 4.2 Estimated Bias of the Jackknife (σ_{UJ}) and Bootstrap (σ_{UB}) Estimators with n=50 and Original Strata by Type of Initial Proportions: Initial Proportions Based on Data (Top), Informal Estimate Based on School Proportions (Middle), and Informal Estimate Based on Equal Proportions (Bottom)

The estimated biases of standard error without sampling weight were also examined (see Appendix). The two results are not very different. The standard error estimators without weight tend to slightly less underestimate the empirical standard error than those with weight. This difference shows the influence of using sampling weight upon standard error estimation.

Relative Bias. Table 4.8 presents the relative biases of the standard error estimates with weights. Like the previous section, only the jackknife and bootstrap estimators were used.

On average, SICSUP worked as well as SC in terms of relative bias. When the sample size is small (n = 50), SICSUP reported slightly smaller relative biases than SC did. Considering the smaller relative bias is the better, SICSUP performed better than SC under such condition.

As the sample size increases, the relative biases tend to increase regardless of sample design, type of standard error estimator, and other simulation conditions. The standard error decreases as the sample size increases, and this reduction in standard error might cause the increase in relative bias. For instance, with sample size of 50, if empirical standard error is .10 and a standard error estimate is .11, the bias is .01 (.11 − .10). With sample size of 1,000, if the empirical standard error is .01 and a standard error estimate is .02, the bias is also .01 (.02 − .01). Although the biases are the same, the relative biases are different: .1 (.01/.10) for the former case and 1.0 (.01/.01) for the latter case. This caused the increase in relative bias with increasing sample size in Table 4.9. Therefore, the results should be interpreted given the same sample size.

On average, in SICSUP, the jackknife and bootstrap estimator worked similarly to each other in terms of relative bias. With very small sample size (n = 50), the jackknife estimator might be slightly better than the bootstrap estimator in SICSUP. These results agree with those of estimated bias.

Table 4.8 Relative Bias of the Standard Error Estimators with Original Strata and Weight

n	ρ	σ_{UJ}	σ_{IJ}	σ_{SJ}	σ_{UB}	σ_{IB}	σ_{SB}
Initial Proportions Based on Data							
50	0.0	-0.04	0.06	0.07	-0.06	0.03	0.07
50	0.4	0.01	0.16	-0.09	-0.01	0.12	-0.13
50	0.7	-0.03	0.00	-0.07	-0.04	-0.03	-0.09
100	0.0	-0.07	-0.03	-0.11	-0.10	-0.05	-0.12
100	0.4	-0.03	0.05	-0.05	-0.03	0.04	-0.05
100	0.7	0.01	0.00	0.04	0.00	-0.02	0.05
500	0.0	0.17	0.11	0.16	0.15	0.11	0.15
500	0.4	0.15	0.14	0.16	0.15	0.14	0.16
500	0.7	0.19	0.20	0.20	0.19	0.19	0.19
1000	0.0	0.44	0.41	0.48	0.45	0.40	0.47
1000	0.4	0.37	0.40	0.48	0.38	0.37	0.49
1000	0.7	0.45	0.46	0.49	0.43	0.45	0.47
Informal Estimate Based on School Proportions							
50	0.0	-0.04	-0.09	-0.07	-0.08	-0.10	-0.08
50	0.4	0.00	-0.02	-0.06	-0.03	-0.06	-0.09
50	0.7	0.00	0.09	0.13	-0.02	0.05	0.11
100	0.0	0.03	0.00	-0.11	0.02	-0.02	-0.12
100	0.4	0.02	-0.03	-0.05	0.00	-0.02	-0.06
100	0.7	-0.05	0.03	0.01	-0.07	0.03	0.01
500	0.0	0.18	0.10	0.09	0.18	0.09	0.08
500	0.4	0.12	0.17	0.13	0.10	0.15	0.13
500	0.7	0.20	0.10	0.14	0.20	0.09	0.14
1000	0.0	0.41	0.31	0.27	0.42	0.31	0.24
1000	0.4	0.42	0.36	0.40	0.42	0.36	0.39
1000	0.7	0.42	0.25	0.27	0.42	0.24	0.29
Informal Estimate Based on Equal Proportions							
50	0.0	0.06	-0.16	-0.08	0.04	-0.18	-0.10
50	0.4	-0.06	0.06	0.08	-0.08	0.04	0.05
50	0.7	-0.12	0.15	-0.10	-0.12	0.12	-0.13
100	0.0	0.02	-0.01	0.03	0.03	-0.07	-0.08
100	0.4	-0.08	-0.10	-0.04	-0.06	-0.01	-0.03
100	0.7	-0.19	-0.20	-0.12	-0.14	-0.06	-0.08
500	0.0	0.18	0.17	0.16	0.15	0.10	0.09
500	0.4	0.21	0.21	0.22	0.21	0.07	0.06
500	0.7	0.14	0.14	0.13	0.13	0.11	0.12
1000	0.0	0.43	0.41	0.42	0.41	0.26	0.27
1000	0.4	0.43	0.43	0.42	0.41	0.26	0.27
1000	0.7	0.44	0.44	0.43	0.45	0.28	0.28

The relative biases without sampling weight were also investigated (see Appendix). Although SICSUP does not show a substantial difference in relative bias between the two results, SICS and SC show a noticeable difference in relative bias under some simulation conditions, with the maximum difference of 30% in SICS and 28% in SC. This indicates the influence of using sampling weight for standard error estimation.

Relative MSE. Table 4.9 presents the relative MSEs of the standard error estimates with sampling weights. Like the previous sections, only the jackknife and bootstrap estimators were used.

On average, SICSUP worked as well as SC in terms of relative MSE except under some simulation conditions. For example, under the condition of $n = 50$, $\rho = .0$, and initial proportions based on data, the relative MSE in SICSUP (.17) is fairly greater than that in SC (.05). This result doesn't seem to agree with the previous results of estimated bias and relative bias. The MSE is a sum of the variance of estimates and squared bias of estimates. Under such condition, the standard errors may be widely spread out and lead to increase the MSE here.

In general, in SICSUP, the jackknife and bootstrap estimators worked similarly to each other. No substantial difference between the two standard error estimators was found, with the maximum difference of .03. Using either of the jackknife or bootstrap estimator would not make a big difference in estimating standard errors in SICSUP.

The relative MSEs without weight were also calculated (see Appendix). The difference between the two results is not significant in general although some simulation conditions produced a relatively large difference between results. This indicates the effect of using sampling weight upon standard error estimation.

Table 4.9 Relative MSE for the Standard Error Estimators with Original Strata and Weight

n	ρ	σ_{UJ}	σ_{IJ}	σ_{SJ}	σ_{UB}	σ_{IB}	σ_{SB}
Initial Proportions Based on Data							
50	0.0	0.17	0.05	0.05	0.17	0.06	0.05
50	0.4	0.06	0.06	0.04	0.05	0.05	0.05
50	0.7	0.02	0.04	0.04	0.02	0.04	0.04
100	0.0	0.02	0.02	0.03	0.03	0.03	0.03
100	0.4	0.03	0.03	0.01	0.03	0.03	0.02
100	0.7	0.02	0.03	0.07	0.03	0.04	0.07
500	0.0	0.04	0.02	0.03	0.03	0.02	0.03
500	0.4	0.03	0.02	0.03	0.04	0.03	0.03
500	0.7	0.04	0.05	0.05	0.04	0.04	0.04
1000	0.0	0.20	0.17	0.23	0.21	0.17	0.23
1000	0.4	0.14	0.16	0.23	0.15	0.14	0.24
1000	0.7	0.21	0.21	0.24	0.18	0.20	0.23
Informal Estimate Based on School Proportions							
50	0.0	0.11	0.07	0.05	0.12	0.06	0.05
50	0.4	0.14	0.11	0.08	0.12	0.09	0.06
50	0.7	0.02	0.05	0.08	0.02	0.04	0.07
100	0.0	0.01	0.09	0.05	0.02	0.09	0.04
100	0.4	0.03	0.03	0.02	0.03	0.03	0.02
100	0.7	0.04	0.05	0.04	0.04	0.07	0.04
500	0.0	0.04	0.02	0.01	0.04	0.02	0.01
500	0.4	0.02	0.04	0.03	0.01	0.03	0.03
500	0.7	0.05	0.03	0.03	0.05	0.03	0.03
1000	0.0	0.17	0.10	0.08	0.18	0.10	0.06
1000	0.4	0.18	0.13	0.16	0.18	0.13	0.15
1000	0.7	0.18	0.07	0.08	0.19	0.06	0.09
Informal Estimate Based on Equal Proportions							
50	0.0	0.05	0.07	0.05	0.05	0.07	0.05
50	0.4	0.05	0.07	0.09	0.05	0.05	0.07
50	0.7	0.07	0.20	0.06	0.07	0.17	0.06
100	0.0	0.05	0.04	0.04	0.04	0.04	0.05
100	0.4	0.04	0.03	0.02	0.03	0.04	0.03
100	0.7	0.06	0.06	0.03	0.04	0.04	0.04
500	0.0	0.04	0.03	0.03	0.03	0.02	0.03
500	0.4	0.06	0.06	0.06	0.05	0.01	0.01
500	0.7	0.02	0.02	0.02	0.02	0.02	0.03
1000	0.0	0.19	0.17	0.18	0.17	0.07	0.08
1000	0.4	0.19	0.18	0.18	0.17	0.07	0.08
1000	0.7	0.20	0.20	0.19	0.21	0.08	0.08

Confidence Interval Coverage Probability. Table 4.10 presents the confidence interval coverage probabilities of the standard error estimates with weights. Like the previous sections, only the jackknife and bootstrap estimators were used.

In terms of confidence interval coverage probability, SICSUP worked as well as SC. Under the condition of $n = 50$, $\rho = .0$, informal estimate based on school proportions, and the bootstrap estimator, SICSUP worked much better than SC: 1.0 in SICSUP and .7 in SC. The corresponding probabilities using jackknife estimator are 1.0 in SICSUP and .8 in SC. Under such condition, SICSUP worked better than SC in terms of confidence interval coverage probability.

In SICSUP, the jackknife and bootstrap worked almost identically. Although most of the coverage probabilities are either of .9 or 1.0, there are some conditions that show the coverage probability of .8, which is lower than the preferred value (.9 or higher). For example, under the condition of $n = 50$, informal estimate based on equal proportions, and $\rho = .7$, either of the jackknife or bootstrap estimator reported the coverage probability of .8.

Table 4.10 Confidence Interval Coverage Probability of the Standard Error Estimators with Original Strata and Weight

n	ρ	σ_{UJ}	σ_{IJ}	σ_{SJ}	σ_{UB}	σ_{IB}	σ_{SB}
Initial Proportions Based on Data							
50	0.0	1.00	1.00	0.90	1.00	1.00	0.90
50	0.4	0.90	0.80	0.90	0.90	0.90	0.90
50	0.7	0.90	1.00	0.80	0.80	0.90	0.80
100	0.0	0.90	1.00	1.00	0.90	1.00	1.00
100	0.4	0.90	0.90	1.00	0.90	0.90	0.90
100	0.7	1.00	0.90	1.00	1.00	0.90	1.00
500	0.0	0.80	1.00	1.00	0.80	1.00	1.00
500	0.4	1.00	0.90	1.00	1.00	0.90	1.00
500	0.7	0.90	1.00	1.00	0.90	1.00	1.00
1000	0.0	1.00	1.00	1.00	1.00	1.00	1.00
1000	0.4	0.90	1.00	1.00	0.90	1.00	1.00
1000	0.7	1.00	1.00	1.00	1.00	1.00	1.00
Informal Estimate Based on School Proportions							
50	0.0	1.00	1.00	0.80	1.00	1.00	0.70
50	0.4	0.80	1.00	0.80	0.90	1.00	0.80
50	0.7	1.00	0.90	1.00	1.00	0.90	1.00
100	0.0	1.00	1.00	1.00	0.90	1.00	1.00
100	0.4	0.90	1.00	0.90	0.90	0.90	0.90
100	0.7	0.90	1.00	1.00	0.90	1.00	1.00
500	0.0	1.00	1.00	1.00	1.00	1.00	1.00
500	0.4	0.90	1.00	1.00	0.90	1.00	1.00
500	0.7	1.00	1.00	1.00	1.00	1.00	1.00
1000	0.0	1.00	1.00	1.00	1.00	1.00	1.00
1000	0.4	1.00	1.00	1.00	1.00	1.00	1.00
1000	0.7	0.90	1.00	1.00	0.90	1.00	1.00
Informal Estimate Based on Equal Proportions							
50	0.0	1.00	0.90	1.00	1.00	0.90	1.00
50	0.4	0.90	0.90	1.00	0.90	0.90	1.00
50	0.7	0.80	1.00	0.90	0.80	1.00	0.90
100	0.0	1.00	1.00	1.00	1.00	1.00	1.00
100	0.4	0.90	0.90	0.80	0.80	1.00	1.00
100	0.7	0.90	0.90	0.90	0.80	1.00	1.00
500	0.0	1.00	1.00	1.00	1.00	0.90	0.90
500	0.4	1.00	1.00	1.00	1.00	1.00	1.00
500	0.7	1.00	1.00	1.00	1.00	1.00	1.00
1000	0.0	1.00	1.00	1.00	1.00	1.00	1.00
1000	0.4	1.00	1.00	1.00	1.00	1.00	1.00
1000	0.7	1.00	1.00	1.00	1.00	0.90	0.90

On average, SICS reported higher confidence interval coverage probabilities than the other two sampling designs did across the simulation conditions. Under some conditions, the coverage probabilities in SICSUP are slightly worse than those in SICS. This can be explained by two reasons: first, the mean was underestimated; second, standard error was underestimated. The range of confidence interval is determined by the sample mean and the standard error estimate. If both of the mean and the standard error are underestimated, the confidence interval coverage probability would decrease.

Additionally, the confidence interval coverage probabilities were computed using the estimated standard errors and unweighted means (see Appendix). In SICSUP, on average, the confidence interval coverage probabilities without sampling weight are similar to those with sampling weight.

Three Sample Designs with Pseudo-Strata. The BRR and Fay's methods require using a special type of strata. Each stratum should have two PSUs. In reality, such populations are rarely found, and hence, the standard error estimators based on the BRR and Fay's methods are often employed with pseudo-strata.

Estimated Bias. The estimated bias is the difference between the empirical standard error and the average of standard error estimates from the 10 sets of samples and can be a negative value.

As shown in Table 4.11, SICSUP worked as well as SC in most simulation conditions with respect to standard error estimation. Under the condition of $n = 50$, $\rho = .7$, and initial information based on data, SICSUP worked relatively worse than SC. The estimated biases are −.17 and −.04 in SICSUP and SC, respectively. This result is different from that with original strata. When original strata were used, under the same condition, SICSUP worked slightly better

than SC in terms of estimated bias. The use of pseudo-strata didn't change much the standard errors in SC while it made a noticeable change to standard errors when SICSUP was used, with the difference about .15. This suggests that using pseudo-strata may influence SICSUP more than SC.

Under most of the conditions, the four standard error estimators tend to underestimate the empirical standard error. The estimated biases are mostly negative. Why this happened? When original strata were used, biases were either of positive or negative. When $n \geq 500$, the jackknife and bootstrap estimators tended to overestimate the empirical standard error. When pseudo-strata were used, biases tended to be negative. Some previous studies reported similar results. For small sample sizes, Fay's estimator had tendency to underestimate the standard error (Paben, 1999).The jackknife estimator also seemed to underestimate the standard error with pseudo-strata (Folsom, 2014). The results of estimated bias in this study show that not only the jackknife and Fay's estimators but also the bootstrap and BRR estimators tended to underestimate the empirical standard errors. The underestimation may be related to the use of pseudo-strata.

In SICSUP, on average, the four standard error estimators worked similarly to each other. No substantial difference in estimated bias was found among the four standard error estimators. When informal estimate based on equal proportions was used, the BRR estimator performed slightly better than the other standard error estimators, especially with $n \geq 100$.

Table 4.11 Estimated Bias of the Standard Error Estimators with Pseudo-Strata and Weight

n	ρ	σ_{UJ}	σ_{IJ}	σ_{SJ}	σ_{UB}	σ_{IB}	σ_{SB}	σ_{UR}	σ_{IR}	σ_{SR}	σ_{UF}	σ_{IF}	σ_{SF}
Initial Proportions Based on Data													
50	0.0	-0.11	-0.05	-0.10	-0.10	-0.04	-0.10	-0.11	-0.04	-0.10	-0.11	-0.05	-0.10
50	0.4	0.05	-0.02	-0.11	0.05	-0.01	-0.11	0.05	-0.01	-0.11	0.05	-0.02	-0.11
50	0.7	-0.17	-0.08	-0.04	-0.17	-0.09	-0.04	-0.17	-0.09	-0.04	-0.17	-0.09	-0.04
100	0.0	-0.10	-0.12	-0.08	-0.10	-0.12	-0.08	-0.10	-0.12	-0.08	-0.10	-0.12	-0.08
100	0.4	-0.05	-0.04	-0.08	-0.06	-0.04	-0.08	-0.05	-0.04	-0.08	-0.05	-0.04	-0.08
100	0.7	-0.04	-0.07	-0.09	-0.04	-0.07	-0.09	-0.04	-0.07	-0.09	-0.04	-0.07	-0.09
500	0.0	-0.04	-0.05	-0.04	-0.04	-0.05	-0.04	-0.04	-0.05	-0.04	-0.04	-0.05	-0.04
500	0.4	-0.05	-0.04	-0.04	-0.05	-0.04	-0.04	-0.05	-0.04	-0.04	-0.05	-0.04	-0.04
500	0.7	-0.05	-0.08	-0.05	-0.05	-0.08	-0.05	-0.05	-0.08	-0.05	-0.05	-0.08	-0.05
1000	0.0	-0.03	-0.04	-0.02	-0.03	-0.03	-0.02	-0.03	-0.04	-0.02	-0.03	-0.04	-0.02
1000	0.4	-0.03	-0.03	-0.02	-0.03	-0.03	-0.02	-0.03	-0.03	-0.02	-0.03	-0.03	-0.02
1000	0.7	-0.03	-0.04	-0.03	-0.03	-0.04	-0.03	-0.03	-0.04	-0.03	-0.03	-0.04	-0.03
Informal Estimate Based on School Proportions													
50	0.0	-0.15	-0.13	-0.11	-0.15	-0.13	-0.10	-0.15	-0.13	-0.10	-0.15	-0.13	-0.11
50	0.4	-0.15	-0.09	-0.10	-0.15	-0.09	-0.09	-0.15	-0.08	-0.10	-0.15	-0.09	-0.10
50	0.7	-0.12	0.06	0.07	-0.11	0.06	0.07	-0.11	0.06	0.07	-0.12	0.06	0.07
100	0.0	-0.07	-0.02	-0.09	-0.08	-0.02	-0.09	-0.07	-0.02	-0.09	-0.07	-0.02	-0.09
100	0.4	-0.07	-0.03	-0.06	-0.07	-0.03	-0.06	-0.07	-0.03	-0.06	-0.07	-0.03	-0.06
100	0.7	-0.02	-0.03	-0.09	-0.02	-0.04	-0.09	-0.02	-0.03	-0.09	-0.02	-0.03	-0.09
500	0.0	-0.05	-0.05	-0.03	-0.05	-0.05	-0.04	-0.05	-0.05	-0.03	-0.05	-0.05	-0.03
500	0.4	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05
500	0.7	-0.05	-0.06	-0.06	-0.05	-0.06	-0.06	-0.05	-0.06	-0.06	-0.05	-0.06	-0.06
1000	0.0	-0.02	-0.02	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02	-0.03
1000	0.4	-0.03	-0.04	-0.02	-0.03	-0.04	-0.02	-0.03	-0.04	-0.02	-0.03	-0.04	-0.02
1000	0.7	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04
Informal Estimate Based on Equal Proportions													
50	0.0	-0.07	-0.15	-0.09	-0.06	-0.15	-0.10	-0.06	-0.15	-0.10	-0.06	-0.15	-0.10
50	0.4	-0.09	-0.08	-0.04	-0.09	-0.08	-0.04	-0.09	-0.09	-0.04	-0.09	-0.09	-0.04
50	0.7	-0.16	-0.02	-0.07	-0.16	-0.02	-0.07	-0.16	-0.02	-0.07	-0.16	-0.02	-0.07
100	0.0	-0.05	-0.05	-0.05	-0.05	-0.04	-0.04	-0.04	-0.04	-0.08	-0.09	-0.09	-0.09
100	0.4	-0.06	-0.06	-0.06	-0.06	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.03	-0.04
100	0.7	-0.11	-0.11	-0.11	-0.11	-0.07	-0.07	-0.07	-0.07	-0.06	-0.06	-0.06	-0.06
500	0.0	-0.05	-0.05	-0.05	-0.05	-0.04	-0.04	-0.04	-0.04	-0.03	-0.03	-0.03	-0.03
500	0.4	-0.04	-0.03	-0.04	-0.04	-0.03	-0.03	-0.03	-0.03	-0.06	-0.06	-0.06	-0.06
500	0.7	-0.07	-0.08	-0.07	-0.07	-0.06	-0.06	-0.06	-0.06	-0.08	-0.08	-0.08	-0.08
1000	0.0	-0.04	-0.04	-0.04	-0.04	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
1000	0.4	-0.03	-0.03	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02	-0.04	-0.04	-0.04	-0.04
1000	0.7	-0.04	-0.03	-0.04	-0.04	-0.03	-0.03	-0.03	-0.03	-0.04	-0.04	-0.04	-0.04

Additionally, estimated biases of the standard errors without sampling weight were examined (see Appendix). For most of the simulation conditions, the estimated biases without weight are slightly smaller than those with weight and are more underestimated than those with weight.

Relative Bias. As I mentioned previously in the results with original strata, relative biases of the standard error estimates do not necessarily decrease as the sample size increases.

On average, SICSUP worked as well as SC in terms of relative bias. Under some simulation conditions, SICSUP worked slightly worse than SC (e.g., $n = 50$, $\rho = .7$, and initial information based on data) while under other conditions, SICSUP worked slightly better than SC (e.g., $n = 100$, $\rho = .7$, and informal estimate based on school proportions). It is hard to find a pattern that explains the different performances of SICSUP with respect to relative bias. Under the condition of $n = 50$, $\rho = .7$, and initial information based on data, SICSUP worked relatively worse than SC, and this result agrees with that of estimated bias.

Table 4.12 Relative Bias of the Standard Error Estimators with Pseudo-Strata and Weight

n	ρ	σ_{UJ}	σ_{IJ}	σ_{SJ}	σ_{UB}	σ_{IB}	σ_{SB}	σ_{UR}	σ_{IR}	σ_{SR}	σ_{UF}	σ_{IF}	σ_{SF}
Initial Proportions Based on Data													
50	0.0	-0.21	-0.09	-0.20	-0.20	-0.08	-0.20	-0.21	-0.08	-0.20	-0.21	-0.09	-0.20
50	0.4	0.09	-0.04	-0.21	0.10	-0.03	-0.20	0.09	-0.02	-0.21	0.09	-0.03	-0.21
50	0.7	-0.27	-0.14	-0.07	-0.27	-0.14	-0.08	-0.27	-0.14	-0.07	-0.27	-0.14	-0.07
100	0.0	-0.27	-0.33	-0.23	-0.27	-0.34	-0.22	-0.27	-0.33	-0.24	-0.27	-0.33	-0.24
100	0.4	-0.14	-0.09	-0.22	-0.15	-0.11	-0.22	-0.13	-0.09	-0.23	-0.13	-0.09	-0.22
100	0.7	-0.09	-0.17	-0.22	-0.10	-0.16	-0.22	-0.09	-0.16	-0.22	-0.09	-0.16	-0.22
500	0.0	-0.27	-0.37	-0.28	-0.27	-0.37	-0.28	-0.28	-0.37	-0.28	-0.28	-0.37	-0.28
500	0.4	-0.31	-0.24	-0.24	-0.31	-0.23	-0.24	-0.31	-0.24	-0.24	-0.31	-0.24	-0.24
500	0.7	-0.31	-0.47	-0.29	-0.31	-0.47	-0.29	-0.31	-0.47	-0.29	-0.31	-0.47	-0.29
1000	0.0	-0.34	-0.41	-0.24	-0.34	-0.40	-0.25	-0.34	-0.41	-0.24	-0.34	-0.41	-0.24
1000	0.4	-0.38	-0.33	-0.29	-0.38	-0.33	-0.29	-0.37	-0.33	-0.29	-0.37	-0.33	-0.29
1000	0.7	-0.34	-0.39	-0.31	-0.34	-0.39	-0.31	-0.34	-0.39	-0.31	-0.34	-0.39	-0.31
Informal Estimate Based on School Proportions													
50	0.0	-0.29	-0.25	-0.22	-0.29	-0.25	-0.21	-0.28	-0.25	-0.21	-0.29	-0.26	-0.22
50	0.4	-0.29	-0.16	-0.19	-0.29	-0.16	-0.18	-0.28	-0.15	-0.19	-0.29	-0.16	-0.19
50	0.7	-0.20	0.09	0.12	-0.19	0.09	0.13	-0.19	0.10	0.12	-0.20	0.09	0.12
100	0.0	-0.20	-0.05	-0.25	-0.21	-0.06	-0.25	-0.19	-0.05	-0.24	-0.20	-0.05	-0.25
100	0.4	-0.17	-0.07	-0.17	-0.17	-0.07	-0.15	-0.17	-0.07	-0.17	-0.17	-0.07	-0.17
100	0.7	-0.05	-0.08	-0.23	-0.05	-0.09	-0.23	-0.05	-0.08	-0.22	-0.05	-0.08	-0.23
500	0.0	-0.33	-0.36	-0.23	-0.34	-0.36	-0.24	-0.33	-0.36	-0.23	-0.33	-0.36	-0.23
500	0.4	-0.33	-0.30	-0.35	-0.33	-0.30	-0.36	-0.33	-0.30	-0.35	-0.33	-0.30	-0.35
500	0.7	-0.31	-0.33	-0.36	-0.31	-0.34	-0.36	-0.31	-0.33	-0.36	-0.31	-0.33	-0.36
1000	0.0	-0.18	-0.23	-0.29	-0.18	-0.25	-0.27	-0.18	-0.23	-0.29	-0.18	-0.23	-0.29
1000	0.4	-0.32	-0.40	-0.21	-0.32	-0.40	-0.22	-0.31	-0.40	-0.21	-0.31	-0.40	-0.21
1000	0.7	-0.38	-0.36	-0.41	-0.38	-0.35	-0.41	-0.38	-0.36	-0.41	-0.38	-0.36	-0.41
Informal Estimate Based on Equal Proportions													
50	0.0	-0.13	-0.30	-0.20	-0.12	-0.29	-0.20	-0.12	-0.30	-0.20	-0.13	-0.30	-0.20
50	0.4	-0.16	-0.16	-0.08	-0.16	-0.16	-0.07	-0.16	-0.16	-0.08	-0.16	-0.16	-0.08
50	0.7	-0.27	-0.04	-0.12	-0.27	-0.04	-0.12	-0.27	-0.04	-0.12	-0.27	-0.04	-0.12
100	0.0	-0.14	-0.14	-0.14	-0.14	-0.12	-0.14	-0.12	-0.12	-0.23	-0.23	-0.23	-0.23
100	0.4	-0.16	-0.15	-0.16	-0.16	-0.13	-0.12	-0.13	-0.13	-0.10	-0.10	-0.09	-0.09
100	0.7	-0.26	-0.25	-0.26	-0.26	-0.20	-0.19	-0.20	-0.20	-0.14	-0.15	-0.14	-0.14
500	0.0	-0.34	-0.33	-0.34	-0.34	-0.34	-0.34	-0.34	-0.34	-0.21	-0.22	-0.21	-0.21
500	0.4	-0.23	-0.22	-0.23	-0.23	-0.22	-0.23	-0.22	-0.22	-0.37	-0.38	-0.37	-0.37
500	0.7	-0.43	-0.44	-0.43	-0.43	-0.38	-0.38	-0.38	-0.38	-0.45	-0.44	-0.45	-0.45
1000	0.0	-0.41	-0.42	-0.41	-0.41	-0.34	-0.35	-0.34	-0.34	-0.28	-0.29	-0.28	-0.28
1000	0.4	-0.30	-0.29	-0.30	-0.30	-0.22	-0.23	-0.22	-0.22	-0.37	-0.37	-0.37	-0.37
1000	0.7	-0.36	-0.36	-0.36	-0.36	-0.36	-0.36	-0.36	-0.36	-0.40	-0.40	-0.40	-0.40

Figure 4.3 shows the differences in relative bias under the condition of ρ = .7 and informal estimate based on equal proportions. The relative biases in SICSUP are expressed in blue, and those in SC are expressed in green regardless of standard error estimator. The same color was used for each sample design regardless of standard error estimator in order to show the difference in relative bias between the two sample designs clearly. With small sample size (*n* = 50), the relative biases in SICSUP (blue lines) are greater in absolute value than those in SC (green lines). However, as the sample size increases, the difference between the two sample designs becomes small, and with the sample size of 1,000, almost standard error estimators work similarly to each other except the Fay's estimator in SC (green dash line). If researchers want to use SICSUP and SC together for their survey under such condition, using sample sizes more than 50 would be recommended in order to keep the estimation precision constant across the sample designs.

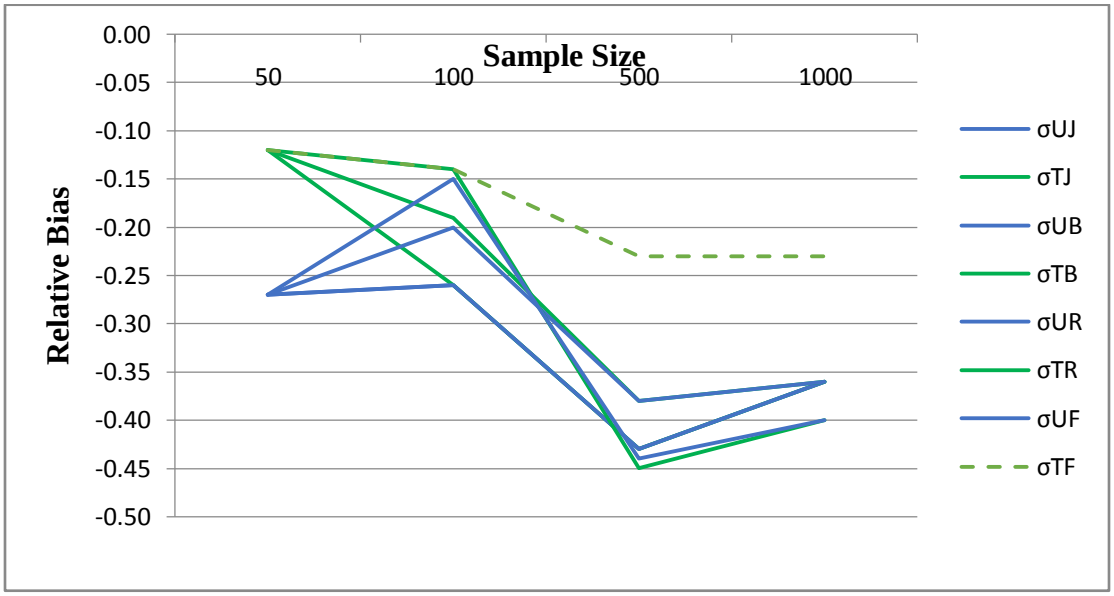


Figure 4.3 Relative Bias of the Standard Error Estimators by Sample Design (ρ= .7 and Informal Estimate Based on Equal Proportions)

In SICSUP, although all standard error estimators work similarly to each other in general, the BRR slightly works better than the other standard error estimators in terms of relative bias. The Fay's estimator also works slightly better than the other standard error estimators under some conditions, but it works worse than the others under different conditions. Therefore, the Fay's estimator seems less stable than the others. As the sample size increases, the difference in relative bias among the four standard error estimators decreases and becomes almost identical except under the condition of informal estimate based on equal proportions.

Additionally, the relative bias without weight was calculated (see Appendix). Many of the simulation conditions produced similar relative biases regardless of whether the weights were used or not. However, there are some conditions where the amount of difference in relative bias is greater than or equal to .1. Since the relative bias is expressed in a proportion, a difference of .1 represents a difference of 10%. About 1.3% (56 out of 432 simulation conditions) of the simulation conditions have differences greater than or equal to .1. Those relatively large differences happened in SICS and SC rather than in SICSUP. That means whether or not using weight has a more significant impact upon standard error estimation for samples in SICS and SC than those in SICSUP.

In SICSUP, although the relative biases with weight and without weight show similar patterns, there are some differences. With sampling weight, as the sample size increases, the relative biases increase. On the other hand, the relative biases without weight report relatively similar values as the sample size increases except under the condition of highly correlated data ($\rho = .7$). This is the same for all four standard error estimators. Figure 4.4 gives the relative biases with weight (blue lines) and those without weight (red lines) by sample size under some conditions. Four standard error estimators including the jackknife, bootstrap, BRR, and the Fay'

estimators were used. In Figure 4.4, all relative biases with weight are in blue and those without weight are in red in order to show the difference between weighted and unweighted samples clearly. The relative biases without weight are more constant across sample sizes than those with weight.

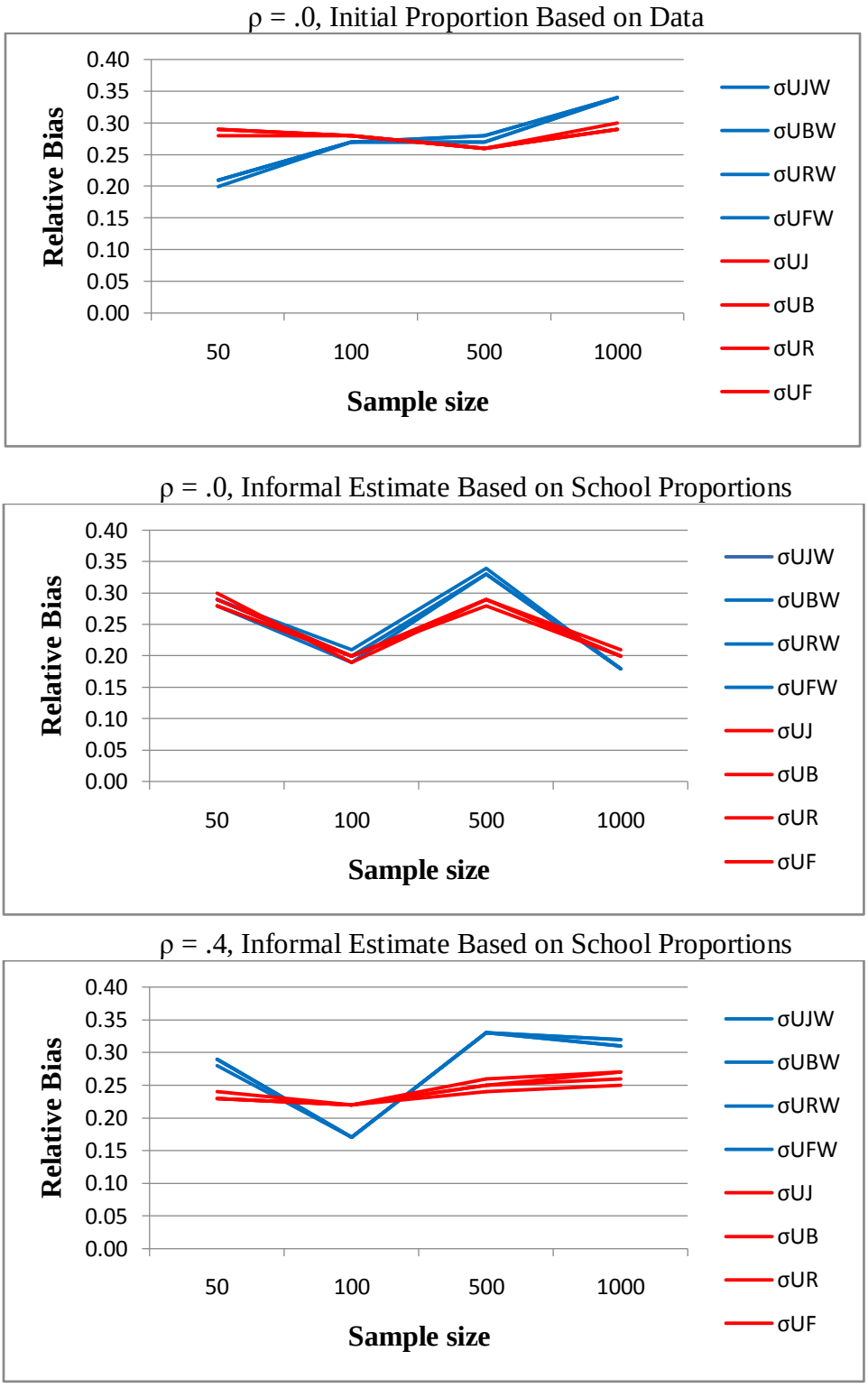


Figure 4.4 Relative Bias of the Standard Error Estimator with Weight (Blue Lines) and without Weight (Red Lines) by Sample Size Using SICSUP.

Relative MSE. On average, SICSUP worked slightly worse than SC did in terms of relative MSE, but the difference between the two sample designs was not substantial. The greatest difference happened under the condition of $n = 50$, $\rho = .0$, and initial proportions based on data. The difference is about .2, meaning about 20% difference in relative MSE. Although this result is different from the results of estimated bias and relative bias, it agrees with the result of relative MSE with original strata.

In SICSUP, throughout the simulation conditions, the jackknife, bootstrap, BRR, and Fay's estimators performed similarly to each other in terms of relative MSE except under the condition of informal estimate based on equal proportions. Under such condition, the BRR performed slightly better than the others. The Fay's estimator worked better than others under some conditions, but under different conditions, it worked worse than the others. It seems the Fay's estimator is less stable than the others under the condition of informal estimate based on equal proportions. As the sample size increases, the difference in relative MSE among the four standard error estimators decreases and becomes almost identical except under the condition of informal estimate based on equal proportions. These results are similar to those of relative bias.

Table 4.13 Relative MSE of the Standard Error Estimators with Pseudo-Strata and Weight

n	ρ	σ_{UJ}	σ_{IJ}	σ_{SJ}	σ_{UB}	σ_{IB}	σ_{SB}	σ_{UR}	σ_{IR}	σ_{SR}	σ_{UF}	σ_{IF}	σ_{SF}
Initial Proportions Based on Data													
50	0.0	0.26	0.07	0.06	0.25	0.07	0.06	0.25	0.07	0.06	0.25	0.07	0.06
50	0.4	0.13	0.06	0.09	0.14	0.05	0.09	0.13	0.06	0.09	0.13	0.06	0.09
50	0.7	0.12	0.13	0.08	0.12	0.12	0.08	0.12	0.13	0.08	0.12	0.13	0.08
100	0.0	0.08	0.13	0.08	0.08	0.14	0.07	0.08	0.13	0.08	0.08	0.13	0.08
100	0.4	0.07	0.06	0.06	0.07	0.06	0.06	0.07	0.05	0.06	0.07	0.06	0.06
100	0.7	0.09	0.11	0.09	0.09	0.11	0.10	0.09	0.11	0.10	0.09	0.11	0.10
500	0.0	0.10	0.15	0.09	0.09	0.15	0.09	0.10	0.15	0.09	0.10	0.15	0.09
500	0.4	0.13	0.08	0.07	0.13	0.08	0.07	0.13	0.08	0.07	0.13	0.08	0.07
500	0.7	0.12	0.25	0.11	0.12	0.25	0.10	0.12	0.25	0.11	0.12	0.25	0.11
1000	0.0	0.16	0.18	0.08	0.16	0.17	0.09	0.16	0.18	0.08	0.16	0.18	0.08
1000	0.4	0.16	0.13	0.09	0.16	0.13	0.10	0.16	0.13	0.09	0.16	0.13	0.09
1000	0.7	0.14	0.20	0.11	0.14	0.20	0.11	0.14	0.20	0.11	0.14	0.20	0.11
Informal Estimate Based on School Proportions													
50	0.0	0.17	0.10	0.14	0.18	0.10	0.14	0.17	0.10	0.14	0.17	0.10	0.14
50	0.4	0.19	0.20	0.16	0.18	0.19	0.16	0.18	0.20	0.16	0.18	0.20	0.16
50	0.7	0.09	0.14	0.14	0.09	0.14	0.14	0.09	0.14	0.14	0.09	0.14	0.14
100	0.0	0.09	0.12	0.10	0.09	0.11	0.10	0.09	0.12	0.10	0.09	0.12	0.10
100	0.4	0.10	0.08	0.07	0.10	0.08	0.06	0.10	0.08	0.07	0.10	0.08	0.07
100	0.7	0.08	0.14	0.07	0.09	0.14	0.07	0.08	0.13	0.07	0.08	0.13	0.07
500	0.0	0.15	0.15	0.07	0.16	0.15	0.07	0.15	0.15	0.07	0.15	0.15	0.07
500	0.4	0.14	0.12	0.14	0.14	0.12	0.14	0.14	0.12	0.14	0.14	0.12	0.14
500	0.7	0.13	0.15	0.15	0.13	0.16	0.15	0.13	0.15	0.15	0.13	0.15	0.15
1000	0.0	0.05	0.08	0.10	0.05	0.08	0.09	0.05	0.08	0.10	0.05	0.08	0.10
1000	0.4	0.13	0.18	0.06	0.12	0.18	0.06	0.13	0.18	0.06	0.13	0.18	0.06
1000	0.7	0.18	0.16	0.17	0.18	0.16	0.18	0.18	0.16	0.17	0.18	0.16	0.17
Informal Estimate Based on Equal Proportions													
50	0.0	0.08	0.14	0.11	0.09	0.14	0.11	0.09	0.14	0.12	0.09	0.14	0.11
50	0.4	0.07	0.17	0.14	0.07	0.18	0.14	0.07	0.18	0.14	0.07	0.18	0.14
50	0.7	0.14	0.18	0.08	0.14	0.18	0.08	0.14	0.18	0.08	0.14	0.17	0.08
100	0.0	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.11	0.11	0.11	0.11
100	0.4	0.13	0.13	0.14	0.13	0.10	0.11	0.10	0.10	0.09	0.10	0.09	0.09
100	0.7	0.15	0.15	0.15	0.15	0.11	0.11	0.11	0.11	0.10	0.10	0.10	0.10
500	0.0	0.15	0.15	0.15	0.15	0.14	0.14	0.14	0.14	0.09	0.08	0.09	0.09
500	0.4	0.07	0.07	0.07	0.07	0.06	0.07	0.06	0.06	0.16	0.16	0.16	0.16
500	0.7	0.20	0.21	0.20	0.20	0.16	0.15	0.16	0.16	0.24	0.24	0.24	0.24
1000	0.0	0.22	0.23	0.22	0.22	0.14	0.14	0.14	0.14	0.11	0.11	0.11	0.11
1000	0.4	0.09	0.09	0.09	0.09	0.05	0.05	0.05	0.05	0.17	0.17	0.17	0.17
1000	0.7	0.16	0.16	0.16	0.16	0.15	0.16	0.15	0.15	0.19	0.19	0.19	0.19

Additionally, the relative MSEs without weight were examined (see Appendix). On average, the relative MSEs without weight and those with weight are similar to each other. The greatest difference in relative MSE occurred in SICSUP under the condition of sample size of 1,000, $\rho = .0$, and informal estimate based on equal proportions. Except that condition, in SICSUP, the relative MSEs with weight and those without weight are similar.

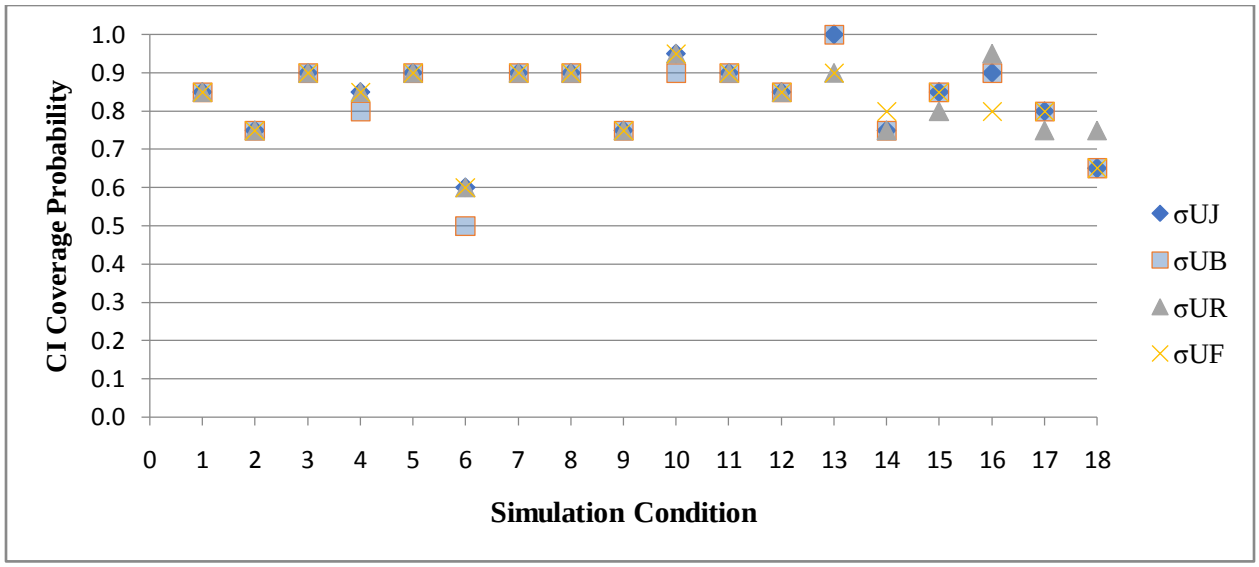
Confidence Interval Coverage Probability. Due to the underestimated standard errors, many of the confidence interval coverage probabilities did not reach the preferred value, .9 or higher. Only 47% of the simulation conditions reached the preferred value, and other conditions reported probabilities less than .9. Underestimated standard errors reduce the range of confidence interval and hence, decrease the coverage probabilities. This is the same regardless of standard error estimator and sample design used.

On average, SICSUP worked as well as SC in terms of confidence interval coverage probability except under some conditions. Under the condition of $n = 50$, $\rho = .0$, and informal estimate based on school proportions, SICSUP reported much higher probability than SC: 1.0 in SICSUP and about .5 in SC. On the other hand, under the condition of $n = 500$, $\rho = .0$, and initial information based on data, SICSUP reported lower probability than SC: .6 in SICSUP and .9 in SC.

Table 4.14 Confidence Interval Coverage Probability of the Standard Error Estimators with Pseudo-Strata and Weight

n	ρ	σ_{UJ}	σ_{IJ}	σ_{SJ}	σ_{UB}	σ_{IB}	σ_{SB}	σ_{UR}	σ_{IR}	σ_{SR}	σ_{UF}	σ_{IF}	σ_{SF}
Initial Proportions Based on Data													
50	0.0	0.80	0.90	0.80	0.80	1.00	0.80	0.80	0.90	0.80	0.80	0.90	0.80
50	0.4	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
50	0.7	0.80	1.00	0.80	0.80	1.00	0.70	0.80	1.00	0.80	0.80	1.00	0.80
100	0.0	0.90	0.80	0.90	0.90	0.80	0.90	0.90	0.80	0.90	0.90	0.80	0.90
100	0.4	0.90	1.00	0.90	0.90	1.00	0.90	0.90	1.00	0.90	0.90	1.00	0.90
100	0.7	1.00	0.90	0.90	1.00	0.90	0.90	1.00	0.90	0.90	1.00	0.90	0.90
500	0.0	0.60	0.70	0.90	0.60	0.70	0.90	0.60	0.70	0.90	0.60	0.70	0.90
500	0.4	0.90	0.80	0.80	0.90	0.80	0.80	0.90	0.80	0.80	0.90	0.80	0.80
500	0.7	0.70	0.70	0.90	0.50	0.80	0.90	0.70	0.70	0.90	0.70	0.70	0.90
1000	0.0	0.90	0.80	0.50	0.90	0.90	0.50	0.90	0.80	0.50	0.90	0.80	0.50
1000	0.4	0.80	0.70	0.50	0.70	0.70	0.50	0.80	0.70	0.50	0.80	0.70	0.50
1000	0.7	0.50	0.60	0.70	0.50	0.60	0.70	0.50	0.60	0.70	0.50	0.60	0.70
Informal Estimate Based on School Proportions													
50	0.0	1.00	0.90	0.50	1.00	0.90	0.50	1.00	0.90	0.60	1.00	0.90	0.60
50	0.4	0.70	0.90	0.80	0.70	0.90	0.80	0.70	0.90	0.80	0.70	0.90	0.80
50	0.7	0.90	0.80	1.00	0.90	0.80	1.00	0.90	0.80	1.00	0.90	0.80	1.00
100	0.0	0.80	0.80	0.70	0.80	0.80	0.70	0.80	0.80	0.70	0.80	0.80	0.70
100	0.4	0.80	0.90	0.90	0.80	0.90	0.90	0.80	0.90	0.90	0.80	0.90	0.90
100	0.7	0.90	1.00	0.80	0.90	1.00	0.80	0.90	1.00	0.80	0.90	1.00	0.80
500	0.0	0.90	0.90	0.90	0.90	0.90	0.80	0.90	0.90	0.90	0.90	0.90	0.90
500	0.4	0.90	1.00	0.60	0.80	1.00	0.60	0.90	1.00	0.60	0.90	1.00	0.60
500	0.7	1.00	0.80	0.50	1.00	0.80	0.50	1.00	0.80	0.50	1.00	0.80	0.50
1000	0.0	0.90	0.70	1.00	0.90	0.70	1.00	0.90	0.70	1.00	0.90	0.70	1.00
1000	0.4	1.00	0.80	0.90	1.00	0.80	0.80	1.00	0.80	0.90	1.00	0.80	0.90
1000	0.7	0.70	0.70	0.60	0.70	0.70	0.60	0.70	0.70	0.60	0.70	0.70	0.60
Informal Estimate Based on Equal Proportions													
50	0.0	1.00	0.90	1.00	1.00	0.90	1.00	1.00	0.90	1.00	1.00	0.90	1.00
50	0.4	0.80	0.90	0.90	0.80	0.90	0.90	0.80	0.90	0.90	0.80	0.90	0.90
50	0.7	0.70	0.90	0.90	0.70	0.80	0.80	0.70	0.90	0.90	0.70	0.90	0.90
100	0.0	1.00	1.00	1.00	1.00	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
100	0.4	0.90	0.90	0.90	0.90	0.80	0.80	0.80	0.80	0.90	0.90	0.90	0.90
100	0.7	0.90	0.90	0.90	0.90	0.80	0.80	0.80	0.80	0.90	0.90	0.90	0.90
500	0.0	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.70	0.70	0.70	0.70
500	0.4	1.00	1.00	1.00	1.00	1.00	0.90	1.00	1.00	0.80	0.80	0.80	0.80
500	0.7	0.60	0.60	0.60	0.60	0.70	0.60	0.70	0.70	0.50	0.50	0.50	0.50
1000	0.0	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.90	0.90	0.90	0.90
1000	0.4	0.80	0.80	0.80	0.80	0.90	0.90	0.90	0.90	0.80	0.80	0.80	0.80
1000	0.7	0.70	0.70	0.70	0.70	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80

In SICSUP, the four standard error estimators performed similarly to each other except under the condition of $n \geq 100$ and informal estimate based on equal proportions. Under such condition, the BRR and Fay's estimators worked differently as compared to the jackknife and bootstrap estimators. Figure 4.5 illustrates the difference in confidence interval coverage probabilities in SICSUP. Under most of the simulation conditions, the four standard error estimators work almost identically except under conditions of informal estimate based on equal proportions (condition 13 to 18 in Figure 4.5).



#	Simulation Condition	#	Simulation Condition
1	Small sample, ρ = .0, initial proportions based on Data	10	Non-small sample, ρ = .4, informal estimate based on school proportions
2	Non-small sample, ρ = .0, initial proportions based on Data	11	Small sample, ρ = .7, informal estimate based on equal proportion
3	Small sample, ρ = .4, initial proportions based on Data	12	Non-small sample, ρ = .7, informal estimate based on equal proportion
4	Non-small sample, ρ = .4, initial proportions based on Data	13	Small sample, ρ = .0, informal estimate based on equal proportion
5	Small sample, ρ = .7, initial proportions based on Data	14	Non-small sample, ρ = .0, informal estimate based on equal proportion
6	Non-small sample, ρ = .7, initial proportions based on Data	15	Small sample, ρ = .4, informal estimate based on equal proportion
7	Small sample, ρ = .0, informal estimate based on school proportions	16	Non-small sample, ρ = .4, informal estimate based on equal proportion
8	Non-small sample, ρ = .0, informal estimate based on school proportions	17	Small sample, ρ = .7, informal estimate based on equal proportion
9	Small sample, ρ = .4, informal estimate based on school proportions	18	Non-small sample, ρ = .7, informal estimate based on equal proportion

Figure 4.5 Confidence Interval Coverage Probability with Pseudo-Strata and Weight

Additionally, the confidence interval coverage probabilities without sampling weight were examined (see Appendix). In SICSUP, on average, coverage probabilities with weight tend to be slightly higher than those without weight.

In line with the findings presented so far, on average, the performance of SICSUP was as good as SC in estimating the population mean, the population standard deviation, and the standard error of the sample mean.

For mean and standard deviation estimation, with $n \geq 500$, SICSUP worked as well as SC. In addition, with $n = 1{,}000$, the performance of the three sample designs became close to that of SRS, with only slight difference. For standard error estimation, SICSUP worked as well as SC except under some conditions. The conditions are different by evaluation criteria or type of strata used, but the common factor is small sample size ($n = 50$). Therefore, very small sample size should be avoided when SICSUP is used.

4.2 Research Question 2

One critical issue in applying a complex sample design is the determination of sample size. This is typically done by determining amount of error that a researcher would allow. The second research question is about how the appropriate sample size for SICSUP can be determined. To address this research question, first, the design effects and corresponding sample sizes were computed for each sample design, based on the standard errors that were obtained from the first research question; second, given the margin of error, required sample sizes for SICSUP, SICS, and SC were examined.

4.2.1 Design Effect and Sample Size

Table 4.15 shows the design effects of SICSUP, SICS, and SC by sample size and initial information about proportions of novice teachers over strata. The standard errors based on the

four replication methods were averaged because of the only slight difference in the estimates among the replication methods. In addition, the levels of correlation between school size and the variable of interest were averaged because of the same reason, and medians were used due to some outliers. On average, the design effects based on the weighted samples are around 2.30, 2.55, and 2.21 in SICSUP, SICS, and SC, respectively. The design effects based on the samples without weight are around 1.86, 2.01, and 1.89 in SICSUP, SICS, and SC, respectively.

As expected, in general, the design effects seem to decrease as the sample size increases regardless of the type of sample design. The design effect measures relative efficiency between a complex sample design and SRS. As the sample size increases, the effect of a complex sample design decreases and the design effect approaches to 1, meaning that its efficiency becomes close to that of SRS.

The type of initial proportions of novice teachers over strata made noticeable differences in design effect when the sample size is small (n = 50), especially in SICSUP and SICS. In the weighted samples, with initial proportions based on data, the designs effects for SICSUP and SICS are 3.03 and 3.63, respectively; with informal estimate based on school proportions, 2.30 and 3.30, respectively; with informal estimate based on equal proportions, 2.74 and 3.01, respectively. The use of informal estimates led to reduce the design effects of SICSUP as compared to the design effect when initial proportions based on data used. On the other hand, the design effects of SICS remained similar regardless of type of initial proportions used. This shows the effect of the updating process in SICSUP on the efficiency of sample design. The updating process is beneficial when initial proportions are different from those in the population, especially for small sample size. With very large sample size (n = 1,000), the effect of the updating process disappears, and, on average, the design effects tend to be similar among the

three sample designs. When the sample size is 1,000, regardless of type of sample design, a half of novice teachers are taken from the population.

Table 4.15 Design Effect for the Variable of Interest

n	Weighted			Unweighted		
	SICSUP	SICS	SC	SICSUP	SICS	SC
Initial Proportions Based on Data						
50	3.03	3.63	2.38	2.00	2.72	2.07
100	2.89	3.27	2.14	2.30	2.52	1.94
500	2.08	1.67	2.03	1.63	1.53	1.74
1000	1.70	1.69	1.85	1.42	1.42	1.57
Informal Estimate Based on School Proportions						
50	2.30	3.30	2.56	1.97	2.32	2.14
100	2.90	3.58	2.36	1.97	2.37	1.88
500	1.90	2.04	1.73	1.65	1.38	1.50
1000	1.79	2.29	2.09	1.59	1.49	1.70
Informal Estimate Based on Equal Proportions						
50	2.74	3.01	3.03	2.25	2.12	2.42
100	2.61	2.55	2.75	2.11	2.42	2.22
500	1.93	1.88	1.84	1.68	1.91	1.77
1000	1.75	1.69	1.77	1.71	1.89	1.77

Table 4.16 presents desired sample sizes based on the design effects in Table 4.15. Under the condition of relatively small sample size ($n \leq 100$), on average, samples of SICSUP three times larger than those of SRS can achieve the same precision in estimation (e.g., 50 for SRS and 152 for SICSUP). Under the condition of medium to large sample size ($n \geq 500$), samples of SICSUP about two times larger than those of SRS can achieve the same precision in estimation (e.g., 500 for SRS and 1,040 for SICSUP). In order to achieve the same level of accuracy in estimation with 1,000 samples of SRS, SICSUP needs to take more than 85% of novice teachers in the population (more than 1,700 novice teachers). The required sample size for SICSUP in this situation seems hard to carry out in practice.

With weighted samples, 50% of the simulation conditions require more samples in SICSUP than for SC. With unweighted samples, 25% of the simulation conditions require more samples in SICSUP than in SC.

Table 4.16 Desired Sample Size

n	Weighted			Unweighted		
	SICSUP	SICS	SC	SICSUP	SICS	SC
Initial Proportions Based on Data						
50	152	181	126	100	136	103
100	289	327	225	230	252	194
500	1040	836	913	815	763	868
1000	1701	1686	1719	1416	1419	1571
Informal Estimate Based on School Proportions						
50	115	165	125	99	116	107
100	290	358	237	197	237	188
500	949	1018	823	823	691	750
1000	1788	2294	2012	1595	1492	1700
Informal Estimate Based on Equal Proportions						
50	137	151	140	113	106	121
100	261	255	275	211	242	222
500	963	939	921	840	953	885
1000	1746	1693	1772	1710	1890	1768

Figure 4.6 to Figure 4.9 illustrate the differences in desired sample sizes, that can provide parameter estimates as accurate as the given SRS samples would, among SICSUP, SICS, and SC. The leftmost point on the horizontal axis in each figure represents the sample size of SRS. The vertical axis represents simulation conditions; the first term, ρ, denotes the correlation coefficient between school size and the variable of interest, and the second term denotes the type of initial proportions of novice teachers over strata (e.g., "Data" for initial proportions based on data, "Informal 1" for informal estimate based on school proportions, and "Informal 2" for informal estimate based on equal proportions). Figure 4.6 to Figure 4.9 give the required sample sizes for SICSUP, SICS, and SC given SRS samples of 50, 100, 500, and 1000, respectively.

In the figures, there are some odd sample sizes that are substantively different from the other sample sizes. For example, in Figure 4.8, the required sample size for SICS, under the condition of $\rho = .7$ and initial proportions based on data, seems too small ($n = 623$) as compared to those for SICSUP ($n = 1{,}040$) and SC ($n = 890$). In this dissertation, 10 sets of samples were generated for each simulation condition and some sets with outliers might affect the results.

Without sampling weight, the required sample sizes are smaller than with sampling weight. This is because the unweighted samples have smaller design effects than the weighted samples have. The differences in sample sizes among SICSUP, SICS, and SC are not significant as compared to the differences with sampling weight.

As shown in Figure 4.9, some simulation conditions produce the required sample sizes greater than the population size of 2,000 (blue vertical line in Figure 4.9). These happen when either of informal estimates is used as initial proportions. The deviation of required sample size for SICSUP from the sample size of SRS becomes great as the sample size increases because the sample size for SICSUP is a product of the sample size of SRS and the corresponding design effect. For example, with SRS samples of 50, the design effect of 3 for SICSUP gives the required sample size of 150, and the difference in sample size between two sample designs is 100. On the other hand, with SRS samples of 1,000, the design effect of 3 for SICSUP gives the required sample size of 3,000, and the difference in them is 3,000. Although the design effect is not changed, the difference in sample size between SICSUP and SRS increases.

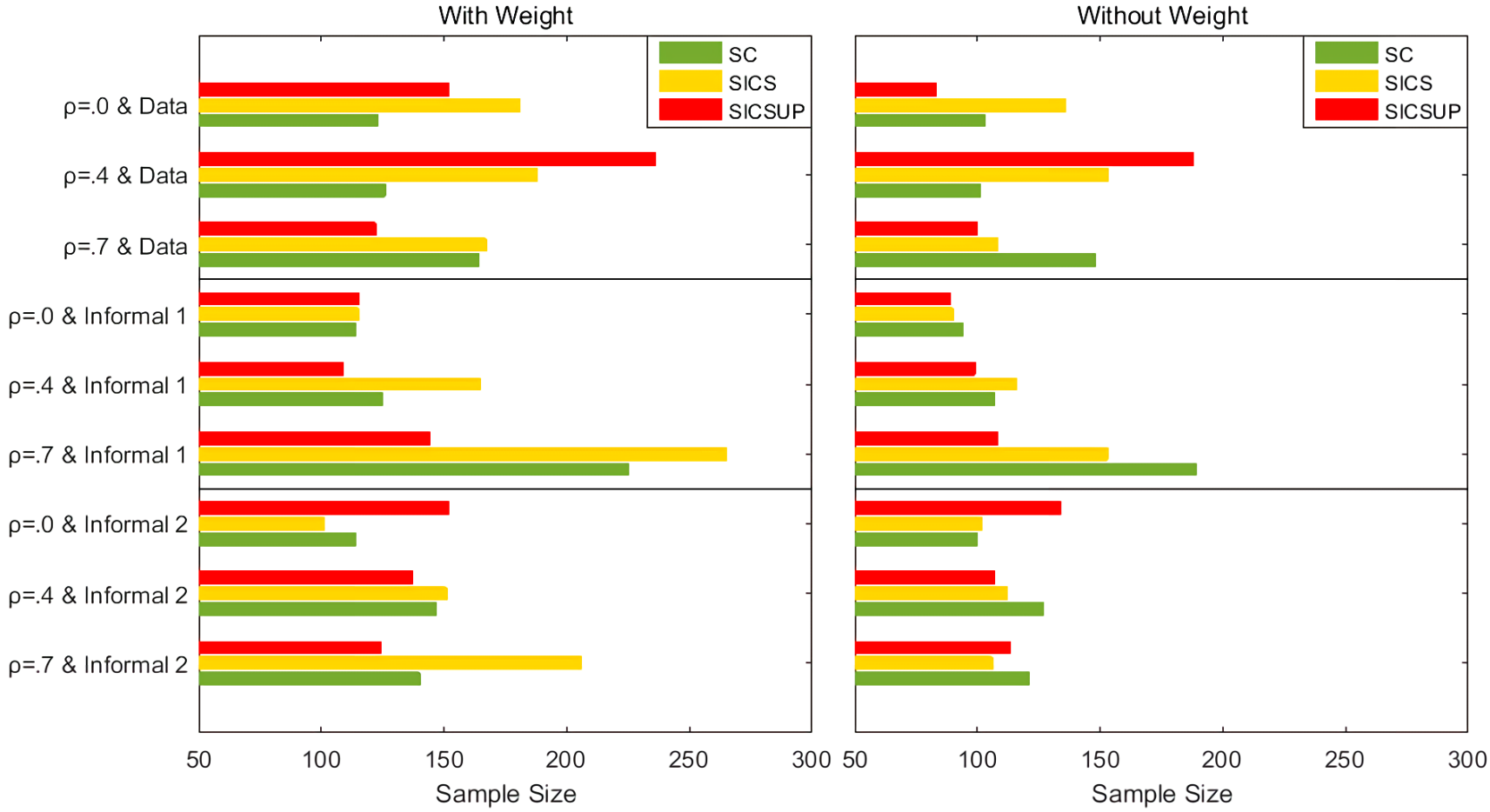


Figure 4.6 Sample Size for SICSUP, SICS, and SC That Yields the Same Precision as SRS of 50

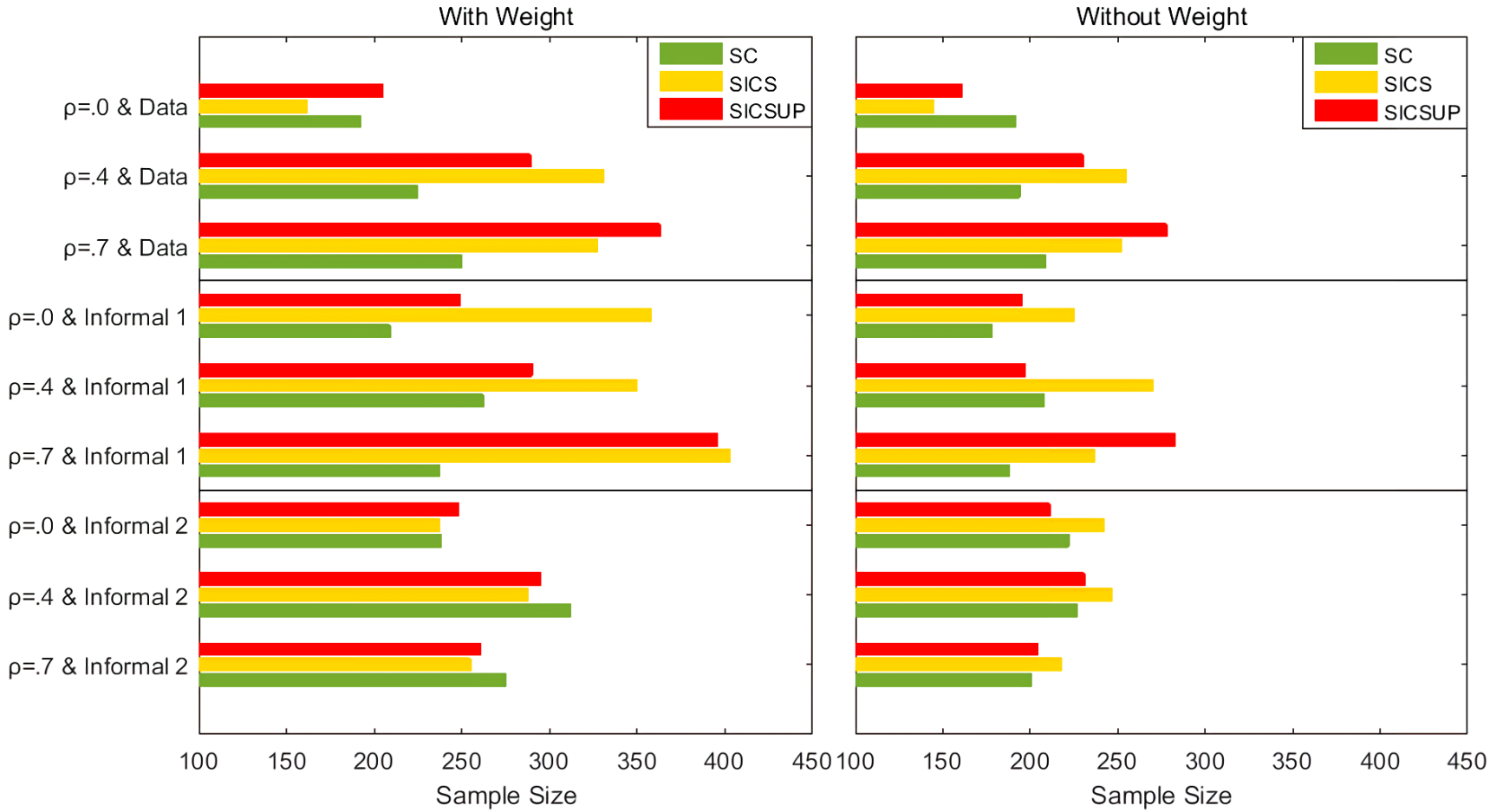


Figure 4.7 Sample Size for SICSUP, SICS, and SC That Yields the Same Precision as SRS of 100

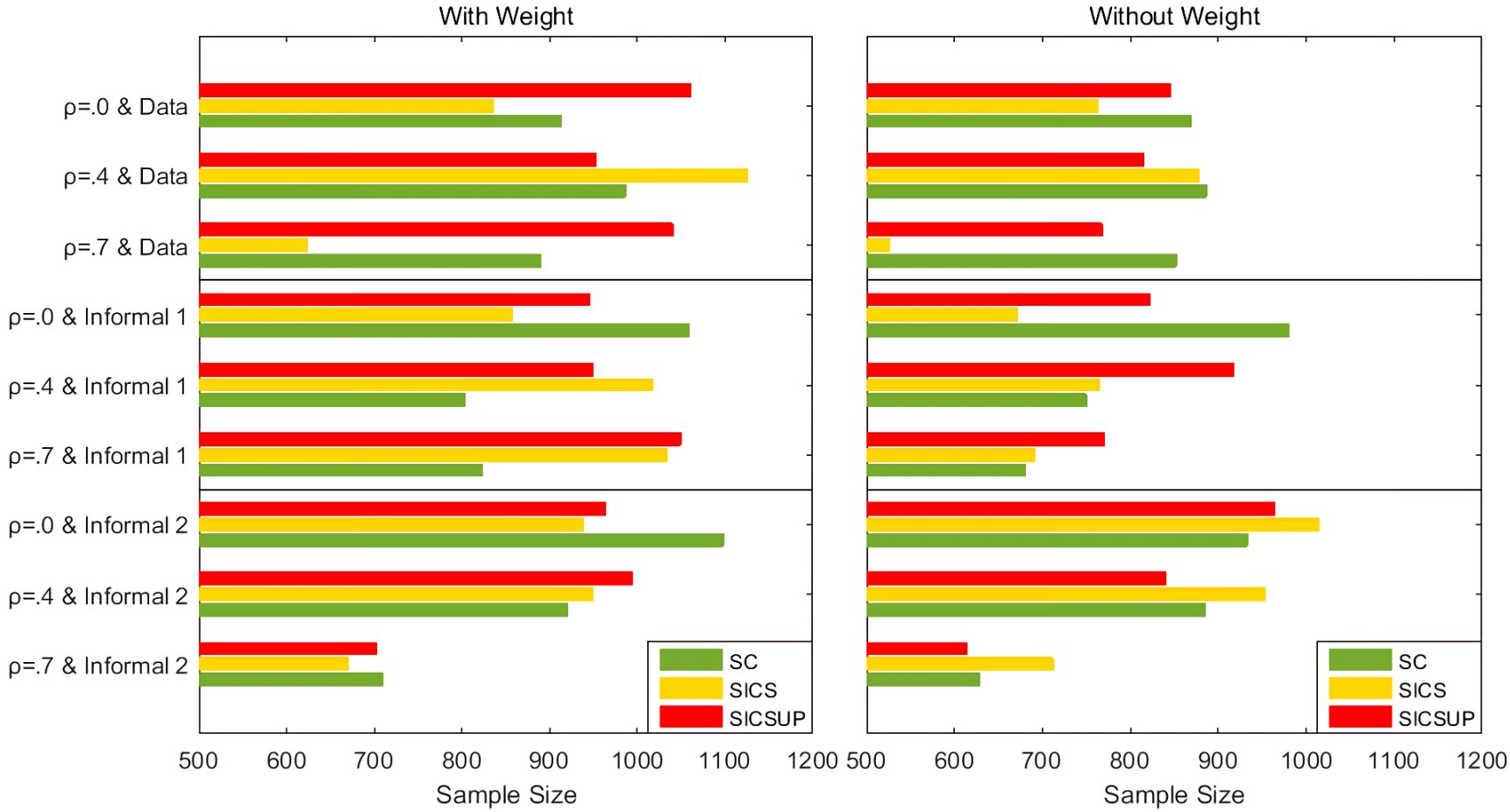


Figure 4.8 Sample Size for SICSUP, SICS, and SC That Yields the Same Precision as SRS of 500

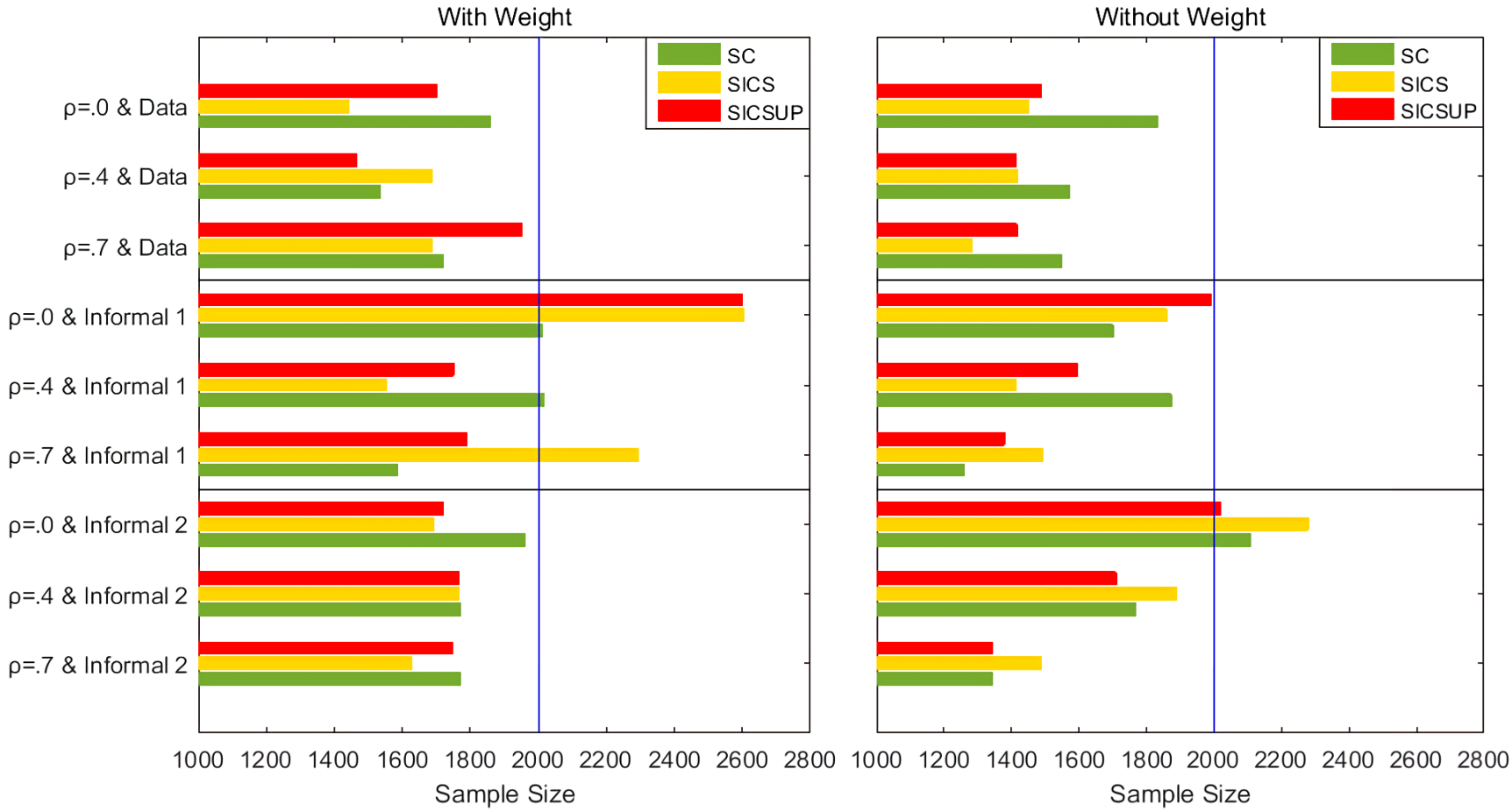


Figure 4.9 Sample Size for SICSUP, SICS, and SC That Yields the Same Precision as SRS of 1,000

4.2.2 Margin of Error and Sample Size

The margin of error refers to the limit of accuracy of a sample estimate of a population parameter (Agresti & Finlay, 2009). In other words, it shows how many points the results can be differ from the population parameter. In this research question, it is the population mean. Table 4.17 presents the required sample sizes of SRS given the level of margin of error.

Table 4.17 Margin of Error for a Sample Mean and Required Sample Size for SRS

ρ	Margin of Error				
	0.5	0.4	0.3	0.2	0.1
ρ = .0	56	87	149	307	840
ρ = .4	63	96	165	337	895
ρ = .7	69	106	180	365	943

The required sample sizes for SICSUP, SICS, and SC were obtained (see Table 4.18) by multiplying the samples sizes for SRS in Table 4.17 and the design effects in Table 4.15. The types of initial proportions of novice teachers over strata and the levels of correlation between school size and the variable of interest were averaged. In this population, with sampling weight, it seems that one cannot use SICSUP with the margin of error of .1 because the required sample size is larger than the population size of 2,000. The minimum margin of error (the maximum precision) that SICSUP can achieve is .2 in this population. Therefore, under this situation, for SICSUP, drawing 761 novice teachers is recommended if the resources such as cost and time are enough to carry out this sampling plan. If sampling weights are not used, the required sample size for SICSUP is 622 given .2 margin of error.

Table 4.18 Margin of Error for a Sample Mean and Required Sample Size for SICSUP, SICS, and SC

Margin of Error	Weighted			Unweighted		
	SICSUP	SICS	SC	SICSUP	SICS	SC
0.5	142	152	140	116	119	117
0.4	218	233	215	178	183	179
0.3	372	398	367	304	313	307
0.2	761	813	750	622	638	626
0.1	2018	2154	1992	1652	1695	1664

Figure 4.10 to Figure 4.12 illustrate the required sample sizes for SICSUP, SICS, and SC as compared to the sample sizes for SRS. In the figures, the left and right panels represent (1) weighted samples and (2) unweighted samples. The top, middle, and bottom panels represent (a) initial proportion based on data, (b) informal estimate based on school proportions, and (c) informal estimate based on equal proportions, respectively. In the figures, the dotted line in black represents the number of SRS samples that can achieve the given margin of error.

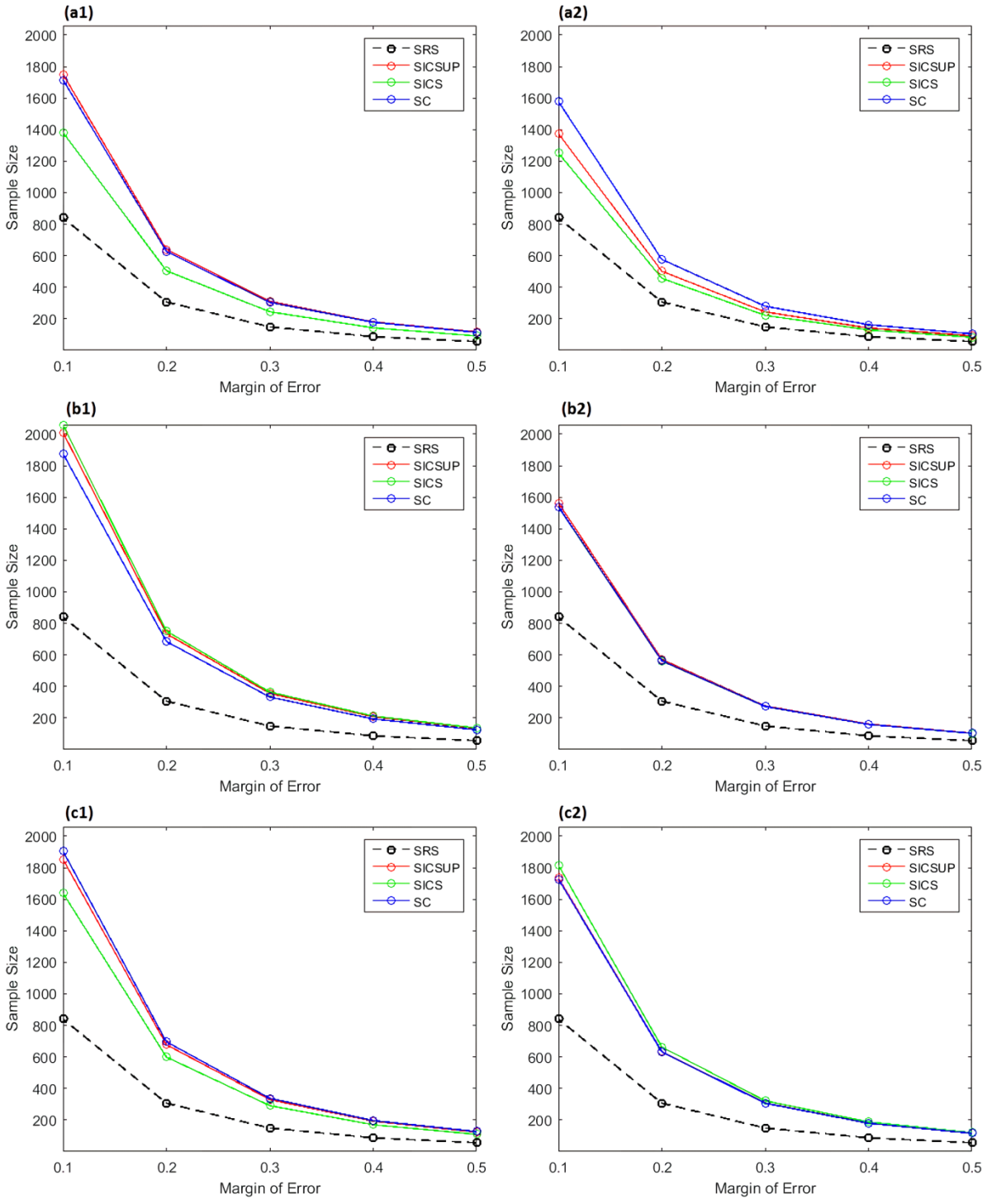


Figure 4.10 Margin of Error for a Sample Mean and Required Sample Size for SICSUP, SICS, and SC under the Condition of $\rho = .0$

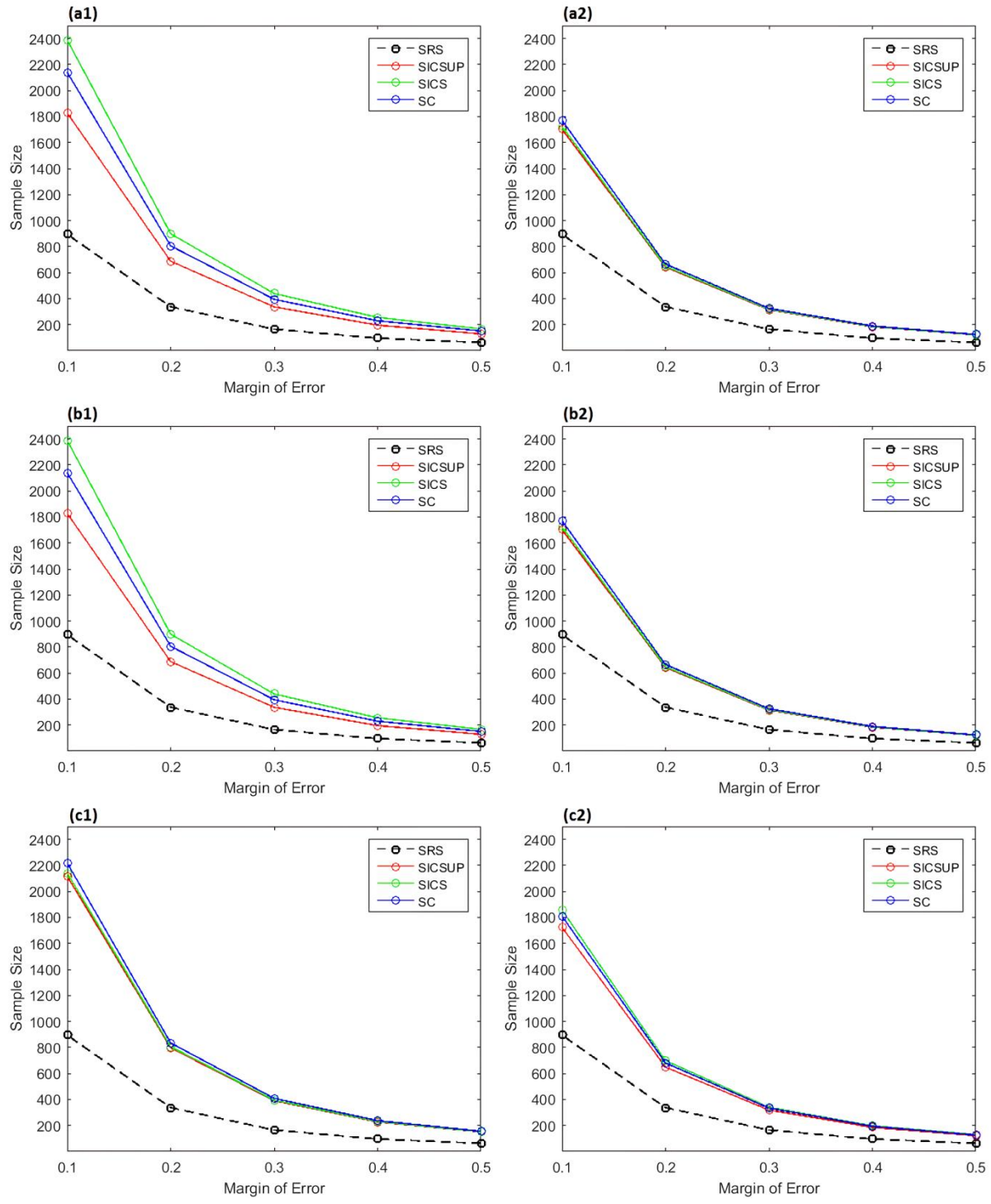


Figure 4.11 Margin of Error for a Sample Mean and Required Sample Size for SICSUP, SICS, and SC under the Condition of $\rho = .4$

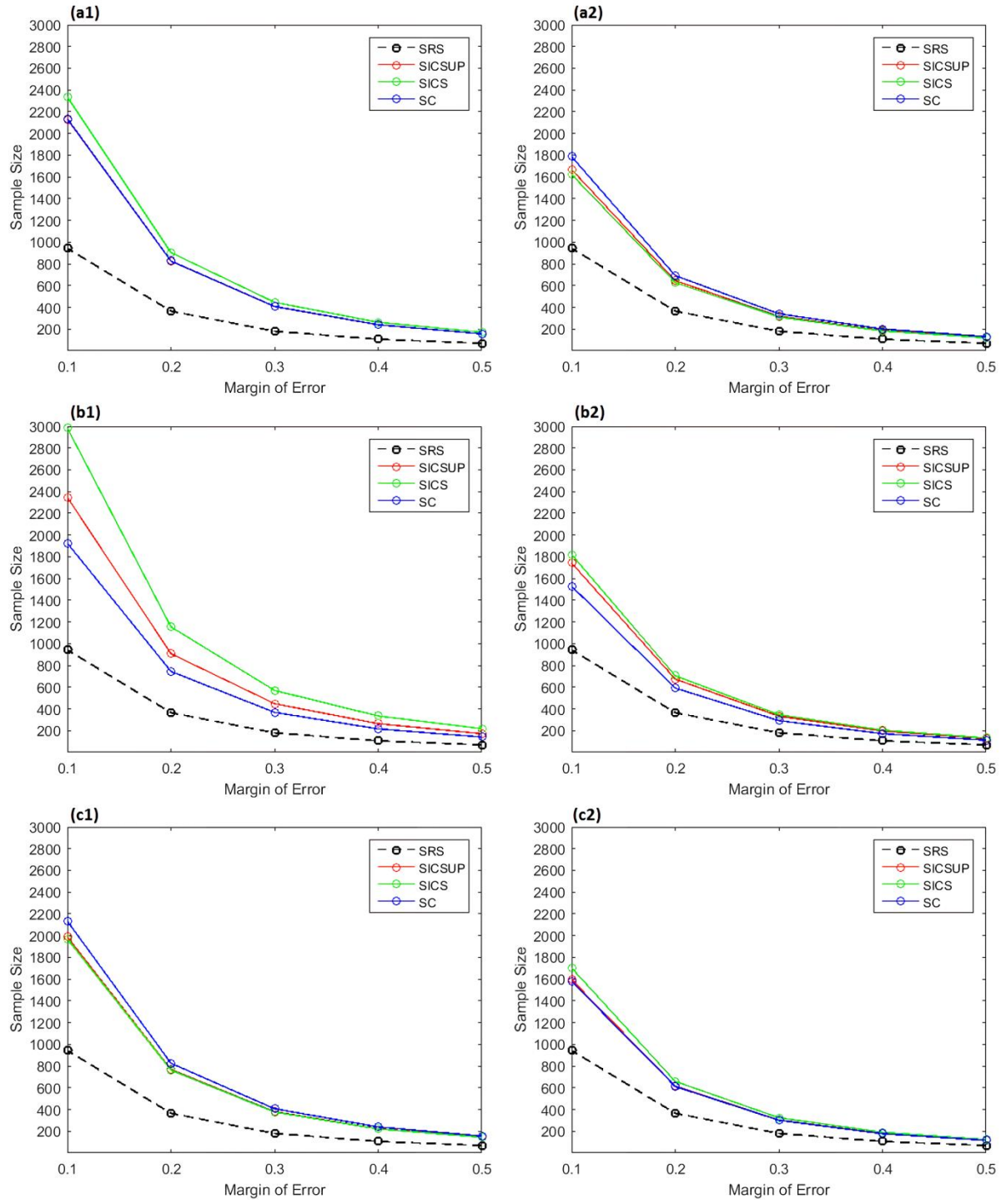


Figure 4.12 Margin of Error for a Sample Mean and Required Sample Size for SICSUP, SICS, and SC under the Condition of $\rho = .7$

All three figures reveal that as the margin of error increases, the required sample size decreases. In order to achieve the margin of error of .1, SICSUP as well as SICS and SC needs large sample sizes, close to or larger than the population size of 2,000. Therefore, it seems impossible to achieve .1 margin of error in this population. The required sample size decreases rapidly as the margin of error increases.

For some conditions, SICSUP, SICS, and SC require similar number of sample size. For example, under the condition of $\rho = .0$, informal estimate based on school proportions, and unweighted sample (b2 of Figure 4.10), the three lines are almost overlapped each other. This implies that one can use SICSUP and SC together for their survey with the same sample size, and the samples of SICSUP and SC would provide similar precision in estimating the mean.

Unlike the cases mentioned above, under some conditions, there are visible differences in sample sizes among SICSUP, SICS, and SC. Under the three conditions of $\rho = .0$, initial proportions based on data, and unweighted sample (a2 of Figure 4.10),$\rho = .4$, initial proportions based on data, and weighted sample (a1 of Figure 4.11), and $\rho = .4$, informal estimate based on school proportions, and weighted sample (b1 of Figure 4.11), SICSUP requires less samples than SC does. On the other hand, under the two conditions of $\rho = .4$, informal estimate based on school proportions, and weighted or unweighted samples (b1 and b2 of Figure 4.12), SICSUP requires more samples than SC.

In line with the findings presented thus far in this section, in order to apply SICSUP to this population of novice teachers, the sample sizes of about 760 and 620 seem the best choices with and without sampling weight, respectively, in terms of estimation precision. However, one should pay attention to the type of initial proportions of novice teachers over strata that are used for SICSUP and the correlation between school size and the variable of interest because they

may influence the expected estimation precision either of positively or negatively given the sample size.

4.3 Research Question 3

The third research question is about whether SICSUP works well in terms of estimating group difference. For each of the five selected countries, 10 sets of samples were taken in order to estimate the population mean. In addition to means, standard errors were also estimated using the jackknife estimator with original strata and the BRR estimator with pseudo-strata. This section provides the results of 95% confidence interval coverage probabilities and rankings of the five countries based on the estimated means.

As shown in Table 4.19, the three sample designs tend to estimate the mean well for all countries. It was assumed that the approximate design effects of SICSUP, SICS, and SC for the populations are less than three for all countries, so the sample size of 600 could achieve the margin of error, .3. It seems that all estimates regardless of simulation condition, sample design, and country achieve the margin of error, .3.

Table 4.19 Sample Means by Country

Country	Population mean	With Weight			Without Weight		
		SICSUP	SICS	SC	SICSUP	SICS	SC
Initial Proportions Based on Data							
Country 1	12.37	12.35	12.35	12.43	12.35	12.35	12.40
Country 2	11.31	11.23	11.21	11.18	11.33	11.32	11.29
Country 3	11.46	11.40	11.44	11.39	11.44	11.48	11.45
Country 4	11.70	11.78	11.72	11.73	11.71	11.68	11.68
Country 5	11.92	11.93	11.88	11.90	11.92	11.88	11.90
Informal Estimate Based on School Proportions							
Country 1	12.37	12.32	12.37	12.33	12.31	12.36	12.34
Country 2	11.31	11.19	11.22	11.20	11.30	11.31	11.31
Country 3	11.46	11.44	11.40	11.43	11.51	11.46	11.49
Country 4	11.70	11.76	11.74	11.74	11.70	11.71	11.69
Country 5	11.92	11.90	11.95	11.89	11.91	11.95	11.89

4.3.1 Confidence Interval Coverage Probability

Confidence interval coverage probability at a 95% confidence level was investigated (see Table 4.20 and Table 4.21). In general, SICSUP works slightly better than SC in terms of confidence interval coverage probability.

Country 1 has the highest proportion of schools with no novice teacher among the five countries, meaning the rarest population. SICSUP is supposed to work well with this type of populations, and it did for country 1. Under the condition of initial proportions based on data, the coverage probability is 1.0; under the condition of informal estimate based on school proportions, the coverage probability is still 1.0 while the corresponding coverage probabilities in SC are 1.0 and .8, respectively.

For country 2, SICSUP did not work well under the condition of informal estimate based on school proportions, with the low coverage probability of .5 with weight. Why did this happen? A possible reason is that country 2 has only two strata (public or private), and there is quite difference in stratum mean, 11.08 for novice teachers in public schools and 11.78 for those in private schools. In the population, about 67% of novice teachers are in public schools and 33% are in private schools. If the updating process of SICSUP produces adjusted proportions which are different from those in the population, SICSUP would not work well in terms of estimating mean.

With respect to type of standard error estimator, in general, the jackknife estimator performed better than the BRR estimator. In SICSUP samples, the jackknife estimator works slightly better than the BRR estimator, but the difference is not substantial. However, in SC samples, the difference in coverage probability is clearer for some countries. For instance, the coverage probabilities for country 1 under the condition of informal estimate based on school

proportions are .8 with the jackknife estimator and .3 with the BRR estimator (or .9 and .5 without weight). It seems that the BRR estimator underestimated the standard errors, so they reduced the range of 95% confidence interval and caused the low coverage probability as compared to the other.

Table 4.20 Coverage Probability of Confidence Interval for the Country Mean Using Weighted Samples

Country	Jackknife			BRR		
	SICSUP	SICS	SC	SICSUP	SICS	SC
Initial Proportions Based on Data						
1	1.0	1.0	1.0	1.0	1.0	1.0
2	0.9	0.9	0.7	0.9	0.9	0.7
3	1.0	1.0	0.9	1.0	1.0	0.9
4	0.8	0.9	0.7	0.8	0.9	0.7
5	0.9	1.0	1.0	0.9	1.0	1.0
Informal Estimate Based on School Proportions						
1	1.0	0.9	0.8	0.9	0.6	0.3
2	0.5	0.8	0.8	0.5	0.6	0.5
3	1.0	1.0	1.0	0.8	0.5	0.9
4	0.9	0.9	0.8	0.7	0.7	0.7
5	1.0	0.9	1.0	1.0	0.6	1.0

Table 4.21 Coverage Probability of Confidence Interval for the Country Mean Using Unweighted Samples

Country	Jackknife			BRR		
	SICSUP	SICS	SC	SICSUP	SICS	SC
Initial Proportions Based on Data						
1	1.0	1.0	1.0	1.0	1.0	1.0
2	1.0	0.8	0.9	1.0	0.8	0.9
3	0.9	1.0	1.0	0.9	1.0	1.0
4	0.9	1.0	1.0	0.9	1.0	1.0
5	0.9	1.0	1.0	0.9	1.0	1.0
Informal Estimate Based on School Proportions						
1	1.0	0.9	0.9	0.6	0.6	0.5
2	0.9	0.9	1.0	0.5	0.8	0.8
3	0.9	0.9	0.9	0.7	0.8	0.8
4	1.0	1.0	1.0	0.9	1.0	0.8
5	1.0	0.9	1.0	1.0	0.8	0.9

4.3.2 Rank Order of Five Countries

Table 4.22 gives the coverage probabilities of producing country rankings that are identical with the rankings based on the population means using the samples of SICSUP, SICS, SC, and the combination of SICSUP and SC. From the second to fourth columns in Table 4.22 represent that a single sample design was applied to all five countries; the last column represents that either of SICSUP or SC was applied to the five countries. Specifically, country 1, 4, and 5 drew samples using SICSUP; country 2 and 3 drew samples using SC. In the populations, country 1 has the highest mean, and county 2 has the lowest mean. The rank order of the five countries is as follows: country 1, country 5, country 4, country 3, and country 2.

As shown in Table 4.22, with weighted samples, SICSUP works as well as SC regardless of type of initial proportions used. Under the condition of initial proportions based on data, SICSUP performs slightly better than SC in terms of coverage probability: .9 for SICSUP and .8 for SC.

An interesting finding is the coverage probabilities based on the combination of two sample designs. Under the condition of informal estimate based on school proportions, the combination works as well as the cases in which a single sample design was used. On the other hand, under the condition of initial proportions based on data, the weighted samples from the combination works slightly worse than those from SICSUP: .8 for the combination and .9 for SICSUP. The coverage probability of the combination is equal to that of SC. This result suggests that SICSUP might be advantageous for all five countries in terms of estimating country rankings.

Given the weighted samples, having informal estimate of proportions of novice teachers over strata at the beginning of the sampling procedure does not cause a significant impact upon the coverage probability of rankings as compared to the coverage probability of population means. Under that condition, the difference in coverage probability of rankings between SICSUP and SC is smaller than that of population means. In terms of differentiating countries, SICSUP works as well as SC under that condition. Thus, the combination of SICSUP and SC produced the identical coverage probability with the cases in which SICSUP or SC was used alone.

Table 4.22 Rates of Producing Rankings That Are Identical with the Rankings Based on the Population Means Using SICSUP, SICS, SC, and the Combination of Two Designs

Weight	SICSUP	SICS	SC	Combination
Initial Proportions Based on Data				
With Weight	0.9	1	0.8	0.8
Without Weight	0.7	0.9	0.9	0.8
Informal Estimate Based on School Proportions				
With Weight	0.9	0.9	0.9	0.9
Without Weight	0.9	0.9	0.9	0.9

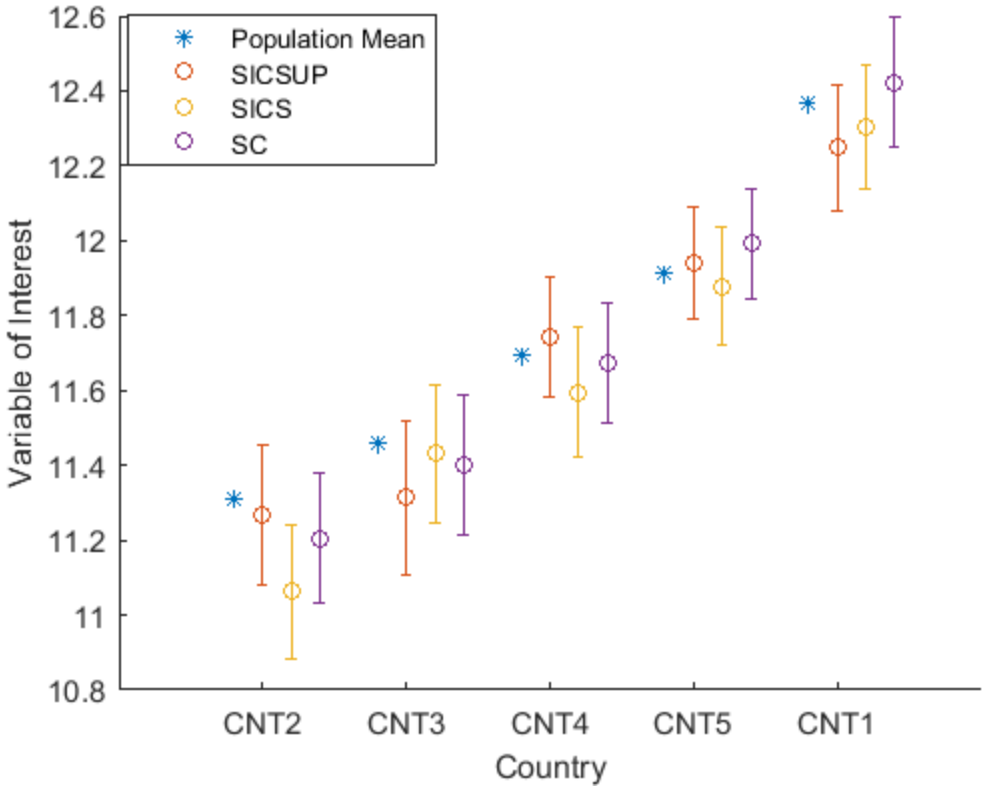


Figure 4.13 Estimated Means with 95% Confidence Interval by Country under the Condition of Initial Proportions Based on Data and with Weight: The First Scenario

Figure 4.13 illustrates the population and sample means using the three sample designs, with 95% confidence intervals under the condition of initial proportions based on data and weighted samples. The jackknife estimator was used to compute standard errors. For all three sample designs, each population mean falls into 95% confidence intervals of the estimates except a single case (SICS samples in country 2). In addition, the three sample designs provide the rankings that are identical with the rankings based on the population means. In this case, using either of SICSUP or SC does not make any difference in rankings of the countries. This illustrates the best scenario for all three sample designs and the combination with respect to coverage of the population means and country rankings.

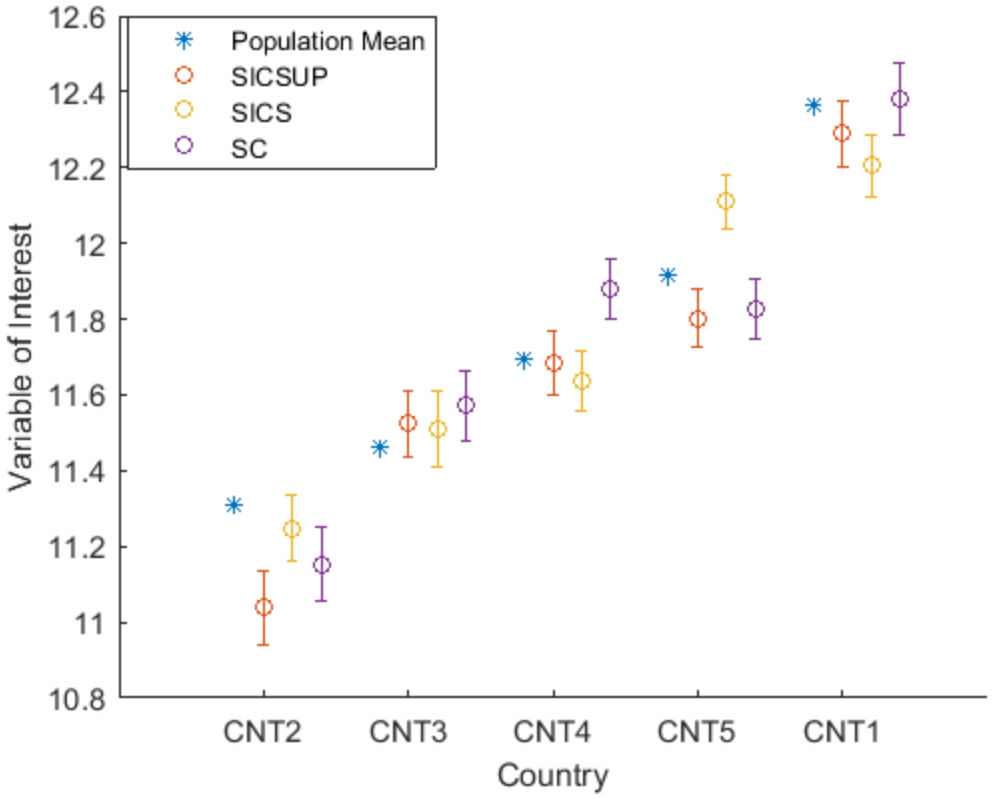


Figure 4.14 Estimated Means with 95% Confidence Interval by Country under the Condition of Informal Estimate of Proportions Based on School Proportion and with Weight: The Second Scenario

Figure 4.14 shows the sample means with 95% confidence intervals based on another set of samples, and the jackknife estimator was used to compute standard errors. The results here are quite different from those in the first scenario. For some countries, there are cases that the population means do not fall into the 95% confidence interval of the sample mean, which indicates hypothesis testing would reject the null hypothesis at a 95% confidence level. For example, the sample means using SICSUP samples are significantly different from the parameters for country 5 and 4. The sample means using SC samples are significantly different from the parameters for most of the countries including country 2 to 5.

It is interesting to observe that the rankings of countries based on SICSUP samples are identical with those based on the population means even though some of the sample means are not very accurate. That is not the case when SC was used. For the samples of SC, the rankings are different from those based on the population means. The combination of SICSUP and SC provides the rankings that are identical with those based on the population means.

These results imply that rankings should be interpreted with caution although they are frequently reported as results of national or international surveys and assessments. For example, when SICSUP is used, the sample means of country 4 and 5 are not statistically different from each other at a 95% confidence level. However, their rank order positions are different. County 5 is ranked higher than country 4.

To sum, although there are some limitations, and the results should be interpreted cautiously, in this scenario, SICSUP performs better than SC with respect to coverage of the population means and country rankings.

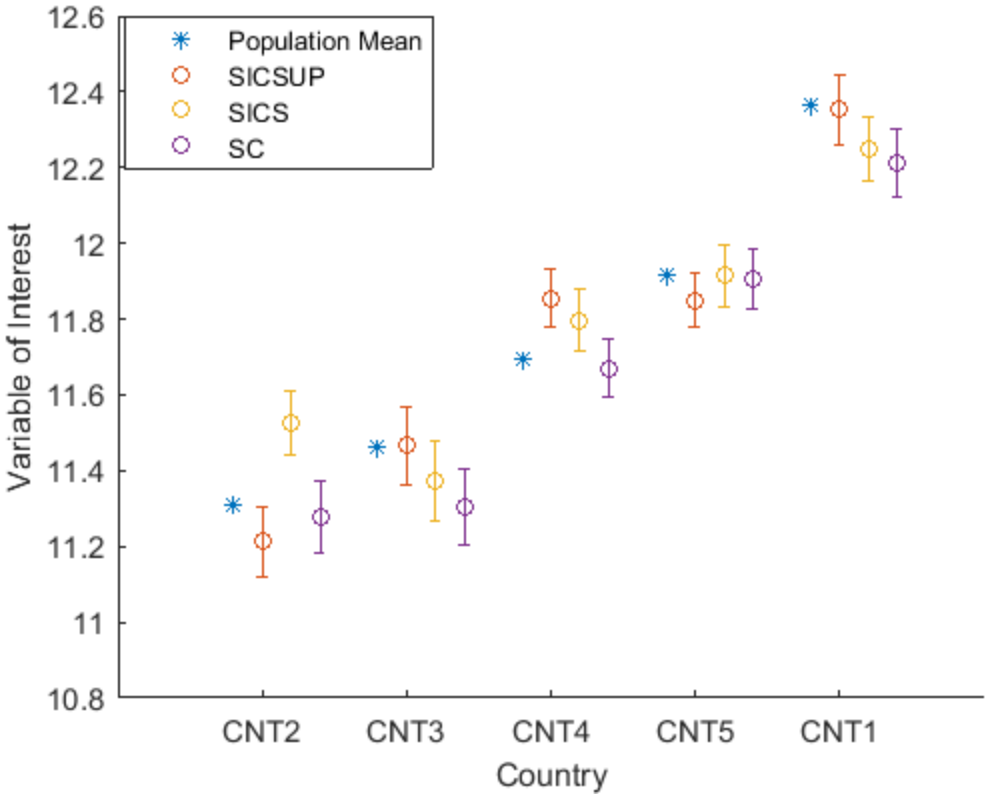


Figure 4.15 Estimated Means with 95% Confidence Interval by Country under the Condition of Informal Estimate of Proportions Based on School Proportion: The Third Scenario

The results shown in Figure 4.15 illustrate another scenario that the sample designs do not work well with the five countries. The simulation conditions here are exactly the same as those in the second scenario. SICSUP here works slightly better than in the second scenario (see the red circles and lines Figure 4.14). Only the population mean of country 4 does not fall into the 95% confidence interval of the sample mean. SC here also works slightly better than those in the second scenario. In two out of the five counties (country 1 and 3), the populations means do not fall into the 95% confidence intervals of the sample means.

How about the rankings of the countries in this scenario? The rankings based on the SICSUP samples are not identical with those based on the population means. That is the same for

the combination of SICSUP and SC. On the other hand, the rankings based on SC are identical with those based on the population means. If one focuses on the sample means for country 4 and 5 based on the SICSUP samples (see the red circles and lines in Figure 4.15), they are almost equal to each other: 11.86 for country 4 and 11.85 for country 5. After rounding the sample means to the nearest tenth, they become identical. The two countries might be the same in rank depending on the decimal places reported. Because SICSUP is applied to country 4 and 5 when the combination of two sample designs is used, the rankings under this combination are also not identical with those based on the population means.

This scenario shows that although SICSUP performs slightly better than SC with respect to coverage of the population means, it does not work as well as SC with respect to country rankings.

Given the results presented thus far in this section, SICSUP functions as well as, or, depending on the condition, slightly better than, SC in the rare populations across the five countries with respect to coverage of the population means. That is the same with respect to coverage of country rankings. However, the three scenarios mentioned in this section suggest that country rankings should be interpreted with caution.

4.4 Research Question 4

The last research question evaluates the economic aspect of SICSUP, and comparisons of SICSUP with SICS and SC were made on the basis of the number of contacted schools during the sampling procedure and the number of schools in the final set of samples. The results in this section are based on the 500 replications.

4.4.1 Results Based on Dataset 1

Table 4.23 gives the numbers of contacted schools during the sampling procedure (n*) and schools in the final set of samples (n) by sample size. For all sample sizes, SICSUP contacted fewer schools than SICS and SC did. The numbers of schools in the final set of samples are similar across the three sample designs. Therefore, the ratio of the schools in the final set of samples to the number of contacted schools in SICSUP is higher than those in SICS and SC, meaning SICSUP is more economical than these two sample designs. In SICSUP, 77% of contacted schools were added to the final set of schools, 76% in SICS and 72% in SC. These show that both of the updating process and sequential selection have a positive effect on the reduction in the number of contacted schools in this population.

Table 4.23 Number of Contacted Schools and Schools in the Sample, Based on Dataset 1

Sample Size	SICSUP			SICS			SC		
	n*	n	$\frac{n}{n^*}$	n*	n	$\frac{n}{n^*}$	n*	n	$\frac{n}{n^*}$
50	28.78	22.11	0.77	30.07	22.79	0.76	33.84	24.39	0.72
100	55.52	42.65	0.77	58.58	44.45	0.76	65.99	47.67	0.72
500	269.06	207.35	0.77	286.87	217.68	0.76	322.40	233.67	0.73
1000	536.52	413.39	0.77	572.02	434.08	0.76	636.88	464.02	0.73

Table 4.24 illustrates the difference in the number of contacted schools by using the ratio of two sample designs. The second column of Table 4.24 presents the effect of updating process and sequential selection on the number of contacted schools; the third column for the effect of the updating process, and the last column for the effect of sequential selection. Small ratio values indicate large effects on the number of contacted schools, meaning the sample design in the top of the ratio is more beneficial than the sample design in the bottom of the ratio in terms of economic aspect. Ratio equal to 1 indicates no effect on the number of contacted schools.

The combination of the updating process and sequential selection is the most effective in the reduction of the number of contacted schools, with the ratio of about .85. However, the third column of Table 4.24 suggests that this effect might be mostly due to the sequential selection rather than the updating process. The ratio in this column is close to 1, meaning that the updating process reduced the number of contacted schools only slightly. The last column shows that the sequential selection reduced the number of contacted schools by 10%. The ratios tend to be constant across the different sample sizes.

Table 4.24 Difference in the Number of Contacted Schools, Based on Dataset 1

Sample Size	$\frac{n^*_{SICSUP}}{n^*_{SC}}$	$\frac{n^*_{SICSUP}}{n^*_{SICS}}$	$\frac{n^*_{SICS}}{n^*_{SC}}$
50	0.85	0.96	0.89
100	0.84	0.95	0.89
500	0.84	0.94	0.89
1000	0.84	0.94	0.90

Table 4.25 Number of Contacted Schools and Schools in the Sample by Strata, Based on Dataset 1

m	Stra.	SICSUP			SICS			SC		
		n*	n	$\frac{n}{n^*}$	n*	n	$\frac{n}{n^*}$	n*	n	$\frac{n}{n^*}$
50	1	9.26	6.41	0.77	11.36	7.86	0.76	13.23	8.42	0.72
	2	8.66	6.69	0.76	9.34	7.20	0.77	10.33	7.65	0.73
	3	10.87	9.01	0.76	9.31	7.70	0.77	10.28	8.31	0.73
100	1	17.23	11.97	0.76	22.38	15.56	0.76	25.88	16.67	0.74
	2	16.18	12.41	0.76	18.01	13.91	0.76	20.23	14.92	0.73
	3	22.11	18.26	0.76	18.14	14.99	0.76	19.88	16.08	0.73
500	1	80.89	56.16	0.76	110.17	76.66	0.76	127.87	82.47	0.73
	2	77.20	59.51	0.76	88.30	68.12	0.76	98.57	73.16	0.74
	3	110.96	91.68	0.76	88.14	72.85	0.76	95.97	78.04	0.73
1000	1	161.24	111.78	0.76	220.48	153.11	0.76	249.86	163.13	0.73
	2	153.05	117.88	0.76	175.83	135.57	0.76	196.38	145.68	0.73
	3	222.23	183.73	0.76	175.99	145.43	0.76	190.63	155.22	0.74

Table 4.25 gives the numbers of contacted schools during the sampling procedure and schools in the final set of samples by strata. Dataset 1 uses location of school, such as (1) rural, (2) town, and (3) city, as stratification. Rural schools contain the smallest number of novice teachers (20%), and city schools contain the largest number of novice teachers (55%). The ratios of schools in the final set of samples to the contacted schools are constant across different sample sizes. In SICSUP, stratum 3 (city) contacted the largest number of schools because this stratum tended to have largest sample sizes as compared to the other strata. Stratum 1 (rural) seems to contact more schools than stratum 2 (town) although the proportion of stratum 2 (25%) is slightly larger than stratum 1 (20%). This is due to the large number of small schools in rural area (stratum 1).

Unlike SICSUP, stratum 3 did not contacted more schools than the other strata in SICS and SC. For all sample sizes, the first stratum (rural) contacted more schools than the other

strata. If drawing novice teachers in rural schools is more expensive than that in town and city schools, the larger number of contacted schools in this stratum might increase the resource consumption in SICS and SC.

Some interesting results are found in Table 4.26. The ratios are quite different between strata. In rural area, the effect of the updating process and sequential selection is substantial, reducing the number of contacted schools by about 30% in SICSUP as compared to SC. However, that is not the same in city. The combination of updating process and sequential selection caused a negative effect, and SICSUP contacted more schools than SC did. The ratios are greater than 1. Distance between schools in rural tends to be greater than that in city, and this may increase the cost for sampling in rural area. If SICSUP requires fewer contacted schools than SC in order to reach the predetermined sample size of elements especially in rural area, SICSUP might significantly reduce the cost for sampling as compared to SC.

The similar pattern is observed in the fourth column of Table 4.26. In city, there are more large schools than in rural or town. In other words, average school size in city is larger than that in rural or town. These results suggest that SICSUP might not have advantages with large schools or clusters.

Table 4.26 Difference in the Number of Contacted Schools by Strata, Based on Dataset 1

Sample Size	Location of School	$\frac{n^*_{SICSUP}}{n^*_{SC}}$	$\frac{n^*_{SICSUP}}{n^*_{SICS}}$	$\frac{n^*_{SICS}}{n^*_{SC}}$
50	Rural	0.73	0.85	0.86
	Town	0.85	0.94	0.90
	City	1.09	1.20	0.91
100	Rural	0.70	0.80	0.86
	Town	0.81	0.91	0.89
	City	1.15	1.27	0.91
500	Rural	0.66	0.77	0.86
	Town	0.79	0.89	0.90
	City	1.21	1.31	0.92
1000	Rural	0.67	0.77	0.88
	Town	0.79	0.88	0.90
	City	1.22	1.32	0.92

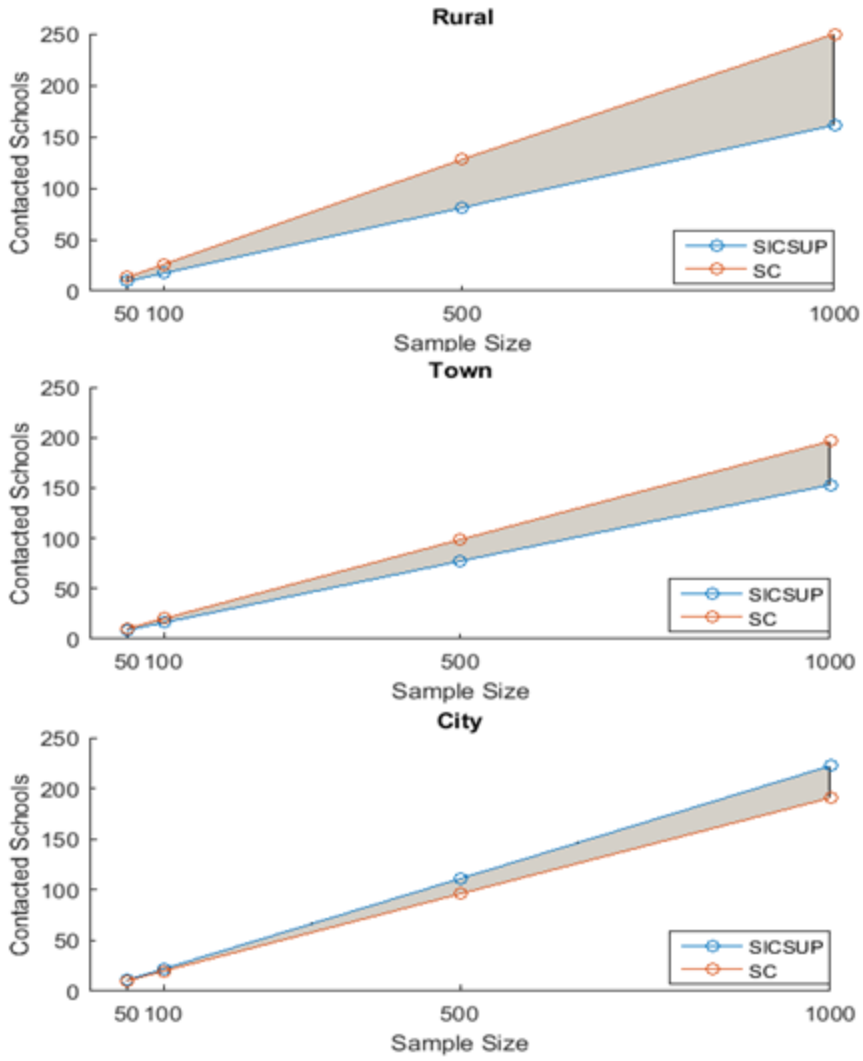


Figure 4.16 Difference in the Number of Contacted Schools by Strata, Based on the Dataset 1

Figure 4.16 illustrates the difference in the number of contacted schools by strata. The blue circles refer to the number of contacted schools in SICSUP and the red circles refer to the number of contacted schools in SC. The gray area between the two lines represents the difference in the number of contacted schools between SICSUP and SC. As the sample size increases, the difference becomes greater, showing SC contacted many more schools than SICSUP did.

However, the bottom panel of Figure 4.16 shows the opposite pattern. As the sample size increases, SICSUP needed more schools to contact than SC did.

With respect to the amount of difference, expressed by gray area, the effect of the updating process and sequential selection on the number of contacted schools is greatest in rural area.

4.4.2 Results Based on Dataset 2

Given the sample size of 600, the numbers of contacted schools during the sampling procedure by country are reported in Table 4.27. Country 1 has the rarest population, meaning a large portion of schools (about 60%) does not have any novice teacher. Because of this fact, country 1 contacted many more schools than the other countries, and the ratio of schools in the final set of samples to contacted schools is very low, about 40% in SICSUP. This means that more than a half of the contacted schools were discarded and researchers should keep contacting schools in order to achieve the predetermined sample size of novice teachers. Country 2 also has a fairly small portion of novice teachers in the general population, and about 30% of schools do not contain any novice teacher. This leads country 2 to have the second-worst ratio among the five countries, .68, .68, .64 for SICSUP, SICS, and SC, respectively. In SICSUP, only 68% of the contacted schools were added to the final set of samples. Country 3 to 5 have relatively high proportions of novice teachers in the general population, and around 77% of schools include at least one novice teacher. Therefore, the ratios for these three countries are higher than country 1 and 2. In country 3 to 5, more than 70% of the contacted schools were added to the final set of samples.

Table 4.27 Number of Contacted Schools and Schools in the Sample, Based on Dataset 2

CNT	SICSUP			SICS			SC		
	n*	n	$\frac{n}{n^*}$	n*	n	$\frac{n}{n^*}$	n*	n	$\frac{n}{n^*}$
1	974.94	383.76	0.39	981.84	389.85	0.40	1270.37	413.96	0.33
2	374.11	255.03	0.68	373.83	255.16	0.68	437.32	280.45	0.64
3	325.03	241.99	0.74	336.94	249.35	0.74	380.66	267.76	0.70
4	298.06	228.18	0.77	311.10	235.41	0.76	352.03	251.93	0.72
5	261.90	206.06	0.79	276.27	211.17	0.77	308.23	223.76	0.73

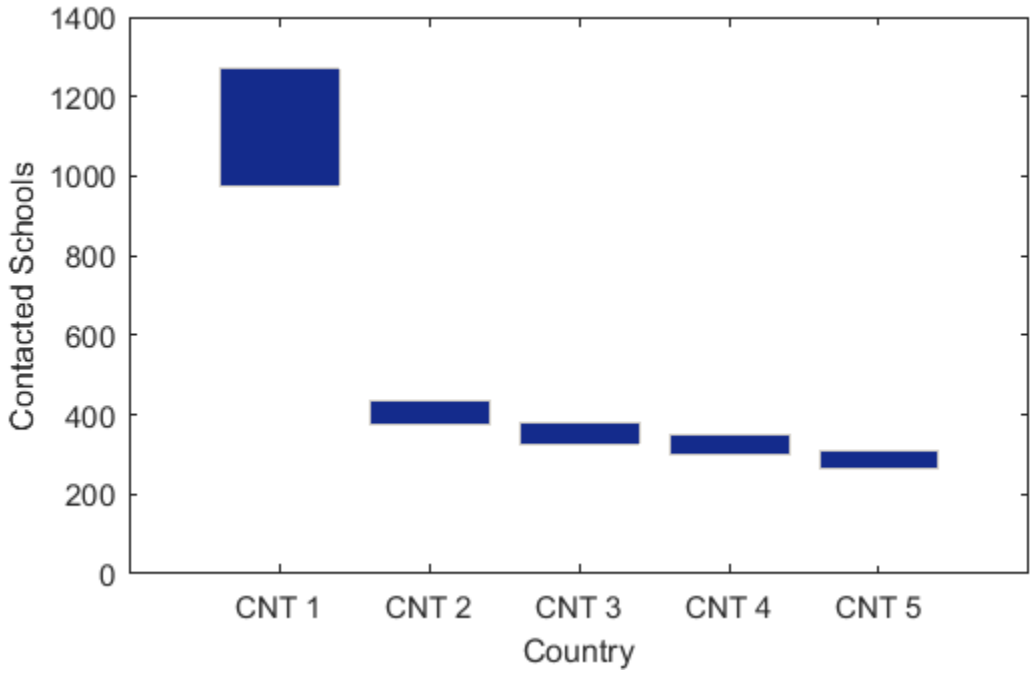


Figure 4.17 Difference in the Number of Contacted Schools by Country: SC (Top Line) and SICSUP (Bottom Line)

Figure 4.17 describes the difference in the number of contacted schools by the five countries. For each box, the top line represents the number of contacted schools in SICSUP, and the bottom line represents those in SC. Country 1 (CNT1) shows the biggest difference in the number of contacted schools between SICSUP and SC as compared to the other countries does.

Because of the high proportion of "blank" schools, country 1 had to contact many more schools than the other countries did.

For the five countries, the updating process of SICSUP seems not very beneficial while the sequential process is fairly advantageous (see Table 4.28). The sequential selection reduced the number of contacted schools by 10 to 20% depending on the country (see the last column in Table 4.28). The third column describes the effect of the updating process on the number of contacted schools, and the values are very close to 1, meaning no effect. For some countries, the updating process was not helpful to reduce the number of contacted schools; for the other countries, the updating process worked differently for each stratum, and the effects were canceled out when it came to the whole sample.

Table 4.28 Difference in the Number of Contacted Schools, Based on Dataset 2

Country	$\frac{n^*_{SICSUP}}{n^*_{SC}}$	$\frac{n^*_{SICSUP}}{n^*_{SICS}}$	$\frac{n^*_{SICS}}{n^*_{SC}}$
Country 1	0.77	0.99	0.77
Country 2	0.86	1.00	0.85
Country 3	0.86	0.97	0.89
Country 4	0.85	0.96	0.88
Country 5	0.85	0.95	0.90

Table 4.29 and Table 4.30 provide detailed descriptions of what happened within each stratum in each country. In country 1 (CNT1) with SICSUP, more than a half of the contacted schools in stratum 1 to 3 were discarded because they were "blank" schools, meaning schools with no novice teacher. There were less "blank" schools in stratum 4, and about a half of the contacted schools were added to the final set of samples. Despite of these results, SICSUP is still more economic than SC in country 1. SC discarded more schools than SICSUP did. The ratios in Table 4.29 show that SICSUP is more economical than SC in country 1.

Each of country 4 (CNT4) and 5 (CNT5) have relatively small stratum as compared to others. For example, stratum 3 and 4 have very small proportions in country 4, and most of the schools contain at least one novice teacher, meaning that novice teachers are not rare in such stratum. The ratios in the three designs are almost identical with each other. In this situation, sequential process and the updating process do not have substantial impact upon the reduction in the number of contacted schools. This suggests that the updating process and sequential selection is effective for rare populations, in which a large portion of clusters does not satisfy the selection criterion.

Table 4.29 Number of Contacted Schools and Schools in the Sample by Strata, Based on Dataset 2

CNT.		SICSUP			SICS			SC		
	St.	n*	n	$\frac{n}{n^*}$	n*	n	$\frac{n}{n^*}$	n*	n	$\frac{n}{n^*}$
CNT1	1	61.78	20.74	0.34	45.79	15.26	0.33	61.54	17.57	0.29
	2	396.88	163.97	0.41	415.75	172.22	0.41	534.16	181.76	0.34
	3	369.69	120.76	0.33	366.49	119.98	0.33	485.79	129.50	0.27
	4	146.59	78.29	0.53	153.73	81.92	0.53	188.87	85.14	0.45
CNT2	1	261.05	171.43	0.66	263.17	173.03	0.66	310.52	190.78	0.61
	2	113.06	83.59	0.74	112.17	82.81	0.74	126.81	89.67	0.71
CNT3	1	64.58	42.87	0.66	77.25	51.37	0.67	89.95	55.18	0.61
	2	101.59	74.88	0.74	114.74	84.70	0.74	131.13	90.33	0.69
	3	158.85	124.25	0.78	144.03	112.83	0.78	159.57	122.26	0.77
CNT4	1	41.04	36.52	0.89	36.73	32.48	0.88	39.01	34.34	0.88
	2	107.08	75.41	0.70	114.94	80.75	0.70	132.53	87.27	0.66
	3	13.23	13.23	1.00	9.71	9.71	1.00	10.10	10.08	1.00
	4	119.26	87.28	0.73	135.46	98.90	0.73	154.16	105.74	0.69
	5	17.44	15.74	0.90	15.42	13.85	0.90	16.23	14.51	0.89
CNT5	1	43.99	21.82	0.50	60.28	30.10	0.50	74.91	32.21	0.43
	2	55.88	39.12	0.70	59.50	41.70	0.70	68.39	45.74	0.67
	3	69.62	57.43	0.82	67.68	55.81	0.82	73.13	59.52	0.81
	4	24.60	24.41	0.99	22.43	22.24	0.99	23.32	22.90	0.98
	5	67.81	63.28	0.93	65.40	61.10	0.93	68.49	63.39	0.93

As shown in Table 4.30, in country 2 (CNT2), the number of contacted schools in SICSUP is very similar with that in SICS with the ratios close to 1. This simply shows that the updating process did not work well because there are only two strata in country 2. The updating process is based on the proportions of novice teachers over strata. If there are only two strata, the updated proportions may not make many changes to the initial sampling plan. Country 3 to 5 show the similar patterns. For some strata, the updating process was effective in reducing the number of contacted schools, showing small ratios, while, for the others, it was not very effective, showing large ratios (see the fourth column in Table 4.30). These differences disappear when they are combined into the whole set of samples (see the third column in Table 4.28).

Table 4.30 Difference in the Number of Contacted Schools by Strata, Based on Dataset 2

Country	Strata	$\frac{n^*_{SICSUP}}{n^*_{SC}}$	$\frac{n^*_{SICSUP}}{n^*_{SICS}}$	$\frac{n^*_{SICS}}{n^*_{SC}}$
Country 1	1	1.00	1.35	0.74
	2	0.74	0.95	0.78
	3	0.76	1.01	0.75
	4	0.78	0.95	0.81
Country 2	1	0.84	0.99	0.85
	2	0.89	1.01	0.88
Country 3	1	0.72	0.84	0.86
	2	0.77	0.89	0.87
	3	1.00	1.10	0.90
Country 4	1	1.05	1.12	0.94
	2	0.81	0.93	0.87
	3	1.31	1.36	0.97
	4	0.77	0.88	0.88
	5	1.07	1.13	0.95
Country 5	1	0.59	0.73	0.80
	2	0.82	0.94	0.87
	3	0.95	1.03	0.93
	4	1.05	1.10	0.96
	5	0.99	1.04	0.95

4.4.3 Probability of Using Substitute Schools in SC

In addition to the three evaluation criteria including the number of contacted schools, ratio of schools in the final set of samples to the contacted schools, and ratio of the contacted schools between sample designs, the probability of using substitute schools in SC was investigated. When cluster or multi-stage sample design is employed, surveys often prepare substitute or replacement clusters in advance. For example, the replacement schools in the PISA are the two neighboring schools of the initially sampled school in the sampling frame (OECD, 2017). These replacement schools are majorly due to non-response. In rare populations, substitute schools are required because sampling units are hard to locate, and there is a high proportion of "blank" clusters in such populations.

Timing is another important economic aspect when selecting a sample design for a survey. Usually surveys have strict closeout dates and publication deadlines. If multiple sets of substitute schools are necessary to reach the fixed sample size, it may take a long time and delay the plan of the survey. This can be considered a disadvantage of SC over SICSUP and SICS. Table *4.31* and Table 4.32 report the probability of using substitute schools in SC. As expected, in general, more than 80% of sets of samples used substitute schools in order to reach the predetermined sample size. In Table 4.32, for some countries, such as country 1, 4, and 5, almost all sets of samples needed substitute schools.

Although these results do not directly provide evidence that SICSUP is economically advantageous over SC, these suggest applications of alternative sample design instead of SC in such rare populations, such as SICSUP.

Table 4.31 Probability of Using Substitute Schools, Based on Dataset 1

Sample Size	50	100	500	1000
Probability of Using Substitute Schools	0.83	0.85	0.87	0.85

Table 4.32 Probability of Using Substitute Schools, Based on Dataset 2

Sample Size	CNT 1	CNT 2	CNT 3	CNT 4	CNT 5
Probability of Using Substitute Schools	0.94	0.78	0.86	0.97	0.96

To sum, in general, SICSUP requires smaller number of contacted schools during the sampling procedure in order to reach the predetermined sample size than SICS and SC do due to the updating process and sequential selection. This suggests that SICSUP may be beneficial for rare populations in terms of economic aspect as compared to SICS and SC.

CHAPTER 5.

CONCLUSION AND DISCUSSION

5.1 Summary of Findings

The aim of this dissertation was twofold. Firstly, it attempted to investigate the performance of stratified inverse cluster sampling with updating process (SICSUP) as compared to that of stratified cluster sampling (SC) with respect to statistical and economic aspects. The comparison was made because SICSUP was expected to serve as an alternative to SC for rare populations in education. Secondly, it was an attempt to provide guidelines for applying SICSUP to rare populations.

Based on these aims, the research questions were the following:

1. Does SICSUP work as well as SC regarding parameter estimation?
2. How can the appropriate sample size for SICSUP be determined?
3. Can the samples from SICSUP determine whether the means of groups are different from each other?
4. Is SICSUP more economic than SC?

The first to third research questions evaluated the statistical aspect of SICSUP, and the last research question evaluated the economic aspect. From the simulation studies, four key findings were drawn.

First, the results of simulation studies in Research Question 1examined the performance of SICSUP with respect to the level of precision in estimating the population mean, the population standard deviation, and standard error of the sample mean as compared to that of SC.

In terms of mean and standard deviation estimation, SICSUP worked as well as SC when sample size was not very small ($n \geq 100$). SICSUP worked worse than SC under the condition of

very small sample size (n = 50) and initial proportions of novice teachers based on data. However, if informal estimates of proportions were used, SICSUP performed better than SC even though the sample size was small. Although the updating process with small sample size is not very helpful for estimating parameters, if researchers do not know the proportions of novice teachers over strata in the population, the updating process at least provides some useful information about the proportions.

In terms of standard error estimation, in general, SICSUP performed as well as SC except with very small sample size (n = 50). With n = 50, SICSUP worked better or worse than SC depending on the evaluation criteria and type of strata used.

The results of the simulation studies showed that the jackknife, bootstrap, BRR, and Fay's estimators provided similar standard errors in SICSUP. The difference in standard errors among the four standard error estimators was not substantial. If one wants to choose one of them, the choice of standard error estimator in SICSUP would depend on the type of strata used for standard error estimators. When original strata were used, the jackknife estimator was slightly better than the bootstrap estimator with very small sample size (n = 50). When pseudo-strata were used, the BRR worked slightly better than the others when informal estimate based on equal proportions was used. However, the difference among the four standard error estimators was not great.

Second, the simulation studies in Research Question 2 suggested some guidelines for sample-size determination for SICSUP. On average, the design effects based on the weighted samples were around 2.30 and 2.21 in SICSUP and SC, respectively. The design effects based on the samples without weight were around 1.86 and 1.89 in SICSUP and SC, respectively. These results indicated that the desired sample size in SICSUP was 2.30 times larger than that in SRS

with weight, or 1.86 times larger than that in SRS without weight, in order to produce estimates as accurate as those in SRS. The required sample sizes in SICSUP were similar to those in SC.

Different margin of errors required different sample sizes. In the studied population, in order to achieve the margin of error of .1, SICSUP as well as SC needed large sample sizes, close to or larger than the population size of 2,000. Therefore, it seemed impractical or impossible to achieve the margin of error of .1 in this population. The best choice of margin of error in this population was the margin of error of .2, and hence, in SICSUP, the sample sizes of about 760 and 620 seemed the best choices with and without sampling weight, respectively. However, one should pay attention to type of initial proportions used and the correlation between school size and the variable of interest because they may influence sample-size determination either of positively or negatively.

Third, the study in Research Question 3 examined the performance of SICSUP for multiple populations (e.g., statewide or international surveys) as compared that of SC with respect to rankings. For each country, the population mean (or country mean) was estimated and confidence interval coverage probability at a 95% confidence level was investigated. In general, SICSUP worked slightly better than SC across the five countries in terms of confidence interval coverage probability of the population mean. Especially, for the country with the highest proportion of schools with no novice teacher, or the rarest population, among the five countries, SICSUP worked fairly better than SC.

In terms of providing country rankings that are identical with those based on the population means, SICSUP worked as well as or, depending on the condition, slightly better than SC and the combination of SICSUP and SC. In Research Question 3, some interesting results were found. For example, the sample means in SICSUP were not very accurate, but it was able to

produce the county rankings that were identical with those based on the population means. This also occurred when SC or the combination of SICSUP and SC was used. These results implied that rankings should be interpreted with caution although they were frequently reported as results of national or international surveys and assessments.

Last but not least, Research Question 4 evaluated the economic aspect of SICSUP in terms of number of contacted schools in order to achieve the predetermined sample size as compared to those in SC.

Based on the dataset in Research Question 1, SICSUP contacted fewer schools than SC did. The numbers of schools in the final set of samples were similar between the two sample designs. Thus, the ratio of the number of schools in the final set of samples to the number of contacted schools during the sampling procedure, ($\frac{n_{Schools\ in\ the\ final\ sample\ )}}{n_{Contacted\ schools}}$), showed that SICSUP was more economical than SC. However, the different ratios by strata suggested that SICSUP might not be advantages for populations with large clusters.

Based on the datasets in Research Question 3, SICSUP required smaller number of contacted schools during the sampling procedure than SC did. However, the updating process of SICSUP seemed not very beneficial while the sequential selection of SICSUP was fairly advantageous. The sequential selection reduced the number of contacted schools by 10 to 20% depending on the country examined. For some countries, the updating process of SICSUP was not helpful to reduce the number of contacted schools (e.g., country with small number of strata); for the other countries, the updating process worked differently for each stratum, and the effects were canceled out when it came to the whole sample.

In this study, the probability of using substitute schools in SC was also investigated. As expected, in general, more than 80% of sets of samples used substitute schools in order to reach

the predetermined sample size. Although these results did not directly provide evidence that SICSUP was economically advantageous over SC, these suggested applications of alternative sample design instead of SC in rare populations, such as SICSUP.

The findings of the entire study reported that SICSUP worked at least as well as SC in terms of statistic aspect and was more economic than SC. SICSUP was sensitive to sample size and type of initial proportions of elements when it comes to parameter estimation. In terms of economic aspect, it was sensitive to number of strata and average cluster size in the population. The use of small number of strata or populations with large clusters could make SICSUP less economic.

5.2 Implications

As societies become more complex and heterogeneous, the field of education also becomes broader and more diverse. This leads growing interest in groups of individuals who have not attracted enough educational practitioners and researchers' attention, such as students who share a distinctive culture or religion. Therefore, a substantial amount of studies have been conducted with these groups of individuals. As such studies increase, new challenges arise. Researchers who attempt to survey these groups of people often experience difficulty in locating them. These groups of individuals, such as those who are in a distinctive culture or religion group (e.g., migrant students), those who experienced a rare event (e.g., students who experienced cyber harassment), and those who share a special characteristic (e.g., students with special educational needs), are usually rare in the general populations and hence, hard to sample. The common characteristic of the groups mentioned above is that they are students in schools. This is the same for teachers in rare populations: they are found in schools. Researchers know where to find these individuals in general, but they cannot exactly locate them. SICSUP could be

advantageous especially to such situations. The simulation studies in this dissertation suggest that SICSUP could provide results as precise as conventional SC would with contacting fewer clusters, mostly schools.

Another advantage of SICSUP is the similarity in procedure to conventional SC. Both of the designs use stratification and clusters. If researchers are familiar with SC, the procedure of SICSUP would be easy to understand, and they may be less hesitant to give it a try as compared to unfamiliar sample designs. The results of this dissertation indicate that SICSUP works as well as SC. Existing educational surveys that have used SC can change their sample design to SICSUP without facing many challenges. Existing statewide or international surveys can employ SICSUP for a part of participating states or countries that have experienced difficulties due to rarity of elements in their populations.

As its name indicates, SICSUP has a close relationship with adaptive sampling, especially with inverse sampling. Adaptive sampling has a solid foundation within sampling theory (Seber & Salehi, 2012; Thompson, 2002). There are well-established theories and sample designs that are related to adaptive sampling, and inverse sampling is one of them. These well-found bases would support SICSUP theoretically and may facilitate the understanding of concepts of the updating process and sequential selection in SICSUP. At the same time, the evaluation of SICSUP in this dissertation would contribute to the literature on adaptive sampling. Despite of a good foundation and the popularity of adaptive sampling in sampling theory, it has been hardly used in the field of educational research. SICSUP may be able to make a connection between the two areas and encourage educational researchers to employ adaptive sampling including SICSUP in their studies.

5.3 Limitation and Future Research

This section briefly discusses limitations of the study and proposes some of the directions for future research. First, a major limitation of this study is that the performance of SICSUP was evaluated only based on the results from simulations with generated datasets. Although I tried to generate datasets as realistic as possible using the TALIS2018 datasets, the results still lack in realism. Future research needs to examine the performance of SICSUP with empirical datasets. Since the development of SICSUP, it was used only once in practice for the field trial of the FIRSTMATH (First Five Years of Mathematics) Study (Tatto et al., 2020). In addition to simulation studies with real datasets, empirical evidence is required in order to evaluate the performance of SICSUP.

Second, the simulation conditions that were examined in this dissertation were (1) sample size, (2) type of initial proportions of elements over strata, (3) level of correlation between cluster size and the variable of interest, and (4) number of strata. In surveys, response rate is one of the important considerations. Response rates for surveys seem to decrease each year in general (Tourangeau et al., 2014), and increasing non-responses have caused difficulties in operation of survey, determination of sample size, and parameter estimation. With respect to rare populations, response rates of some groups of individuals tend to be low (e.g., parents with very high or low income). Evaluating the performance of SICSUP based on different levels of response rate would be suggested for future research.

Third, the number of variable of interest used in this study was one for each population, and mainly the population mean and standard deviation were estimated using SICSUP. In practice, questionnaires, tests, and interview questions include many items, so the number of variable of interest in surveys and assessments is more than one. Statistical factors such as

estimation precision and required sample size tend to be different by variables of interest within a survey (OECD, 2017, 2019). Future studies could evaluate SICSUP with multiple variable of interest. Type of standard error estimator is also an important statistical consideration when evaluating SICSUP. Along with the mean and standard deviation, different standard error estimators such as ratio, regression coefficient, and plausible value could be examined in order to evaluate the performance of SICSUP in terms of estimation precision.

Finally, future studies may explore how the point at which the updating process takes place affects the performance of SICSUP. The updating process relies on current samples collected to the updating point. If the size of the current samples is too small, the updating process may not work well. Based on the dataset generated for the first research question, with $n = 50$, some updating process occurred with few samples (less than 5 for a stratum), and such cases usually failed to produce accurate proportions. Another factor that affects the updating process is which stratum first reaches the initial sample size. With small sample size (e.g., $n = 50$), when the smallest stratum reached the initial sample size first, the updating process tended to produce less accurate proportions than when the largest stratum reached first. These may be topics for future research and provide directions to improve current SICSUP.

APPENDIX

Standard Error of the Sample Mean Using Samples without Weight

The estimated bias of a standard error estimator is the difference between the average of standard error estimates from the 10 sets of samples and the empirical standard error. A positive value indicates that the standard error estimator tends to overestimate the empirical standard error and a negative value indicates that the standard error estimator tends to underestimate the empirical standard error. The relative bias is the estimated bias divided by the empirical standard error. Because an estimated bias can be a negative or positive value, the relative bias can also be a negative or positive value.

Table A.1 Estimated Bias for the Standard Error Estimators with Original Strata and without Weight

n	ρ	σ_{UJ}	σ_{IJ}	σ_{SJ}	σ_{UB}	σ_{IB}	σ_{SB}
Initial Proportions Based on Data							
50	0.0	-0.03	0.02	0.01	-0.05	0.01	0.01
50	0.4	0.02	0.08	-0.03	0.02	0.07	-0.04
50	0.7	0.02	-0.02	0.00	0.01	-0.03	-0.01
100	0.0	-0.02	0.00	-0.02	-0.02	0.00	-0.02
100	0.4	0.01	0.01	0.00	0.00	0.00	-0.01
100	0.7	0.02	0.00	0.02	0.01	0.00	0.01
500	0.0	0.02	0.01	0.02	0.02	0.01	0.02
500	0.4	0.02	0.02	0.02	0.02	0.02	0.02
500	0.7	0.03	0.03	0.03	0.03	0.03	0.03
1000	0.0	0.03	0.03	0.03	0.03	0.03	0.03
1000	0.4	0.03	0.03	0.04	0.03	0.03	0.04
1000	0.7	0.04	0.04	0.04	0.04	0.04	0.04
Informal Estimate of Based on School Proportions							
50	0.0	-0.02	-0.01	-0.03	-0.03	-0.01	-0.03
50	0.4	0.03	-0.02	0.00	0.03	-0.02	-0.01
50	0.7	0.01	0.02	0.05	0.00	0.02	0.05
100	0.0	0.01	0.00	-0.02	0.01	0.00	-0.02
100	0.4	-0.01	0.00	-0.01	-0.02	0.00	-0.01
100	0.7	-0.01	0.01	0.01	-0.01	0.00	0.00
500	0.0	0.02	0.02	0.02	0.02	0.02	0.02
500	0.4	0.02	0.03	0.02	0.02	0.03	0.02
500	0.7	0.03	0.02	0.02	0.03	0.02	0.02
1000	0.0	0.03	0.03	0.03	0.03	0.03	0.03
1000	0.4	0.03	0.04	0.04	0.03	0.04	0.04
1000	0.7	0.04	0.03	0.03	0.03	0.03	0.03
Informal Estimate Based on Equal Proportions							
50	0.0	0.04	-0.02	-0.02	0.03	-0.02	-0.03
50	0.4	-0.02	0.04	0.02	-0.03	0.04	0.01
50	0.7	-0.03	0.01	-0.01	-0.04	0.01	-0.01
100	0.0	-0.01	-0.01	0.00	0.00	0.01	0.01
100	0.4	-0.01	-0.01	-0.01	-0.02	0.01	0.01
100	0.7	-0.01	-0.01	-0.05	-0.06	-0.01	-0.01
500	0.0	0.02	0.02	0.02	0.02	0.03	0.03
500	0.4	0.02	0.02	0.03	0.03	0.03	0.03
500	0.7	0.02	0.02	0.02	0.02	0.02	0.02
1000	0.0	0.03	0.03	0.03	0.03	0.04	0.04
1000	0.4	0.03	0.04	0.02	0.02	0.03	0.03
1000	0.7	0.03	0.03	0.03	0.03	0.04	0.04

Table A.2 Relative Bias of the Standard Error Estimators with Original Strata and without Weight

n	ρ	σ_{UJ}	σ_{IJ}	σ_{SJ}	σ_{UB}	σ_{IB}	σ_{SB}
Initial Proportions Based on Data							
50	0.0	-0.07	0.05	0.03	-0.11	0.02	0.01
50	0.4	0.05	0.17	-0.07	0.04	0.14	-0.08
50	0.7	0.03	-0.05	0.00	0.02	-0.05	-0.02
100	0.0	-0.06	0.02	-0.05	-0.07	0.01	-0.06
100	0.4	0.02	0.02	0.00	0.00	0.00	-0.02
100	0.7	0.05	0.00	0.04	0.02	0.00	0.04
500	0.0	0.16	0.11	0.17	0.17	0.10	0.16
500	0.4	0.17	0.13	0.15	0.17	0.13	0.16
500	0.7	0.18	0.21	0.19	0.19	0.21	0.18
1000	0.0	0.44	0.41	0.47	0.44	0.41	0.47
1000	0.4	0.40	0.38	0.48	0.38	0.37	0.49
1000	0.7	0.45	0.46	0.49	0.45	0.46	0.52
Informal Estimate Based on School Proportions							
50	0.0	-0.04	-0.02	-0.06	-0.06	-0.02	-0.08
50	0.4	0.06	-0.04	0.01	0.06	-0.05	-0.02
50	0.7	0.01	0.05	0.11	-0.01	0.04	0.10
100	0.0	0.04	0.01	-0.06	0.03	0.00	-0.07
100	0.4	-0.04	0.00	-0.04	-0.05	-0.01	-0.04
100	0.7	-0.01	0.02	0.03	-0.01	0.00	0.01
500	0.0	0.19	0.13	0.13	0.18	0.14	0.13
500	0.4	0.15	0.21	0.16	0.13	0.21	0.15
500	0.7	0.20	0.15	0.17	0.23	0.15	0.18
1000	0.0	0.39	0.43	0.40	0.40	0.44	0.39
1000	0.4	0.43	0.47	0.52	0.42	0.47	0.51
1000	0.7	0.41	0.37	0.41	0.41	0.37	0.40
Informal Estimate Based on Equal Proportions							
50	0.0	0.09	-0.05	-0.05	0.07	-0.06	-0.06
50	0.4	-0.05	0.09	0.05	-0.06	0.09	0.01
50	0.7	-0.06	0.02	-0.01	-0.08	0.02	-0.03
100	0.0	-0.02	-0.03	0.01	-0.01	0.04	0.04
100	0.4	-0.03	-0.04	-0.03	-0.04	0.02	0.02
100	0.7	-0.02	-0.03	-0.13	-0.14	-0.03	-0.03
500	0.0	0.17	0.18	0.14	0.16	0.22	0.22
500	0.4	0.16	0.15	0.17	0.18	0.25	0.23
500	0.7	0.17	0.18	0.11	0.10	0.18	0.17
1000	0.0	0.49	0.47	0.32	0.36	0.56	0.55
1000	0.4	0.48	0.50	0.26	0.25	0.50	0.48
1000	0.7	0.45	0.44	0.31	0.29	0.52	0.51

Table A.3 Relative MSE for the Standard Error Estimators with Original Strata and Weight

n	ρ	σ_{UJ}	σ_{IJ}	σ_{SJ}	σ_{UB}	σ_{IB}	σ_{SB}
Initial Proportions Based on Data							
50	0.0	0.07	0.04	0.04	0.08	0.05	0.03
50	0.4	0.04	0.06	0.03	0.04	0.05	0.03
50	0.7	0.01	0.02	0.02	0.01	0.02	0.03
100	0.0	0.02	0.01	0.02	0.01	0.02	0.02
100	0.4	0.03	0.02	0.01	0.03	0.03	0.01
100	0.7	0.02	0.01	0.04	0.02	0.01	0.04
500	0.0	0.03	0.02	0.03	0.03	0.02	0.03
500	0.4	0.03	0.02	0.03	0.04	0.02	0.03
500	0.7	0.04	0.05	0.04	0.04	0.05	0.04
1000	0.0	0.20	0.17	0.22	0.19	0.17	0.23
1000	0.4	0.16	0.15	0.23	0.14	0.14	0.24
1000	0.7	0.20	0.21	0.24	0.21	0.21	0.28
Informal Estimate Based on School Proportions							
50	0.0	0.08	0.03	0.04	0.08	0.02	0.04
50	0.4	0.09	0.07	0.04	0.10	0.06	0.04
50	0.7	0.02	0.03	0.05	0.02	0.03	0.05
100	0.0	0.01	0.03	0.02	0.01	0.03	0.03
100	0.4	0.01	0.03	0.01	0.01	0.03	0.01
100	0.7	0.02	0.02	0.03	0.03	0.01	0.03
500	0.0	0.04	0.02	0.02	0.04	0.02	0.02
500	0.4	0.02	0.05	0.03	0.02	0.05	0.03
500	0.7	0.05	0.03	0.03	0.06	0.03	0.04
1000	0.0	0.16	0.18	0.17	0.16	0.19	0.16
1000	0.4	0.18	0.22	0.27	0.18	0.23	0.27
1000	0.7	0.17	0.14	0.17	0.17	0.15	0.16
Informal Estimate Based on Equal Proportions							
50	0.0	0.05	0.03	0.04	0.04	0.02	0.04
50	0.4	0.05	0.07	0.04	0.05	0.05	0.03
50	0.7	0.05	0.07	0.03	0.04	0.06	0.02
100	0.0	0.02	0.02	0.03	0.03	0.02	0.02
100	0.4	0.02	0.02	0.04	0.03	0.01	0.01
100	0.7	0.02	0.02	0.05	0.05	0.01	0.01
500	0.0	0.04	0.04	0.04	0.05	0.06	0.06
500	0.4	0.03	0.03	0.05	0.05	0.07	0.06
500	0.7	0.04	0.04	0.02	0.02	0.04	0.04
1000	0.0	0.24	0.22	0.11	0.14	0.32	0.30
1000	0.4	0.23	0.26	0.07	0.07	0.25	0.24
1000	0.7	0.20	0.20	0.11	0.10	0.28	0.27

Table A.4 Confidence Interval Coverage Probability of the Standard Error Estimators with Original Strata and without Weight

n	ρ	σ_{UJ}	σ_{IJ}	σ_{SJ}	σ_{UB}	σ_{IB}	σ_{SB}
Initial Proportions Based on Data							
50	0.0	0.90	1.00	0.90	0.90	1.00	0.90
50	0.4	0.90	0.70	0.90	0.90	0.70	0.90
50	0.7	0.80	0.90	0.90	0.80	0.90	0.90
100	0.0	0.90	1.00	1.00	0.90	1.00	1.00
100	0.4	0.90	0.90	0.90	0.90	0.90	0.90
100	0.7	1.00	0.90	1.00	1.00	0.90	1.00
500	0.0	0.80	1.00	1.00	0.80	1.00	1.00
500	0.4	1.00	0.90	1.00	0.90	0.90	1.00
500	0.7	0.90	1.00	1.00	0.90	1.00	1.00
1000	0.0	1.00	1.00	0.90	1.00	1.00	0.90
1000	0.4	0.90	1.00	1.00	0.90	1.00	1.00
1000	0.7	1.00	0.90	1.00	1.00	0.90	1.00
Informal Estimate Based on School Proportions							
50	0.0	1.00	0.90	0.80	1.00	0.90	0.80
50	0.4	0.90	1.00	0.80	0.90	0.90	0.80
50	0.7	1.00	0.90	1.00	1.00	0.90	1.00
100	0.0	0.80	1.00	1.00	0.80	1.00	0.90
100	0.4	0.80	0.90	0.90	0.80	0.90	0.90
100	0.7	0.90	1.00	0.80	0.90	1.00	0.80
500	0.0	1.00	0.90	1.00	1.00	0.90	1.00
500	0.4	0.90	1.00	1.00	0.90	1.00	0.90
500	0.7	1.00	0.90	0.90	1.00	0.90	1.00
1000	0.0	1.00	1.00	1.00	1.00	1.00	1.00
1000	0.4	1.00	1.00	1.00	1.00	1.00	1.00
1000	0.7	0.90	1.00	0.90	0.90	1.00	0.90
Informal Estimate Based on Equal Proportions							
50	0.0	1.00	0.90	1.00	1.00	0.90	1.00
50	0.4	0.90	0.90	0.90	0.90	0.90	0.90
50	0.7	0.80	0.90	0.90	0.80	0.90	0.90
100	0.0	1.00	1.00	1.00	1.00	1.00	1.00
100	0.4	1.00	1.00	1.00	1.00	1.00	1.00
100	0.7	1.00	1.00	1.00	1.00	1.00	1.00
500	0.0	0.90	0.90	1.00	1.00	1.00	1.00
500	0.4	1.00	1.00	0.80	0.80	0.90	0.90
500	0.7	0.90	0.80	1.00	1.00	1.00	1.00
1000	0.0	1.00	1.00	1.00	1.00	1.00	1.00
1000	0.4	1.00	1.00	0.90	1.00	1.00	1.00
1000	0.7	0.90	0.90	1.00	1.00	1.00	1.00

Table A.5 Estimated Bias of the Standard Error Estimators with Pseudo-Strata and without Weight

n	ρ	σ_{UJ}	σ_{IJ}	σ_{SJ}	σ_{UB}	σ_{IB}	σ_{SB}	σ_{UR}	σ_{IR}	σ_{SR}	σ_{UF}	σ_{IF}	σ_{SF}
Initial Proportions Based on Data													
50	0.0	-0.13	-0.04	-0.09	-0.13	-0.04	-0.09	-0.13	-0.04	-0.08	-0.13	-0.04	-0.09
50	0.4	0.05	-0.02	-0.10	0.05	-0.02	-0.09	0.05	-0.01	-0.10	0.05	-0.01	-0.10
50	0.7	-0.12	-0.09	-0.02	-0.12	-0.09	-0.03	-0.12	-0.09	-0.02	-0.12	-0.09	-0.02
100	0.0	-0.09	-0.09	-0.06	-0.09	-0.09	-0.06	-0.09	-0.09	-0.06	-0.09	-0.09	-0.06
100	0.4	-0.05	-0.04	-0.06	-0.04	-0.04	-0.06	-0.05	-0.04	-0.06	-0.05	-0.04	-0.06
100	0.7	-0.03	-0.05	-0.07	-0.03	-0.05	-0.07	-0.03	-0.05	-0.07	-0.03	-0.05	-0.07
500	0.0	-0.03	-0.04	-0.03	-0.03	-0.04	-0.03	-0.03	-0.04	-0.03	-0.03	-0.04	-0.03
500	0.4	-0.04	-0.03	-0.03	-0.04	-0.03	-0.03	-0.04	-0.03	-0.03	-0.04	-0.03	-0.03
500	0.7	-0.05	-0.06	-0.04	-0.05	-0.06	-0.04	-0.05	-0.06	-0.04	-0.05	-0.06	-0.04
1000	0.0	-0.02	-0.03	-0.01	-0.02	-0.03	-0.01	-0.02	-0.03	-0.01	-0.02	-0.03	-0.01
1000	0.4	-0.02	-0.03	-0.02	-0.03	-0.03	-0.01	-0.02	-0.03	-0.02	-0.02	-0.03	-0.02
1000	0.7	-0.03	-0.03	-0.02	-0.03	-0.03	-0.02	-0.03	-0.03	-0.02	-0.03	-0.03	-0.02
Informal Estimate Based on School Proportions													
50	0.0	-0.13	-0.10	-0.10	-0.14	-0.10	-0.10	-0.13	-0.10	-0.09	-0.13	-0.10	-0.09
50	0.4	-0.11	-0.08	-0.06	-0.12	-0.08	-0.07	-0.11	-0.08	-0.06	-0.11	-0.08	-0.06
50	0.7	-0.10	0.00	0.05	-0.10	0.00	0.05	-0.10	0.00	0.05	-0.10	0.00	0.05
100	0.0	-0.07	-0.03	-0.06	-0.06	-0.03	-0.06	-0.07	-0.03	-0.06	-0.07	-0.03	-0.06
100	0.4	-0.08	-0.02	-0.05	-0.08	-0.02	-0.05	-0.08	-0.02	-0.05	-0.08	-0.02	-0.05
100	0.7	-0.02	-0.04	-0.07	-0.02	-0.04	-0.06	-0.02	-0.04	-0.07	-0.02	-0.04	-0.07
500	0.0	-0.04	-0.04	-0.02	-0.04	-0.04	-0.02	-0.04	-0.04	-0.02	-0.04	-0.04	-0.02
500	0.4	-0.04	-0.04	-0.04	-0.03	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04
500	0.7	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04
1000	0.0	-0.02	-0.01	-0.01	-0.02	-0.01	-0.01	-0.02	-0.01	-0.01	-0.02	-0.01	-0.01
1000	0.4	-0.02	-0.02	-0.01	-0.02	-0.02	-0.01	-0.02	-0.02	-0.01	-0.02	-0.02	-0.01
1000	0.7	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
Informal Estimate Based on Equal Proportions													
50	0.0	-0.03	-0.07	-0.07	-0.03	-0.07	-0.07	-0.03	-0.07	-0.07	-0.03	-0.07	-0.07
50	0.4	-0.08	-0.06	-0.03	-0.08	-0.06	-0.03	-0.08	-0.05	-0.03	-0.08	-0.06	-0.03
50	0.7	-0.10	-0.06	-0.03	-0.10	-0.06	-0.03	-0.10	-0.07	-0.03	-0.10	-0.07	-0.03
100	0.0	-0.07	-0.07	-0.07	-0.07	-0.04	-0.04	-0.03	-0.04	-0.02	-0.02	-0.02	-0.02
100	0.4	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.04	-0.04	-0.04	-0.04
100	0.7	-0.04	-0.04	-0.04	-0.04	-0.09	-0.09	-0.09	-0.09	-0.05	-0.05	-0.05	-0.05
500	0.0	-0.02	-0.02	-0.02	-0.02	-0.03	-0.04	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02
500	0.4	-0.04	-0.04	-0.04	-0.04	-0.03	-0.03	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02
500	0.7	-0.05	-0.05	-0.05	-0.05	-0.06	-0.06	-0.06	-0.06	-0.04	-0.04	-0.04	-0.04
1000	0.0	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00
1000	0.4	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.01	-0.01	-0.01	-0.01
1000	0.7	-0.03	-0.03	-0.03	-0.03	-0.04	-0.04	-0.04	-0.04	-0.02	-0.02	-0.02	-0.02

Table A.6 Relative Bias of the Standard Error Estimators with Pseudo-Strata and without Weight

n	ρ	σ_{UJ}	σ_{IJ}	σ_{SJ}	σ_{UB}	σ_{IB}	σ_{SB}	σ_{UR}	σ_{IR}	σ_{SR}	σ_{UF}	σ_{IF}	σ_{SF}
Initial Proportions Based on Data													
50	0.0	-0.29	-0.08	-0.21	-0.29	-0.09	-0.20	-0.28	-0.08	-0.17	-0.29	-0.09	-0.21
50	0.4	0.10	-0.04	-0.20	0.10	-0.03	-0.20	0.09	-0.01	-0.20	0.09	-0.03	-0.20
50	0.7	-0.23	-0.18	-0.05	-0.23	-0.18	-0.05	-0.23	-0.18	-0.05	-0.23	-0.18	-0.05
100	0.0	-0.28	-0.29	-0.18	-0.28	-0.29	-0.18	-0.28	-0.28	-0.18	-0.28	-0.29	-0.18
100	0.4	-0.14	-0.11	-0.19	-0.13	-0.10	-0.18	-0.13	-0.11	-0.19	-0.13	-0.11	-0.19
100	0.7	-0.07	-0.13	-0.19	-0.07	-0.13	-0.20	-0.07	-0.13	-0.19	-0.07	-0.13	-0.19
500	0.0	-0.26	-0.32	-0.25	-0.26	-0.32	-0.24	-0.26	-0.32	-0.25	-0.26	-0.32	-0.25
500	0.4	-0.27	-0.25	-0.23	-0.27	-0.25	-0.25	-0.27	-0.25	-0.23	-0.28	-0.25	-0.23
500	0.7	-0.31	-0.43	-0.28	-0.31	-0.43	-0.28	-0.31	-0.43	-0.27	-0.31	-0.42	-0.27
1000	0.0	-0.29	-0.33	-0.19	-0.29	-0.33	-0.20	-0.29	-0.33	-0.20	-0.30	-0.33	-0.20
1000	0.4	-0.30	-0.32	-0.21	-0.31	-0.32	-0.20	-0.30	-0.32	-0.21	-0.31	-0.32	-0.21
1000	0.7	-0.35	-0.38	-0.29	-0.34	-0.38	-0.30	-0.34	-0.38	-0.29	-0.34	-0.38	-0.29
Informal Estimate Based on School Proportions													
50	0.0	-0.29	-0.22	-0.22	-0.30	-0.21	-0.22	-0.28	-0.22	-0.21	-0.28	-0.22	-0.22
50	0.4	-0.23	-0.17	-0.14	-0.24	-0.17	-0.15	-0.23	-0.16	-0.13	-0.23	-0.17	-0.14
50	0.7	-0.20	0.01	0.10	-0.20	0.01	0.10	-0.19	0.00	0.10	-0.19	0.01	0.10
100	0.0	-0.20	-0.10	-0.20	-0.19	-0.10	-0.19	-0.20	-0.09	-0.20	-0.20	-0.09	-0.20
100	0.4	-0.22	-0.06	-0.16	-0.22	-0.06	-0.16	-0.22	-0.06	-0.16	-0.22	-0.06	-0.16
100	0.7	-0.06	-0.12	-0.20	-0.06	-0.11	-0.18	-0.06	-0.11	-0.20	-0.06	-0.12	-0.20
500	0.0	-0.28	-0.31	-0.19	-0.29	-0.32	-0.19	-0.29	-0.31	-0.20	-0.29	-0.31	-0.20
500	0.4	-0.25	-0.27	-0.28	-0.24	-0.28	-0.28	-0.25	-0.27	-0.28	-0.26	-0.27	-0.28
500	0.7	-0.30	-0.31	-0.31	-0.30	-0.30	-0.33	-0.30	-0.31	-0.31	-0.30	-0.31	-0.32
1000	0.0	-0.20	-0.19	-0.22	-0.21	-0.19	-0.21	-0.20	-0.19	-0.21	-0.20	-0.18	-0.21
1000	0.4	-0.27	-0.27	-0.13	-0.25	-0.27	-0.11	-0.26	-0.27	-0.13	-0.27	-0.27	-0.13
1000	0.7	-0.36	-0.31	-0.34	-0.36	-0.31	-0.36	-0.36	-0.31	-0.34	-0.36	-0.31	-0.34
Informal Estimate Based on Equal Proportions													
50	0.0	-0.07	-0.17	-0.17	-0.07	-0.15	-0.16	-0.08	-0.16	-0.17	-0.07	-0.17	-0.17
50	0.4	-0.17	-0.12	-0.07	-0.17	-0.12	-0.06	-0.17	-0.12	-0.07	-0.17	-0.12	-0.07
50	0.7	-0.18	-0.13	-0.07	-0.18	-0.13	-0.07	-0.19	-0.14	-0.07	-0.19	-0.14	-0.07
100	0.0	-0.23	-0.21	-0.23	-0.23	-0.10	-0.10	-0.10	-0.10	-0.08	-0.08	-0.08	-0.08
100	0.4	-0.15	-0.14	-0.14	-0.15	-0.14	-0.14	-0.14	-0.14	-0.11	-0.11	-0.11	-0.11
100	0.7	-0.13	-0.12	-0.13	-0.13	-0.23	-0.23	-0.22	-0.23	-0.16	-0.16	-0.16	-0.16
500	0.0	-0.20	-0.20	-0.20	-0.20	-0.23	-0.25	-0.23	-0.23	-0.17	-0.16	-0.17	-0.17
500	0.4	-0.32	-0.32	-0.32	-0.32	-0.19	-0.20	-0.19	-0.19	-0.14	-0.12	-0.14	-0.14
500	0.7	-0.40	-0.40	-0.40	-0.40	-0.34	-0.35	-0.34	-0.34	-0.32	-0.32	-0.32	-0.32
1000	0.0	-0.17	-0.18	-0.17	-0.17	-0.17	-0.16	-0.17	-0.17	-0.06	-0.07	-0.06	-0.06
1000	0.4	-0.23	-0.23	-0.23	-0.23	-0.26	-0.26	-0.26	-0.26	-0.11	-0.10	-0.11	-0.11
1000	0.7	-0.35	-0.36	-0.35	-0.35	-0.37	-0.37	-0.37	-0.37	-0.28	-0.28	-0.28	-0.28

Table A.7 Relative MSE of the Standard Error Estimators with Pseudo-Strata and without Weight

n	ρ	σ_{UJ}	σ_{IJ}	σ_{SJ}	σ_{UB}	σ_{IB}	σ_{SB}	σ_{UR}	σ_{IR}	σ_{SR}	σ_{UF}	σ_{IF}	σ_{SF}
Initial Proportions Based on Data													
50	0.0	0.15	0.05	0.08	0.15	0.05	0.08	0.15	0.04	0.10	0.15	0.04	0.08
50	0.4	0.10	0.08	0.09	0.11	0.09	0.09	0.10	0.09	0.09	0.10	0.08	0.09
50	0.7	0.09	0.10	0.07	0.09	0.10	0.08	0.09	0.10	0.07	0.09	0.10	0.07
100	0.0	0.10	0.10	0.06	0.09	0.10	0.05	0.09	0.10	0.06	0.09	0.10	0.06
100	0.4	0.06	0.06	0.04	0.06	0.05	0.04	0.06	0.05	0.04	0.06	0.05	0.04
100	0.7	0.06	0.07	0.08	0.07	0.06	0.08	0.06	0.07	0.08	0.06	0.07	0.08
500	0.0	0.08	0.11	0.07	0.08	0.11	0.07	0.08	0.11	0.07	0.08	0.11	0.07
500	0.4	0.09	0.07	0.06	0.09	0.07	0.07	0.09	0.07	0.06	0.10	0.07	0.06
500	0.7	0.11	0.20	0.09	0.11	0.20	0.09	0.11	0.20	0.09	0.11	0.20	0.09
1000	0.0	0.10	0.12	0.05	0.10	0.12	0.06	0.10	0.12	0.05	0.11	0.12	0.05
1000	0.4	0.10	0.11	0.05	0.11	0.11	0.04	0.10	0.11	0.05	0.11	0.11	0.05
1000	0.7	0.13	0.18	0.10	0.13	0.18	0.10	0.13	0.18	0.10	0.13	0.18	0.10
Informal Estimate Based on School Proportions													
50	0.0	0.16	0.07	0.09	0.16	0.06	0.09	0.15	0.07	0.09	0.15	0.07	0.09
50	0.4	0.15	0.15	0.09	0.15	0.16	0.10	0.15	0.15	0.09	0.15	0.15	0.09
50	0.7	0.09	0.09	0.11	0.09	0.08	0.11	0.09	0.09	0.11	0.09	0.09	0.11
100	0.0	0.07	0.06	0.07	0.07	0.06	0.06	0.07	0.07	0.07	0.07	0.06	0.07
100	0.4	0.07	0.07	0.06	0.07	0.08	0.06	0.07	0.07	0.06	0.07	0.07	0.06
100	0.7	0.04	0.07	0.05	0.04	0.08	0.05	0.04	0.07	0.05	0.04	0.07	0.05
500	0.0	0.10	0.11	0.05	0.10	0.11	0.05	0.10	0.11	0.05	0.11	0.11	0.05
500	0.4	0.08	0.09	0.09	0.08	0.09	0.09	0.08	0.09	0.09	0.09	0.09	0.09
500	0.7	0.10	0.12	0.11	0.10	0.12	0.13	0.10	0.12	0.12	0.10	0.12	0.12
1000	0.0	0.05	0.05	0.06	0.05	0.05	0.06	0.05	0.05	0.06	0.05	0.05	0.06
1000	0.4	0.09	0.09	0.02	0.08	0.08	0.02	0.09	0.09	0.02	0.09	0.09	0.02
1000	0.7	0.15	0.12	0.13	0.15	0.11	0.14	0.15	0.12	0.13	0.16	0.12	0.13
Informal Estimate Based on Equal Proportions													
50	0.0	0.06	0.07	0.09	0.06	0.07	0.09	0.06	0.07	0.09	0.06	0.07	0.09
50	0.4	0.05	0.09	0.09	0.05	0.11	0.10	0.05	0.10	0.09	0.05	0.10	0.09
50	0.7	0.10	0.08	0.04	0.09	0.08	0.04	0.10	0.08	0.04	0.10	0.08	0.04
100	0.0	0.09	0.08	0.09	0.09	0.06	0.07	0.07	0.07	0.04	0.05	0.05	0.05
100	0.4	0.07	0.07	0.07	0.07	0.09	0.08	0.09	0.09	0.04	0.04	0.04	0.04
100	0.7	0.07	0.07	0.07	0.07	0.11	0.10	0.11	0.11	0.05	0.05	0.05	0.05
500	0.0	0.07	0.07	0.07	0.07	0.08	0.09	0.08	0.08	0.05	0.05	0.05	0.05
500	0.4	0.12	0.12	0.12	0.12	0.04	0.04	0.04	0.04	0.02	0.02	0.02	0.02
500	0.7	0.18	0.18	0.18	0.18	0.13	0.14	0.13	0.13	0.11	0.11	0.11	0.11
1000	0.0	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.02	0.02	0.02	0.02
1000	0.4	0.06	0.07	0.06	0.06	0.09	0.09	0.09	0.09	0.02	0.02	0.02	0.02
1000	0.7	0.14	0.15	0.14	0.14	0.16	0.16	0.16	0.16	0.09	0.09	0.09	0.09

Table A.8 Confidence Interval Coverage Probability of the Standard Error Estimators with Pseudo-Strata and without Weight

n	ρ	σ_{UJ}	σ_{IJ}	σ_{SJ}	σ_{UB}	σ_{IB}	σ_{SB}	σ_{UR}	σ_{IR}	σ_{SR}	σ_{UF}	σ_{IF}	σ_{SF}
Initial Proportions Based on Data													
50	0.0	0.80	0.90	0.80	0.80	0.90	0.80	0.80	0.90	0.90	0.80	0.90	0.80
50	0.4	0.90	0.70	0.90	0.90	0.70	0.80	0.90	0.70	0.90	0.90	0.70	0.90
50	0.7	0.80	0.90	0.60	0.80	0.90	0.60	0.80	0.90	0.60	0.80	0.90	0.60
100	0.0	0.80	0.70	0.90	0.80	0.70	0.90	0.80	0.70	0.90	0.80	0.70	0.90
100	0.4	0.80	0.70	0.70	0.80	0.70	0.80	0.80	0.70	0.70	0.80	0.70	0.70
100	0.7	1.00	0.90	0.90	1.00	0.90	0.90	1.00	0.90	0.90	1.00	0.90	0.90
500	0.0	0.60	0.80	0.90	0.60	0.80	0.90	0.60	0.80	0.90	0.60	0.80	0.90
500	0.4	0.90	0.60	0.80	0.90	0.60	0.80	0.90	0.60	0.80	0.90	0.60	0.80
500	0.7	0.50	0.80	0.80	0.50	0.80	0.80	0.50	0.80	0.80	0.50	0.80	0.80
1000	0.0	0.80	0.90	0.70	0.80	0.90	0.70	0.80	0.90	0.70	0.80	0.90	0.70
1000	0.4	0.70	0.70	0.40	0.70	0.70	0.40	0.70	0.70	0.40	0.70	0.70	0.40
1000	0.7	0.50	0.50	0.60	0.50	0.50	0.60	0.50	0.50	0.60	0.50	0.50	0.60
Informal Estimate Based on School Proportions													
50	0.0	1.00	0.90	0.60	1.00	0.90	0.60	1.00	0.90	0.60	1.00	0.90	0.60
50	0.4	0.60	0.90	0.80	0.60	0.90	0.80	0.60	0.90	0.80	0.60	0.90	0.80
50	0.7	0.70	0.80	1.00	0.70	0.80	1.00	0.70	0.80	1.00	0.70	0.80	1.00
100	0.0	0.80	0.80	0.60	0.80	0.80	0.60	0.80	0.80	0.60	0.80	0.80	0.60
100	0.4	0.70	0.90	0.90	0.80	0.90	0.90	0.70	0.90	0.90	0.70	0.90	0.90
100	0.7	0.90	1.00	0.80	0.90	1.00	0.80	0.90	1.00	0.80	0.90	1.00	0.80
500	0.0	0.80	0.90	0.90	0.90	0.90	0.90	0.80	0.90	0.90	0.80	0.90	0.90
500	0.4	0.90	1.00	0.60	0.90	1.00	0.60	0.90	1.00	0.60	0.90	1.00	0.60
500	0.7	0.80	0.80	0.50	0.80	0.80	0.50	0.80	0.80	0.50	0.80	0.80	0.50
1000	0.0	0.80	0.70	1.00	0.80	0.70	0.90	0.80	0.70	1.00	0.80	0.70	1.00
1000	0.4	1.00	0.80	0.90	1.00	0.80	0.90	1.00	0.80	0.90	1.00	0.80	0.90
1000	0.7	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60
Informal Estimate Based on Equal Proportions													
50	0.0	1.00	0.90	1.00	1.00	0.90	1.00	1.00	0.90	1.00	1.00	0.90	1.00
50	0.4	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
50	0.7	0.70	0.80	0.90	0.70	0.80	0.90	0.70	0.80	0.90	0.70	0.80	0.90
100	0.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
100	0.4	0.90	1.00	0.90	0.90	0.90	0.90	0.90	0.90	1.00	1.00	1.00	1.00
100	0.7	0.90	0.90	0.90	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
500	0.0	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
500	0.4	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
500	0.7	0.70	0.70	0.70	0.70	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
1000	0.0	0.90	0.90	0.90	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1000	0.4	1.00	1.00	1.00	1.00	0.80	0.90	0.80	0.80	0.90	0.90	0.90	0.90
1000	0.7	0.80	0.80	0.80	0.80	0.70	0.70	0.70	0.70	0.50	0.50	0.50	0.50

REFERENCES

REFERENCES

Agresti, A. & Finlay, B., (2009). *Statistical methods for the social sciences*. Upper Saddle River, NJ: Prentice Hall.

Allensworth, E., Ponisciak, S., & Mazzeo, C. (2009). *The schools teachers leave: Teacher mobility in Chicago public schools*. Chicago, IL: Consortium on Chicago School Research. Retrieved from https://consortium.uchicago.edu/sites/default/files/2018-10/CCSR_Teacher_Mobility.pdf

Anscombe, F. J. (1953). Sequential estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, *15*(1), 1-21.

Bailey, T., & Weininger, E. B. (2002). Performance, graduation, and transfer of immigrants and natives in City University of New York community colleges. *Educational Evaluation and Policy Analysis*, *24*(4), 359-377.

Betti, G., Gagliardi, F., & Verma, V. (2018). Simplified jackknife variance estimates for fuzzy measures of multidimensional poverty. *International Statistical Review*, *86*(1), 68-86.

Burke, A. M., Morita-Mullaney, T., & Singh, M. (2016). Indiana emergent bilingual student time to reclassification: A survival analysis. *American Educational Research Journal*, *53*(5), 1310-1342.

Canty, A. J., & Davison, A. C. (1999). Resampling-based variance estimation for labour force surveys. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *48*(3), 379-391.

Chen, T. C., Bobbitt, P. A., Himelein, J. A., Paben, S. P., Cho, M. J., & Ernst, L. R. (2007). Variance estimation for international price program indexes. *2007 Proceedings of the American Statistical Association*, 1427-1434.

Chen, H., & Shen, Q. R. (2019). Variance estimation for survey-weighted data using bootstrap resampling methods: 2013 methods-of-payment survey questionnaire. In P. H. Kim, D. T. Jacho-Chavez, & G. Tripathi (Eds.), *The Econometrics of Complex Survey Data: Theory and Applications* (pp. 144-163). Bingley, England: Emerald Publishing Limited.

Christman, M. C. (2004). Sequential sampling for rare and geographically clustered populations. In W. Thompson (Ed.), *Sampling rare or elusive species: Concepts, designs, and techniques for estimating population parameters* (pp.134-145). Washington, DC: Island Press.

Chubbuck, S. M., Clift, R. T., Allard, J., & Quinlan, J. (2001). Playing it safe as a novice teacher: Implications for programs for new teachers. *Journal of Teacher Education, 52*(5), 365-376.

Clark, R. G., & Steel, D. G. (2000). Optimum allocation of sample to strata and stages with simple additional constraints. *Journal of the Royal Statistical Society: Series D (The Statistician), 49*(2), 197-207.

Daly, F. & Gilligan, R. (2005). *Lives in foster care*: *The educational and social support experiences of young people aged 13 to 14 years in long-term foster case*. Dublin, Ireland: Children's Research Centre, Trinity College, Dublin.

De Róiste, A., & Dinneen, J. (2005). *Young people's views about opportunities, barriers and supports to recreation and leisure.* Dublin, Ireland: Government Publications.

Dippo, C. S., Fay, R. E., & Morganstein, D. H. (1984). Computing variances from complex samples with replicate weights. *Proceedings of the Survey Research Methods Section*, 489-494.

Downing, K., & Ganotice Jr, F. A. (Eds.). (2016). *World university rankings and the future of higher education*. Hershey, PA: IGI Global.

Ghosh, M., Parr, W. C., Singh, K., & Babu, G. J. (1984). A note on bootstrapping the sample median. *Annals of Statistics, 12*(3), 1130-1135.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*(1), 1-26.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia, PA: SIAM.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall.

Folsom, R. (2014). *National assessment approach to sampling error estimation* (Report No. BK-0013-1412). Research Triangle Institute. Retrieved from https://www.rti.org/rti-press-publication/sampling-error-estimation/fulltext.pdf

FrankelL, M. R. (1971). *Inference from survey samples.* Ann Arbor, MI: Institute for Social Research, University of Michigan.

Greenberg Motamedi, J., Singh, M., & Thompson, K. D. (2016). *English learner student characteristics and time to reclassification: An example from Washington state* (Report No. REL 2016–128). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational

Laboratory Northwest. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/northwest/pdf/REL_2016128.pdf

Haas, E., Tran, L., Huang, M., & Yu, A. (2015). *The achievement progress of English learner students in Arizona* (Report No. REL 2015-098). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/west/pdf/REL_2015098.pdf

Haldane, J.B.S. (1945). On a method of estimating frequencies. *Biometrika, 33*(3), 222-225.

Hall, P. (1989). On efficient bootstrap simulation, *Biometrika, 76*(3), 613-617.

Judkins, D. R. (1990). Fay's method for variance estimation. *Journal of Official Statistics, 6*(3), 223-239.

Kamens, D. H., & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review, 54*(1), 5-25.

Kim, S., Ju, U., & Reckase, M. D. (2015). Obtaining a representative sample when the distribution of elements is not known. A poster was presented at the meeting of the Psychometric Society. Beijing, China.

Kinnunen, P., & Malmi, L. (2006). Why students drop out CS1 course?. *Proceedings of the Second International Workshop on Computing Education Research*, 97-108.

Kirsch, I., Lennon, M., von Davier, M., Gonzalez, E., & Yamamoto, K. (2013). On the growing importance of international large-scale assessments. In M. Von Davier, E. Gonzales, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 1-11). Dordrecht, Netherlands: Springer.

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

Kish, L. (1985). Sample surveys versus experiments, controlled observations, censuses, registers, and local studies. *Australian Journal of Statistics, 27*(2), 111-122.

Kish, L. (1991). Taxonomy of elusive populations. *Journal of Official Statistics, 7*(3), 340-347.

Kish, L., & Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society: Series B (Methodological), 36*(1), 1-22.

Kovar, J. G., Rao, J. N. K., & Wu, C. F. J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics, 16*(S1), 25-46.

Lassibille, G. & Navarro Gómez, L. (2008). Why do higher education students drop out? Evidence from Spain. *Education Economics, 16*(1), 89-105.

Lee, S. E., Lee, P. R., & Shin, K. I. (2016). A composite estimator for stratified two stage cluster sampling. *Communications for Statistical Applications and Methods*, *23*(1), 47-55.

Lee, V. E., Ready, D. D., & Johnson, D. J. (2001). The difficulty of identifying rare samples to study: The case of high schools divided into schools-within-schools. *Educational Evaluation and Policy Analysis*, *23*(4), 365-379.

Lesaux, N. K., & Kieffer, M. J. (2010). Exploring sources of reading comprehension difficulties among language minority learners and their classmates in early adolescence. *American Educational Research Journal*, *47*(3), 596-632.

Lumley, T. (2020). Survey: Analysis of complex survey samples. R package version 4.0.

Lo, N., Griffith, D. & Hunter, J. (1997). Using a restricted adaptive cluster sampling to estimate pacific hake larval abundance. *California Cooperative Oceanic Fisheries Investigations Report, 38,* 103-113.

Mach, L., Dumais, J., & Robinson, A. A. (2005). Study of the properties of a bootstrap variance estimator under sampling without replacement. A paper was presented at the Federal Committee on Statistical Methodology (FCSM) Research Conference. Arlington, VA.

Martin, M. O., Mullis, I. V., & Hooper, M. (2016). *Methods and procedures in TIMSS 2015*. Boston College, TIMSS & PIRLS International Study Center. Retrieved from http:// timssandpirls.bc.edu/publications/timss/2015-methods.html

The MathWorks, Inc. (1984-2015). MATLAB version 10.1. Natick, MA: The MathWorks Inc.

McDonald, L. L. (2004). Sampling rare populations. In W. Thompson (Ed.), *Sampling rare or elusive species: Concepts, designs, and techniques for estimating population parameters* (pp.11-42). Washington, DC: Island Press.

Mitchell, R. E. (2006). How many deaf people are there in the United States? Estimates from the Survey of Income and Program Participation. *Journal of deaf studies and deaf education*, *11*(1), 112-119.

Murthy, M. N. (1967). *Sampling theory and methods*. Calcutta, Canada: Eka Press.

National Academies of Sciences, Engineering, and Medicine (2018). *Improving health research on small populations: Proceedings of a workshop*. Washington, DC: The National Academies Press.

OECD (2016). Country note: Key findings from PISA 2015 for the United States. Retrieved from https://www.oecd.org/pisa/PISA-2015-United-States.pdf

OECD (2017). *PISA 2015 technical report*. Paris, France: PISA, OECD Publishing. Retrieved from http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

OECD (2019). *TALIS starting strong 2018 technical report*. Paris, France: TALIS, OECD Publishing. Retrieved from https://www.oecd.org/education/talis/TALIS_2018_Technical_Report.pdf

Paben, S. P. (1999). Comparison of variance estimation methods for the national compensation survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 709-795.

Pathak, P. K. (1976). Unbiased estimation in fixed cost sequential sampling schemes. *The Annals of Statistics, 4*(5), 1012–1017.

Quenouille, M. H. (1949). The joint distribution of serial correlation coefficients. *The Annals of Mathematical Statistics, 20*(4), 561–571.

Rao, J. N. K., & Wu, C. J. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association, 80*(391), 620-630.

Rao, J. N. K., Wu, C. F. J., & Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology, 18*(2), 209-217.

R Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria. URL https://www.R-project.org/.: R Foundation for Statistical Computing.

Reckase, M. D., Kim, S., & Ju, U. (2016). Sequential cluster sampling for international studies. A paper was presented at the meeting of the Psychometric Society. Asheville, NC.

Riniolo, T. C. (1999). Using a large control group for statistical comparison: Evaluation of a between-groups median test. *The Journal of Experimental Education, 68*(1), 75-88.

Robinson, A. P., & Hamann, J. D. (2008). Correcting for spatial autocorrelation in sequential sampling. *Journal of Applied Ecology, 45*(4), 1221-1227.

Rust, K. F., & Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research, 5*(3), 283-310.

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). (2013). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. New York: CRC Press.

Shalizi, C. R. (2016). *Advanced data analysis from an elementary point of view*. Cambridge, England: Cambridge University Press.

Smith, P. J., Srinath, C. K., & Battaglia, M. P. (2000). Issues relating to the use of jackknife methods in the National Immunization Survey. A paper was presented at the American Statistical Association Meetings. Indianapolis, IN.

Salehi, M. & Seber, G. A. (1997). Two-stage adaptive cluster sampling. *Biometrics*, *53*(3), 959-970.

Salehi, M. M. & Smith, D. R. (2005). Two-stage sequential sampling: A neighborhood-free adaptive sampling procedure. *Journal of Agricultural, Biological, and Environmental Statistics*, *10*(1), 84-103.

Scott, J. A. & Hoffmeister, R. J. (2016). American sign language and academic English: Factors influencing the reading of bilingual secondary school deaf and hard of hearing students. *The Journal of Deaf Studies and Deaf Education*, *22*(1), 1-13.

Seber, G. A. & Salehi, M. M. (2012). *Adaptive sample designs: Inference for sparse and clustered populations*. New York: Springer Science & Business Media.

Shin, H. S. (1995). Estimating future teacher supply: Any policy implications for educational reform?. *International Journal of Educational Reform*, *4*(4), 422-433.

Simon, N. S. & Johnson, S. M. (2015). Teacher turnover in high-poverty schools: What we know and can do. *Teachers College Record*, *117*(3), 1-36.

Skinner, C. J., Holt, D., & Smith, T. M. F. (1989). *Analysis of complex surveys*. Chichester, England: Wiley.

Smith, T. M. & Ingersoll, R. M. (2004). What are the effects of induction and mentoring on beginning teacher turnover?. *American Educational Research Journal*, *41*(3), 681-714.

Smith, W. C. (Ed.) (2016). *The global testing culture: Shaping educational policy, perceptions, and practice*, Oxford, England: Symposium Books.

Stapleton, L. M. (2008). Variance estimation using replication methods in structural equation modeling with complex sample data. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(2), 183-210.

Statistics Canada (2018). *General social survey cycle 30: Canadians at work and home*. Minister of Industry. Retrieved from http://sda.chass.utoronto.ca/sdaweb/dli2/gss/gss30/gss30/more_doc/GSSC30ENgid.pdf

Statistics Canada (2019). National Cannabis Survey, third quarter 2019, The Daily, October 30. Retrieved from https://www150.statcan.gc.ca/n1/en/daily-quotidien/191030/dq191030a-eng.pdf?st=AzIUltJh

Tatto, M. T. (2014). Teacher Education Development Study-Mathematics (TEDS-M). In S. Lerman (Ed.), *Encyclopedia of Mathematics Education* (pp. 586-592). Dordrecht, Netherlands: Springer.

Tatto, M. T., Rodriguez, M. C., Reckase, M. D., Smith, W. M., Bankov, K., & Pippin, J. (2020). *The First Five Years of Teaching Mathematics (FIRSTMATH): Concepts, methods and strategies for comparative international research*. Dordrecht, Netherlands: Springer Nature.

Tourangeau, R., Edwards, B., Johnson, T. P., Wolter, K. M., & Bates, N. (Eds.). (2014). *Hard-to-survey populations*. Cambridge, England: Cambridge University Press.

Thompson, S. K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association, 85*(412), 1050-1059.

Thompson, S. K. (1991a). Adaptive cluster sampling: Designs with primary and secondary units. *Biometrics, 47*(3), 1103-1115.

Thompson, S. K. (1991b). Stratified adaptive cluster sampling. *Biometrika, 78*(2), 389-397.

Thompson, S. K. (2002). *Sampling.* New York: John Wiley & Sons.

Thompson, W. (Ed.). (2004). *Sampling rare or elusive species: Concepts, designs, and techniques for estimating population parameters*. Washington, DC: Island Press.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics, 16* (2), 117-186

Wald, A. (1947). *Sequential analysis*. New York: John Wiley.

Westerman, D. A. (1991). Expert and novice teacher decision making. *Journal of teacher education, 42*(4), 292-305.

Wolter, K. (1985). *Introduction to variance estimation*. New York: Springer-Verlag.