Research Report

Model-Based Weighting and Comparisons

Jiahe Qian

April 2008

ETS RR-08-17

Model-Based Weighting and Comparisons

Jiahe Qian

ETS, Princeton, NJ

April 2008

Abstract

In survey research, sometimes the formation of groupings, or aggregations of cases on which to make an inference, are of importance. Of particular interest are the situations where the cases aggregated carry useful information that has been transferred from a sample employed in a previous study. For example, a school to be included in the sample of the High School Effectiveness (HSES) study must contain one or more cases transferred from the National Educational Longitudinal Study of 1988 (NELS:88). To calculate the aggregation inclusion probabilities, this study investigated three statistical models and, based on these models, derived the school weights for the HSES study. This study also assessed the effects of weighting by comparing the statistics yielded from different sets of weights: (a) those from an empirical population database and (b) those from data generated from simulation based on the principles of a superpopulation. Both categorical data and continuous variables were analyzed in the comparison.

Key words: Aggregation probability, model-based weighting, goodness-of-weighting, the High School Effectiveness Study

Acknowledgments

The author thanks Shelby Haberman, Bruce Spencer, Marty Frankel, and Dan Eignor for their suggestions and comments. The author also thanks Kim Fryer for editorial assistance. The work reported herein was partially supported under the National Assessment of Educational Progress (Grant No. R999G50001, CFDA No. 84.999G) as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The opinions expressed herein are solely those of the author and do not necessarily represent those of ETS.

1. Introduction

For a typical sample design, design-based weights are generally developed as the inverse of the inclusion probability for sampled units. If nondesign-based adjustments, such as adjustments for nonresponse, are ignored, the weights will yield design consistent estimates, which are also known as π estimates (Horvitz &Thompson, 1952; Kish, 1990; Särndal, Swensson, & Wretman, 1992). Some surveys, however, define only the rules for including sample units, and the inclusion probabilities are not expressed explicitly. Accordingly, the unit weights have to be derived from statistical models instead of being directly imputed by the inverse of the inclusion probability. One such survey is the High School Effectiveness Study (HSES; Ingels et al., 1994), an independent part of the National Educational Longitudinal Study of 1988 (NELS:88; Ingels, Abraham, Karr, Spencer, & Frankel, 1990). To utilize the longitudinal information of NELS:88, the sample design of the 1990 HSES Study included high schools that had 10th grade classes and that had enrolled at least one student who had been selected in the NELS:88 study. The inclusion probability of an HSES school was not defined directly. Instead, it was determined by the transfer mechanism of the students who had been selected in NELS:88 and included in the school. By this mechanism, it is possible to build a statistical model to drive the inclusion probability of the school. In what follows, each high school that contains a group of student cases is called an *aggregation*. In different contexts, such aggregations could also consist of hospitals, families, resident areas, cod, humpback whales, and so on (Bekkevold, Hansen, & Loeschcke, 2002; Palsbell et al., 1995).

A recent example from the area of educational longitudinal surveys is the supplemental study of the Early Childhood Longitudinal Study Program-kindergarten cohort (ECLS-K). It was included in the design of the eighth grade sample for the 2007 National Assessment of Educational Progress (NAEP; Allen, Donoghue, & Schoeps, 2001). The ECLS-K is an ongoing study that focuses on children's early school experiences beginning with kindergarten and through middle school (Pollack, Najarian, Rock, Atkins-Burnett, & Hausken., 2006; Rock & Pollack, 2002). By design, the ECLS-K schools in the supplemental study were identified by whether a selected NAEP school contained at least one eighth grade student from ECLS-K.

Several statistical models were developed to derive the formulas for the aggregation inclusion probability. The weighting procedure for the High School Effectiveness Study used three models. The Spencer-Foran (SF) weights in the sample were based on the hypergeometric probability model (Spencer & Foran, 1991), the Qian-Frankel (QF) weights in the sample were derived based on the binomial probability model (Qian, 1995) , and the Kaufman (K) weights were estimated by the averaging of the student NELS:88 weights in a high (Frankel & Qian, 1995).

The longitudinal nature of the HSES data provides researchers with a variety of information about family, school, community, and individual factors that are associated with school performance (Lee & Burkam, 2001; Lee, Burkam, Chow-Hoy, Smerdon, & Goverdt, 1998; Perkins, Kleiner, Roey, & Brown, 2004). These far-reaching results will be used to assess the progress of U.S. students in urban and suburban secondary schools in the 30 largest metropolitan statistical areas. But such analysis will yield adequate estimates only by employing weights appropriately provided from the HSES sample, and many users are unclear about how the weights are developed. This paper is intended to investigate the model-based weighting for the HSES school samples and to compare the weights derived from different models.

After this introduction to the surveys in this study and issues in weighting, section 2 describes the HSES study and its sample design. Section 3 describes three statistical models for deriving the aggregation probabilities for the HSES sample. Section 4 assesses the effects of weighting by comparing the statistics yielded from different sets of weights with those (a) from an empirical population database and (b) from the samples generated from simulation based on the principles of a superpopulation. Section 5 summarizes the comparisons and offers some recommendations.

2. The Sample Design of the High School Effectiveness Study

The objective of HSES is to study the effectiveness of education at the school level across different categories, such as school type and ethnicity group, and to document U.S. students" ongoing progress. It is an independent longitudinal survey executed within the NELS:88 data collection. The HSES base year is 1990, when students were high school sophomores. The data in HSES covers the same sources as the data in NELS:88, including students, parents, teachers, and administrators (Ingels et al., 1994).

To utilize existing longitudinal records from NELS:88, the study only sampled the high schools with 10th grade classes that had enrolled at least one student who had been selected in the NELS:88 study as an eighth grade student.. The sample design for this kind of survey is complicated because it involves the transition of sample units from one level of aggregation to another (Scott, Ingels, Sehra, Taylor, & Jergovic, 1996).

For the HSES base year (1990) sample, schools were selected through a two-phase sampling mechanism. In the first phase, there were 724 high schools with 10th grade classes in the 30 largest metropolitan statistical areas (MSAs) found to contain one or more students selected for NELS:88. This pool of schools was partitioned into eight strata: four types of schools (public, Catholic, NAIS, [1] and other private) at two levels of location (urban, suburban). In the second phase, a sample of 276 schools from this pool was drawn by stratified sampling (Cochran, 1977), resulting in a final baseline sample of 247 schools after the exclusion of illegible, unavailable, and nonparticipating schools.

By design, the students in a selected HSES school can be classified by whether they were in the core sample of the NELS:88 study. Those students who were are called *core cases*. From each sampled HSES school, 10th-grade students were selected through two mechanisms: (a) a subsample of all core cases, and (b) an augmentation sample of additional students who did not participate in NELS:88. The target sample size within an HSES school was approximately 30 students. The probability for inclusion of an aggregation in the first phase of HSES was determined by two factors: the selection of students in NELS:88 and the pattern of core cases that transitioned to high school. The chance for an aggregation to enroll one or more core cases, P_1 , is defined to be the *aggregation probability* (Spencer & Foran, 1991). Because the inclusion mechanism of schools did not obviously provide the aggregation probabilities, they have to be derived from statistical models.

Consistent with the sample design, the inclusion probabilities of the students in the HSES sample were determined in three steps by computation of (a) the aggregation probabilities for the 724 schools in the HSES school frame, (b) the conditional probabilities for the 247 schools selected from the HSES school frame, and (c) the conditional probabilities for students to be selected from each school. The product of these three components forms the inclusion probabilities of the students, but the main task of the weighting process is in deriving the aggregation probabilities for schools. This paper also focuses on comparing the weights derived from different models.

3. Three Probability Models for Weighting

To derive aggregation probabilities, three probability models have been proposed to describe the sampling mechanism of schools for the HSES study. Based on these models, school weights can be generalized. This section also discusses some issues related to the weighting procedure and the applications of weights, such as the effects of missing data and the analysis of cross-sectional and longitudinal data.

3.1 The Spencer-Foran Model

The Spencer-Foran (SF) model (Spencer & Foran, 1991) is based on the probability model that employs the hypergeometric distribution. Schools with eighth grade classes in the sampling frame of NELS:88 are defined as primary sampling units (PSUs) and the student cases enrolled in those schools as secondary sampling units (SSUs). In HSES, high schools are the aggregations of interest. A PSU *feeds* an aggregation if at least one case from the PSU belongs to the aggregation.

Let the SSUs in each PSU be partitioned into L strata. One such kind of partition variable is ethnicity because it affects the transition of SSUs from a school with eighth grade classes to a high school. Note that the partition of the student cases in the PSUs in NELS:88 is identical to the partition of the cases in the aggregations in HSES. Consider the PSUs to be selected with replacement, with selection probabilities proportional to a measure of size, and consider SSUs to be selected from stratum *l* in PSU$_j$ for $1 \le j \le U$ by simple random sampling without replacement. Let N_{jl} be the size of stratum *l* within PSU$_j$ and n_{jl} be the sample size drawn from N_{jl} in NELS:88. Assume $n_{jl} > 0$ unless $N_{jl} = 0$.

Consider the transition of SSUs from PSUs in NELS:88 to aggregations in HSES. Let $B_{kl|jl}$ be the number of the SSUs in stratum *l* in PSU$_j$ that transferred to stratum *l* in aggregation *k*. Then the total cases from the PSU$_j$ that transferred to the aggregation *k* are $B_{k.|j.} = \sum_{l=1}^{L} B_{kl|jl}$. The total number of SSUs in aggregation *k* is $B_{k.|..} = \sum_{j=1}^{U} B_{k.|j.}$, where U is the number of PSUs that feed aggregation *k*. Let $b_{kl|jl}$, transferred from stratum *l* in PSU$_j$ to stratum *l* in aggregation, be the number of core cases in $B_{kl|jl}$. Table 1 displays the transition of the cases from stratum *l* in PSU$_j$ in NELS:88 to an aggregation in HSES.

Table 1

Transition of the Cases From Stratum l in PSU$_j$ and in the National Educational Longitudinal Study of 1988 (NELS:88) to Stratum l in Aggregations in the High School Effectiveness Study (HSES)

In NELS:88	In aggregation k		
	Yes	No	Total
Yes	$b_{kl\|jl}$	$n_{.l\|jl} - b_{kl\|jl}$	$n_{.l\|jl}$
No	$B_{kl\|jl} - b_{kl\|jl}$	$N_{.l\|jl} - B_{kl\|jl} - n_{.l\|jl} + b_{kl\|jl}$	$N_{.l\|jl} - n_{.l\|jl}$
Total	$B_{kl\|jl}$	$N_{.l\|jl} - B_{kl\|jl}$	$N_{.l\|jl}$

Note. To include the aggregation information, the symbols N_{jl} and n_{jl} are expressed as $N_{.l|jl}$ and $n_{.l|jl}$.

Given that PSU$_j$ was selected, the probability model of the hypergeometric distribution can be employed in the computation of the conditional probability of $b_{kl|jl}$. Then, the conditional probability that $b_{kl|jl} = 0$ can be calculated:

$$T_{kl|jl} = H\left(n_{jl}, B_{kl|jl}, N_{jl}\right) = \begin{cases} \dfrac{\left(N_{jl} - B_{kl|jl}\right)!\left(N_{jl} - n_{jl}\right)!}{N_{jl}!\left(N_{jl} - B_{kl|jl} - n_{jl}\right)!}, & \text{if } N_{jl} - B_{kl|jl} - n_{jl} \geq 0; \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Given that PSUj was in NELS:88, the conditional probability that none of the core cases fed by PSUj in aggregation k is $T_{k.|j.} = \prod_{l=1}^{L} T_{kl|jl}$, and the conditional aggregation probability for enrolling at least one core case in aggregation k is

$$P_{k|}^{SF} = 1 - T_{k.|..} = 1 - \prod_{j=1}^{U} T_{k.|j.} \tag{2}$$

Let $R_{k.|j.} = 1 - \pi_j + \pi_j T_{k.|j.}$ By the SF model, the aggregation probability for k equals

$$P_1^{SF} = 1 - \prod_{j=1}^{U} R_{k.|j.} \, , \tag{3}$$

where π_j is the probability that PSUj was selected in NELS:88. An approximation of π_j is the reciprocal of the school weight for PSUj (Kish, 1992).

Based on $\left\{P_{k|}^{SF}\right\}$ in (2) and $\left\{P_1^{SF}\right\}$ in (3), two sets of school weights can be computed separately. When the study is interested in analyzing the schools in the HSES sample, the weights based on $\left\{P_{k|}^{SF}\right\}$should be applied. When the study is interested in analyzing the pool of schools in NELS:88 and in HSES , the weights based on $\left\{P_1^{SF}\right\}$should be applied. The HSES data file only included the weights derived based on $\left\{P_1^{SF}\right\}$.

The SF model for data without partitioning information. In SF model-based weighting, one obstacle is the burden of collecting the information of the L strata in all of the U PSUs that feed the aggregations in HSES, such as N_{jl}, n_{jl}, $B_{kl|jl}$, and so on. Moreover, there are always missing data and coding errors in the data collected. For example, among 1,823 schools with eighth grade classes collected in the feeder data file, there are 61 schools missing MSA indices. To impute the selection probabilities in NELS:88 for these 61 schools, extra information for the sampling frame must be collected. Missing data imputation is a vital part of the HSES weighting process. To address this issue, Spencer and Foran (1991) proposed a simplified model that ignored the stratification of SSUs because nonstratified information, such as $N_{j.}$, $n_{j.}$, $B_{k.|j.}$, are usually readily available. They also suggested some methods to impute the missing data.

The effects of missing data on the SF model. The missing data, including undercounts of the feeder school data or undercounts of the students, will trim the estimates of the aggregation probabilities derived from the SF model. This situation implies that the weight, approximated by the reciprocal of the inclusion probability, will be inflated. Let $\mathrm{U}^{'}$ be the number of feeder PSUs with information in the data file and U be the actual number of feeder PSUs. Assume $\mathrm{U}^{'} < \mathrm{U}$. So

$$P_1^{SF} = 1 - R_{k.|1.} R_{k.|2.} \cdots R_{k.|U.} > P_1^{SF'} = 1 - R_{k.|1.} R_{k.|2.} \cdots R_{k.|U'.} .$$

This shows that the estimates derived from the SF model will be smaller than it should be when there are data missing from the feeder data file. The small probabilities will produce extremely large weights. Accordingly, the design effects of weighting will be large because of the large variations among weights.

Let $\gamma = R_{k.|U'+1.} \cdots R_{k.|U.}$. Thus

$$\frac{1-P_1^{SF'}}{1-P_1^{SF}} = \frac{R_{k.|1.}R_{k.|2.}\cdots R_{k.|U'.}}{R_{k.|1.}R_{k.|2.}\cdots R_{k.|U'.}R_{k.|U'+1.}\cdots R_{k.|U.}} = \frac{1}{\gamma},$$

and the relationship between P_1^{SF} and $P_1^{SF'}$ can be expressed as $P_1^{SF} = 1 - \gamma + \gamma P_1^{SF'}$.

3.2 The Qian-Frankel Model

The basis for the QF model is the probability model of the binomial distribution. As an alternative approach, the QF model attempts to preserve the design-based properties of the SF model, while eliminating the burden of obtaining the feeder pattern information required by the SF model.

Given the basic sample design, assume that students in each aggregation possess similar characteristics related to the NELS:88 sample design, such as school type, area region, location, and so on. Therefore, the model assumes that the students in the same aggregation have about the same chance to be included in NELS:88. Note that this is different from the assumption that students have the same chance to move to an aggregation.

Let PSU_j be a school selected in NELS:88 and M_k be the size of aggregation *k*. Let $S_k = P\left\{i_k \in \mathrm{PSU}_j\right\}$ be the chance of case *i,* in aggregation *k,* to be included in NELS:88. It is reasonable to assume that two events, a student transitioning to a high school and a student being selected in the NELS:88 sample, are independent. Let ξ_k be the variate of having core cases in aggregation *k*. Assuming that the chance for students in aggregation *k* to be included in the NELS:88 study is homogeneous, ξ_k forms a binomial distribution with index M_k and probability S_k:

$$P\left(\xi_k = t\right) = \binom{M_k}{t} S_k^t \left(1-S_k\right)^{M_k - t} . \qquad (4)$$

The probability that aggregation k contains no core cases equals

$$Q_k = P\left(\xi_k = 0\right) = \left(1-S_k\right)^{M_k} . \qquad (5)$$

and the probability that aggregation k enrolls at least one core case is

$$P_k^{QF} = P\left(\xi_k > 0\right) = 1 - Q_k \quad . \tag{6}$$

Since HSES is a continual study of NELS:88, P_k^{QF} in (6) is regarded as conditional aggregation probability.

Consider estimation of S_k. Let m_k be the number of core cases in the sample in aggregation *k*. Let $W_{i,k}^{88}$ be the case weight of core case *i* in aggregation *k* and $P_{i,k}^{88}$ be its inclusion probability. The case weights in NELS:88 are approximately equal to the reciprocal of the inclusion probability, $W_{i,k}^{88} = 1/P_{i,k}^{88}$. The inclusion probability of the core cases in aggregation *k,* S_k, is estimated by a harmonic mean:

$$\hat{S}_k = \left(\frac{1}{m_k} \sum_{i=1}^{m_k} \frac{1}{P_{i,k}^{88}} \right)^{-1} \quad . \tag{7}$$

The estimate can also be expressed by the reciprocal of the average of the weights of the core cases:

$$\hat{S}_k = \frac{m_k C}{\sum_{i=1}^{m_k} W_{i,k}^{88}} \quad , \tag{8}$$

where C is a constant used to normalize the summation of school weights. Accordingly, by (8), $\hat{Q}_k = \left(1 - \hat{S}_k\right)^{M_k}$ and $\hat{P}_k^{QF} = 1 - \hat{Q}_k$. As a result, the estimated weight for aggregation *k* will be the reciprocal of $\hat{P}_k^{QF}$. As pointed out before, the school weights derived from $\left\{\hat{P}_k^{QF}\right\}$ on the HSES sample are appropriate for analyzing school data in the HSES study.

The QF model for data with partitioning information. The weighting can be improved by partitioning the students into *H* strata as was done in the SF model. Although certain student groups are sometimes oversampled, the students in the same group have about the same chance to be selected in NELS:88. The variables used for stratification could be feeder schools, ethnicity, a combination of them, and so on. Employing partitioning information in weighting will allow the model design to be fit appropriately. For example, oversampling students from certain ethnic groups will cause unequal probability of selection. But the homogeneity in the probability of selection usually remains on hold within ethnic groups. Let $M_{k,h}$ be the size of stratum *h* in aggregation *k*. Let $S_{k,h}$ be the chance of student *i*, belonging to stratum *h* in aggregation *k*, to be included in NELS:88, and $\xi_{k,h}$ be the binomial variate of having core cases in stratum *h* in aggregation *k*. Thus,

$$Q_{k,h} = P\left(\xi_{k,h} = 0\right) = \left(1 - S_{k,h}\right)^{M_{k,h}} . \tag{9}$$

and the conditional aggregation probability that the aggregation k contains at least one core case is

$$P_{k.}^{QF} = P\left(\xi_{k.} > 0\right) = 1 - \prod_{h=1}^{H} Q_{k,h} . \tag{10}$$

Let h be the index for the PSUs in NELS:88. By the QF model, the aggregation probability for k equals

$$P_{1}^{QF} = 1 - \prod_{h=1}^{H} \left[1 - \pi_h + \pi_h Q_{k,h}\right] , \tag{11}$$

where π_h is the probability that PSU_h was selected in NELS:88.

Consider the estimation of $Q_{k,h}$. Let $m_{k,h}$ be the number of core cases in stratum *h* in aggregation *k*. Let $W_{i,k,h}^{88}$ be the NELS:88 case weight of core case *i* in stratum *h* in aggregation *k*. Similar to Q_k, $Q_{k,h}$ can be estimated by

$$\hat{Q}_{k,h} = \frac{m_{k,h} C}{\sum_{i=1}^{m_{k,h}} W_{i,k,h}^{88}} . \tag{12}$$

So the conditional aggregation probability $P_{k.}^{QF}$ in (10) and the aggregation probability P_{1}^{QF} in (11) can be estimated. Thus, the cross-sectional and the longitudinal school weights for aggregation k equal the reciprocal of $\hat{P}_{k.}^{QF}$ and $\hat{P}_{1}^{QF}$, respectively.

The effects of missing data on the QF model. Differing from the SF model, the QF model is not sensitive to missing data because the aggregation probability derived from the QF model uses only the NELS:88 weights for all the core cases transferred into an aggregation. There are usually no missing data among the base year weights. Therefore, missing data are not an issue for the QF model. Although partition information is needed when applying the improved QF model, these data are easier to collect.

3.3 The Kaufman Model

The K model, independently proposed by Kaufman and by Spencer, estimates school weights by the average of the student NELS:88 weights in a high school. As a means for reducing the cost of collecting a feeder pattern data, the following is an alternative estimator of the total population of X of interest,

$$\hat{X} = \sum_{k=1}^{A} \frac{1}{M_k} \sum_{i=1}^{M_k} W_{i,k}^{88} \; I\,(i_k \in J \mid k \in K) \cdot I(k \in K) X_k \quad . \tag{13}$$

In (13), $W_{i,k}^{88}$ is the weight for the student case selected in NELS:88; J represents the NELS:88 student sample; K represents aggregations on the frame of the second phase of school selection; M_k is the size of aggregation k that contains at least one core case, and i_k represents the student in aggregation k $\left(1 \le k \le A\right)$. In addition, $I(k \in K\,) = 1$ is defined as k in the HSES frame and as 0 otherwise; and $I(i_k \in J \mid k \in K\,) = 1$is defined as i_k in NELS:88 given k in the frame and as 0 otherwise.

Given that aggregation *k* was on the frame, the weight for X_k is its coefficient $M_k^{-1} \sum_{i=1}^{m_k} W_{i,k}^{88}$, where m_k is the number of core cases in aggregation *k*. Use the symbol in (7): $\hat{S}_k = \left(m_k^{-1} \sum_{i=1}^{m_k} W_{i,k}^{88}\right)^{-1}$. So the inclusion chance of the aggregation can be approximated by the reciprocal of its weight

$$\hat{P}_1^K = \frac{M_k}{m_k} \hat{S}_k \quad . \tag{14}$$

The estimated chance $\hat{P}_1^K$ has an inverse relationship with the number of core cases transferred into the aggregation, which seems contrary to what is to be expected. The K weights yield unbiased estimates on average of all possible NELS:88 samples and all possible HSES samples.[2] However, the estimates would not be unbiased for a specific NELS:88 sample that has been collected.

4. Comparison of the Weights Derived From Different Models

All three sets of derived school weights were considered in comparison. To avoid the confounding effects of raking, the weights used in comparison were trimmed but not raked. The final weights on the HSES data file were all adjusted by the trimming[3] and raking[4] procedures.

To compare the weights derived from different models, one can use *goodness-of-weighting,* a measure of the closeness between the weighted sample distributions and the estimate of population distribution. For categorical variables, the χ^2 statistic was used to test the agreement between the weighted sample distributions and population distributions and the *F*-test was used to test the results derived from two different sets of weights in comparison. For continuous variables, bias and mean square error (MSE) were used as the criteria.

In comparison, the sample distributions were obtained by two mechanisms: (a) samples drawn from an empirical population database and (b) simulation-based samples from populations drawn from a superpopulation. For the data from mechanism (a), the weighting effects are examined by comparing the estimates for some characteristics from the sample with the true characteristics from the population. Due to finite population sampling, the variance formulae for survey data must include a finite population correction factor (fpc), either explicitly or implicitly. Moreover, because of complex sampling design, this study estimated the variances by using the delta method, which is based on a second degree Taylor series expansion (Cochran, 1977; Wolter, 1985) and incorporates the fpc implicitly. Under the superpopulation model, the simulation adopts a stochastic viewpoint in comparisons and views each finite population as a random sample drawn with replacement from a hypothetical infinite population (Deming & Stephan, 1941). Each finite population is a reflection of the possible relationships among different variables observed, and the replicate samples may be used to simulate the relationships of interest.

4.1 Comparison Based on Empirical Data

The Quality Education Data's (QED) database[5] had served as the *true population* in the analysis. Note that the sampling frame of the schools for the HSES study was created based on the QED database. The QED file used contains 4,628 high schools in the largest 30 MSAs and a number of school-level characteristics for each school on its list. Some of the variables of the characteristics on the list were categorical while a few were continuous. For each variable of interest, the corresponding values for the 247 schools in the sample were used in conjunction with the school weights to produce three separate weighted estimates.

Table 2 displays the χ^2 statistics for 12 categorical variables under three sets of weights. The comparisons of the different weights are shown by the results of the *F*-test with α level at 0.25. The α level was set differently from regular tests because it was used to evaluate goodness-of-weighting. In general, the SF and QF weighting models provided closer agreement between the weighted sample distributions and population distributions.

Table 2

The χ^2 Test for Some Variables in the Quality Education Data (QED) File

	df	SF weight	K weight	QF weight
Teacher population code	5	$94.86^{K,Q}$	1782.84	190.39^{K}
School type	2	$60.23^{K,Q}$	2091.53	328.14^{K}
Grade level	7	$131.26^{K,Q}$	524.87	298.03^{K}
Instruction dollars per pupil	6	78.86^{K}	277.17	$32.25^{K,S}$
MSA	29	282.57	$82.46^{S,Q}$	154.36^{S}
Location	1	161.87^{K}	1452.05	284.81^{K}
Personnel gender	1	$8.94^{K,Q}$	1631.49	85.54^{K}
# of students code	8	89.03^{K}	1159.55	130.09^{K}
# of teachers code	8	$52.73^{K,Q}$	1003.85	148.09^{K}
State postal code	19	129.24	104.36	$73.79^{K,S}$
Enrollment change (building)	6	118.70^{K}	252.09	113.54^{K}
Region	3	62.57^{K}	515.37	127.00^{K}

Note. The weights are trimmed & nonraked. To compare the goodness of fit of the distribution of a certain variable between two sets of weights, an *F*-test at α=.25 is used to test the significance of difference. The letters SF, K and QF stand for Spencer-Foran, Kaufman, and Qian-Frankel weights, respectively. When a letter appears as a superscript of a number, it means that the distribution under the weights of that column is fitted better than the distribution under the weights represented by the letter in the superscription. For example, in the first column for the SF weight on the table, $94.86^{K,Q}$ shows that the distribution of teacher population code is fitted better under the SF weights than under the K or QF weights. MSA = metropolitan statistical areas.

For the four continuous variables that were available in the QED file, the results are summarized in Table 3. Let $\hat{\theta}_{\mathbf{W}_\tau}$ be the weighted estimate of a school characteristic under a weight vector $\mathbf{W}_\tau$ and τ be the index for the three sets of the weights. The bias of $\hat{\theta}_{\mathbf{W}_\tau}$ is defined as $\mathrm{E}\left(\hat{\theta}_{\mathbf{W}_\tau}\right)-\theta_P$, and the MSE of $\hat{\theta}_{\mathbf{W}_\tau}$ is $\mathrm{E}\left(\hat{\theta}_{\mathbf{W}_\tau}-\theta_P\right)^2$, where θ_P is the true school characteristic of the QED population. For a weighted mean estimate $\bar{\theta}_{\mathbf{W}_\tau}$, the estimate of its bias, $\hat{\mathrm{b}}\left(\bar{\theta}_{\mathbf{W}_\tau}\right)$, is $\bar{\theta}_{\mathbf{W}_\tau}-\theta_P$ and the estimate of its MSE equals $\hat{\mathrm{V}}\left(\bar{\theta}_{\mathbf{W}_\tau}\right)+\hat{\mathrm{b}}^2\left(\bar{\theta}_{\mathbf{W}_\tau}\right)$, where $\hat{\mathrm{V}}\left(\bar{\theta}_{\mathbf{W}_\tau}\right)$, the variance estimate, is calculated by the delta method. Table 3 shows the estimates applying the SF weights yielded the smallest bias for all four variables, while those applying the K weights produced the largest. Although the comparisons in Table 2 and on the top part of Table 3 were based on the weights that were trimmed but not raked, the bottom part of Table 3 also displays the results based on the weights that were nontrimmed and nonraked. Because the QF model is not sensitive to missing data, there were no extreme weights and the trimming procedure was not applied to the QF weights. Under the nontrimmed and nonraked weights, the results are similar to those under the trimmed and nonraked weights. Apparently, to draw certain conclusions, the comparisons needed more data, to be either collected by survey or obtained by simulation.

4.2 Comparison Based on Simulation Approach

To obtain certainty in comparison, Monte Carlo simulation was used to approximate the exact sampling distribution of weighted estimates by drawing a large number of samples from a fixed population for a specific design (Liu, 2001; Särndal et al., 1992). Based on the outcome of the repeated samples, the errors of the statistics of interest can be determined as a function of the number of trials and other quantities. In assessing the weighting effects, the Monte Carlo method needs to be set up differently from regular simulation. Because of substantial computation and lack of auxiliary data for weighting, it is unlikely to impute a separate set of weights for each replicate sample. Instead, the sampling design was preserved, weights were kept fixed, and replicate samples were drawn from finite population with the values of the variable of interest. From the realized samples from finite population, the variation of weighted estimates can be obtained. Note that the population in simulation could be treated as a sample drawn from superpopulation (Cochran, 1977).

Table 3

The Bias and Root Mean Square Errors of Approximation (RMSA) of the Mean Estimates for Some Variables in the Quality Education Data (QED) File

Variable	Population mean	SF weight		K weight		QF weight	
		Bias	RMSA	Bias	RMSA	Bias	RMSA
Trimmed & nonraked weights							
No. of White students	944.26	-54.42	115.39	190.33	199.57	64.47	117.60
No. of Black students	273.94	-72.50	78.01	81.56	90.24	80.14	100.15
No. of Hispanic students	190.52	-63.07	66.64	103.65	108.27	86.17	93.88
No. of teachers	51.59	1.82	4.85	29.51	29.69	13.51	14.56
Nontrimmed & nonraked weights							
No. of White students	944.26	-64.04	120.93	209.26	219.80	64.47	117.60
No. of Black students	273.94	-73.70	79.07	80.95	89.51	80.14	100.15
No. of Hispanic students	190.52	-65.01	68.51	98.01	102.85	86.17	93.88
No. of teachers	51.59	-4.40	7.74	28.39	28.61	13.51	14.56

Note. The calculations for each statistic are based on those cases with nonmissing and nonzero values. To include the impact of the complex sample design and weighting, the standard errors of the weighted means are calculated by the delta method. The results for the average student counts in school are dropped because, according to the QED database that was used, the population mean of the average student counts is smaller than the population mean of the White student counts in school. SF = Spencer-Foran, K = Kaufman, QF = Qian-Frankel.

Let *N* be the size of finite population and *n* be the sample size. Let Y_k be the variable of interest and X_k be the known auxiliary value for $1 \le k \le N$. Let the population values of $\{y_k\}$ be a realization of $\{Y_k\}$ and y_k form a general linear model

$$y_k = X_k + \varepsilon_k \,, \tag{15}$$

where $\mathrm{E}(\varepsilon_k) = 0$ and $\mathrm{V}(\varepsilon_k) = \sigma_k^2$. The auxiliary values of $\{X_k\}$ are based on four continuous variables on the QED file. In particular, assume $\varepsilon_k \overset{iid}{\sim} N(0, \sigma_k^2)$ and, to set the coefficient of variation (cv) as a constant, $\sigma_k = cX_k$. A simulation probed the robustness of weighting by setting cv at values of 0.2 and 0.8 separately.

In assessing the weighting effects for a given sample design, consider the criteria of unbiasedness and MSE. Let θ_ρ be the mean of the school characteristic of interest, which was calculated from $\{X_k\}$. Its weighted estimate,

$$\bar{\theta}_{\mathbf{W}_\tau} = \frac{\sum_{k=1}^{n} w_{\tau,k} y_k}{\sum_{k=1}^{n} w_{\tau,k}} \,, \tag{16}$$

was calculated from $\{y_k\}$ in (15) under $\mathbf{W}_\tau$ with index τ for weight sets. The symbol $\bar{\theta}_{\mathbf{W}_\tau,a}$ was the estimate in *a*th repetition under $\mathbf{W}_\tau$. Define $\bar{\bar{\theta}}_{\mathbf{W}_\tau,\cdot} = A^{-1}\sum_{a=1}^{A} \bar{\theta}_{\mathbf{W}_\tau,a}$, where *A* is the total number of repetitions and *A* = 1,000 in simulation. Given sample design and weights, the bias estimate of $\bar{\theta}_{\mathbf{W}_\tau}$ was $\bar{\bar{\theta}}_{\mathbf{W}_\tau,\cdot} - \theta_\rho$, and the MSE estimate of $\bar{\theta}_{\mathbf{W}_\tau}$ was $(A-1)^{-1}\sum_{a=1}^{A}\left(\bar{\theta}_{\mathbf{W}_\tau,a} - \bar{\bar{\theta}}_{\mathbf{W}_\tau,\cdot}\right)^2 + \left(\bar{\bar{\theta}}_{\mathbf{W}_\tau,\cdot} - \theta_\rho\right)^2$.

The results of the simulation approach are summarized in Table 4. Although the estimates under the SF weights still yielded the smallest bias for all four variables, as was the cases with the empirical approach, they are much closer here to those under the QF weights. The bias of the estimates applying the K weights was still largest. When the cv in the simulation changed from 0.2 to 0.8, the bias estimates under any of the three sets of weights had only very minor changes. This implied that the bias estimates were robust under different weighting procedures. As shown in the simulation results in Table 4, the MSEs for the estimates under the SF weights were slightly smaller than those under the QF weights.

Table 4

The Bias and Root Mean Square Error of Approximation (RMSA) of the Mean Estimates in Simulation

Variable	Population mean	SF weight		K weight		QF weight	
		Bias	RMSA	Bias	RMSA	Bias	RMSA
Coefficient of variation (cv) = 0.2 in models							
No. of White students	944.26	-56.12	74.12	189.48	192.27	63.18	82.53
No. of Black students	273.94	-72.92	73.26	81.36	82.29	79.77	80.90
No. of Hispanic students	190.52	-63.14	63.22	103.67	103.99	86.15	87.07
No. of teachers	51.59	1.80	2.94	29.49	29.04	13.45	13.75
Coefficient of variation (cv) = 0.8 in models							
No. of White students	944.26	-59.36	207.80	186.25	229.70	61.23	223.20
No. of Black students	273.94	-72.34	77.96	82.45	96.56	81.39	102.80
No. of Hispanic students	190.52	-63.20	64.73	103.19	108.48	85.73	100.55
No. of teachers	51.59	1.62	9.75	29.16	30.00	13.03	17.27

Note. Repetition = 1,000. To include the impact of the complex sample design and weighting, an adjustment has been made to the standard errors of the weighted means by multiplying the standard errors by DEFT (Kish, 1965). DEFT were calculated by the delta method from QED data set. The average student counts in school are not included in the simulation because, according to the QED database that was used, the population mean of the average student counts is smaller than the population mean of the White student counts in school. SF = Spencer-Foran, K = Kaufman, QF = Qian-Frankel.

5. Conclusions

In general, the comparison of the effects of different weighting schemes could be summarized based on five factors:

1. Consistency between the weighting probability model and the sample design
2. Consistency between the weighting conventions and the estimation of aggregation probability or conditional aggregation probability
3. Bias in estimation (for continuous variables)
4. Goodness-of-fit (for categorical variables)
5. Cost in data collection for weighting

Table 5 summarizes the results of the comparisons. For Factor 2, both the SF and QF models that yield the aggregation inclusion probabilities are consistent with the sample design. For Factor 2, the school probabilities derived either from the SF model or from the QF model are consistent with the weighting conventions. In general, the relationship between the aggregation probability and the number of the core cases in the aggregation is not simply a linear one or an inverse one. For each set of weights, based on the analysis in section 4, Table 5 lists the characteristics for Factors 3, 4, and 5.

Table 5

Summary of the Comparisons

	Three types of weights		
Property	SF weight	K weight	QF weight
a	Good	N/A	Good
b	Good	N/A	Good
c	Small	Largest	Middle
d	Good	Medium	Good
e	High	Low	Low

Note. SF = Spencer-Foran, K = Kaufman, QF = Qian-Frankel.

The results from the comparison offer some tips to model-based weighting and to application of the weights. In deriving the weights for surveys like HSES that involve the transition of sample units from one level of aggregation to another, the QF model is preferred because the cost to collect information for weighting is low, the computation is not so complicated, and less missing data need to be imputed. Particularly, when needed partitioning information is available, the QF model gains additional accuracy in weighting without the burden of having to collect as much extra data as for the SF model.

In estimation, the two sets of weights, derived either from the QF model or from the SF model, function approximately the same. The simulation results show that the SF weights are as effective as the QF weights for the cross-sectional analysis of the schools in HSES. In theory, the set of the SF weights on the HSES file is more suitable for analyzing the pool of the schools in NELS:88 and in HSES. As a rule of thumb, if a set of weights is predetermined, it should be used through the whole analysis. In addition to educational data, the model-based weighting has the potential to extend its application to other fields, such as health care, agriculture, zoology, and ecology.

References

Allen, N., Donoghue, J., & Schoeps, T. (2001). *The NAEP 1998 technical report* (NCES 2001-509). Washington, DC: National Center for Education Statistics.

Bekkevold, D., Hansen, M. M., & Loeschcke, V. (2002). Male reproductive competition in spawning aggregations of cod. *Molecular Ecology, 11,* 91–102.

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

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal tables are known. *Annals of Mathematical Statistics, 11,* 427–444.

Deming W. E., & Stephan, F. F. (1941). On the interpretation of censuses as samples. *Journal of the American Statistical Association, 36*(213), 45–49.

Frankel, M. R., & Qian, J. (1995, April). *Sample weighting for multi-level analysis in panel designs.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Haberman, S. J. (1979). *Analysis of qualitative data* (vol. 2). New York: Academic Press.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47,* 663–685.

Ingels, S. J., Abraham, S. Y., Karr, R., Spencer, B. D., & Frankel, M. R. (1990). *National Education Longitudinal Study of 1988: Base year: Student component data file user's manual* (NCES 90-464). Washington, DC: National Center for Education Statistics.

Ingels, S. J., Dowd, K. L., Baldridge, J. D., Stipe, J. L., Bartot, V. H, & Frankel, M R. (1994). *National Education Longitudinal Study of 1988. Second follow-up: Student component data file user's manual* (NCES 94-374). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Kish, L. (1965). *Survey sampling.* New York, Wiley.

Kish, L. (1990). Weighting: Why, when, and how? *Proceedings of the section on survey research methods* (pp. 121–130). Alexandria, VA: American Statistical Association.

Kish, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, *8*(2), 183–200.

Lee, V. E., & Burkam, D. T. (2001, January). *Dropping out of high school: The role of school organization and structure.* Paper presented at the Conference of dropouts in America:

How severe is the problem? What do we know about intervention and prevention? Cambridge, MA.

Lee, V. E., Burkam, D. T., Chow-Hoy, T., Smerdon, B.A., & Goverdt, D. (1998). *High school curriculum structure: Effects on coursetaking and achievement in mathematics for high school graduates* (NCES 98-09). Washington DC: National Center for Education Statistics.

Liu, J. S. (2001). *Monte Carlo strategies in scientific computing.* New York: Springer-Verlag.

Palsbell, P. J., Clapham, P. J., Mattila, D. K., Larsen, F., Sears R., Siegismund, H. R., et al. (1995). Distribution of mtDNA haplotypes in North Atlantic humpback whales: the influence of behavior on population structure. *Marine Ecology Progress Series, 116*, 1–10.

Perkins, R., Kleiner, B., Roey, S., & Brown, J. (2004). *The High School Transcript Study: A decade of change in curricula and achievement, 1990-2000*. (NCES 2004-455). Washington DC: National Center for Education Statistics.

Pollack, J. M., Najarian, M., Rock, D. A., Atkins-Burnett, S., & Hausken, E. G. (2006). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric report for the fifth grade* (NCES 2006–036). Washington, DC: National Center for Education Statistics.

Potter, F. J. (1990). A study of procedures to identify and trim extreme sampling weights. *Proceedings of the section on survey research methods* (pp. 225–230). Alexandria, VA: American Statistical Association.

Qian, J. (1995). A comparison of weights derived from different models. *Proceedings of the section on survey research methods* (pp. 912–916). Alexandria, VA: American Statistical Association.

Rock, D. A., & Pollack, J. (2002). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric report for kindergarten through first grade* (NCES 2002-05). Washington, DC: National Center for Education Statistics.

Särndal, C., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Scott, L. A., Ingels, S. J., Sehra, S., Taylor, J. R., & Jergovic, D. (1996, March). *High School Effectiveness Study: Data file user's manual. National Educational Longitudinal Study of 1988*. Washington, DC: National Center for Education Statistics.

Spencer, B. D., & Foran, W. (1991). Sampling probabilities for aggregations, with application to NELS:88 and other educational longitudinal surveys. *Journal of Education Statistics, 16,* 21–34.

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

Notes

[1] Independent private schools are members of the National Association of Independent Schools (NAIS).

[2] The unbiasedness of $\hat{X}$ is verified in the appendix.

[3] The trimming process truncates extreme weights caused by unequal probability sampling or by poststratification adjustment. It reduces variation caused by extremely large weights but introduces some bias in estimates. The process usually employs the criterion of minimum mean squared error (Potter, 1990).

[4] Raking refers to the procedure that uses the Deming-Stephan algorithm to adjust weights in the sample to make the weighted marginal distributions of the sample agree with the marginal distributions of the population on specified demographic variables (Deming & Stephan, 1940; Haberman, 1979). The final weights on the data file were all adjusted by the raking process, matching weights with the marginal distributions specified by six variables obtained from the Quality Education Data's (QED) database.

[5] The National Education Database was constructed and is maintained by Quality Education Data (QED). The database covers U.S. and Canadian educational institutions and includes childcare centers, elementary schools, middle/junior high schools, senior high schools, colleges, libraries, school personnel, and district personnel. It provides the national list of school districts and schools and demographic information and is updated regularly. The QED database has been used as the sampling frame for many educational surveys.

Appendix
The Verification of Unbiasedness of the Kaufman Weights

The estimator of the population total of X based on the K weights is

$$\hat{X} = \sum_{k=1}^{A} \frac{1}{M_k} \sum_{i=1}^{M_k} W_{i,k}^{88} \ I\,(i_k \in J \mid k \in K) \cdot I(k \in K) X_k \,,$$

where

$$I\left(k \in K\right) = \begin{cases} 1\,, & \text{if k is in SES frame,} \\ 0\,, & \text{otherwise;} \end{cases}$$

and

$$I\left(i_k \in J \mid k \in K\right) = \begin{cases} 1\,, & \text{if } i_k \text{ is selected in NELS:88 given } k \text{ in SES,} \\ 0\,, & \text{otherwise.} \end{cases}$$

In the estimate, $W_{i,k}^{88}$ is the weight for the student case selected in NELS:88, which approximately equals the reciprocal of the inclusion probability; J represents the NELS:88 student sample; K represents aggregations with size M_k in the frame of the second phase of school selection that contain at least one core case; and i_k represents the student in aggregation k. The estimate $\hat{\mathrm{X}}$ will be unbiased about the expectation of the random of the NELS:88 sampling and HSES sampling. The expectation of $\hat{\mathrm{X}}$ equals

$$\begin{aligned} E\left(\hat{X}\right) &= E_k\left(E_{j|k}\left(\hat{X}\right)\right) \\ &= E_k\left(\sum_{k=1}^{A} I(k \in K) X_k \frac{1}{M_k} \sum_{i=1}^{M_k} W_{i,k}^{88} \cdot E_{j|k}\left(I\,(i_k \in J \mid k \in K)\right)\right) \\ &= E_k\left(\sum_{k=1}^{A} I(k \in K) X_k \frac{1}{M_k} \sum_{i=1}^{M_k} W_{i,k}^{88} \cdot P\,(i_k \in J \mid k \in K)\right). \end{aligned}$$

Because, in the first stage of the HSES study, all the HSES schools are listed that contain at least one core case, $P\,(k \in K \mid i_k \in J) = 1$. Hence,

$$P\,(i_k \in J \mid k \in K) = \frac{P\,(i_k \in J \wedge k \in K)}{P\,(k \in K)} = \frac{P\,(i_k \in J)}{P\,(k \in K)}\,.$$

Therefore,

$$
\begin{aligned}
E\left(\hat{X}\right) = {} & E_k\left(\sum_{k=1}^{A} I(k \in K) X_k \cdot \frac{1}{M_k}\sum_{i=1}^{M_k} W_{i,k}^{88} \cdot \frac{P(i_k \in J)}{P(k \in K)}\right) \\
& = E_k\left(\sum_{k=1}^{A} I(k \in K) X_k \cdot \frac{1}{P(k \in K)}\right) \\
& = \sum_{k=1}^{A} X_k \\
& = X.
\end{aligned}
$$