NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

The Working Paper Series was initiated to promote the sharing of the valuable work experience and knowledge reflected in these preliminary reports. These reports are viewed as works in progress, and have not undergone a rigorous review for consistency with NCES Statistical Standards prior to inclusion in the Working Paper Series.

U. S. Department of Education
Office of Educational Research and Improvement

NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

A Study of Variance Estimation Methods

Working Paper No. 2001–18 September 2001

Contact: Ralph Lee
Statistical Standards Program
E-mail: ralph.lee@ed.gov

U. S. Department of Education
Office of Educational Research and Improvement

U.S. Department of Education
Rod Paige
Secretary

Office of Educational Research and Improvement
Grover J. Whitehurst
Assistant Secretary

National Center for Education Statistics
Gary W. Phillips
Acting Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Office of Educational Research and Improvement
U.S. Department of Education
1990 K Street NW
Washington, DC 20006

September 2001

The NCES World Wide Web Home Page is
http://nces.ed.gov

Suggested Citation

U.S. Department of Education. National Center for Education Statistics. *A Study of Variance Estimation Methods*. Working Paper No. 2001–18, by Fan Zhang, Stanley Weng, Sameena Salvucci, and Ming-xiu Hu. Project Officer, Ralph Lee. Washington, DC: 2001.

Foreword

In addition to official NCES publications, NCES staff and individuals commissioned by NCES produce preliminary research reports that include analyses of survey results, and presentations of technical, methodological, and statistical evaluation issues.

The ***Working Paper Series*** was initiated to promote the sharing of the valuable work experience and knowledge reflected in these preliminary reports. These reports are viewed as works in progress, and have not undergone a rigorous review for consistency with NCES Statistical Standards prior to inclusion in the Working Paper Series.

Copies of Working Papers can be downloaded as pdf files from the NCES Electronic Catalog (http://nces.ed.gov/pubsearch/), or contact Sheilah Jupiter at (202) 502–7444, e-mail: sheilah_jupiter@ed.gov, or mail: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 1990 K Street NW, Room 9048, Washington, DC 20006.

Marilyn M. Seastrom
Chief Mathematical Statistician
Statistical Standards Program

Ralph Lee
Mathematical Statistician
Statistical Standards Program

This page intentionally left blank.

A Study of

Variance Estimation Methods

Prepared by:

Fan Zhang
Stanley Weng
Sameena Salvucci
Ming-xiu Hu
Synectics for Management Decisions, Inc.

Prepared for:

U.S. Department of Education
Office of Educational Research and Improvement
National Center for Education Statistics

September 2001

This page intentionally left blank.

Table of Contents

This page intentionally left blank.

An Empirical Study of Poststratified Estimator

Fan Zhang

Introduction

In National Center for Education Statistics (NCES) surveys, ordinary poststratification and raking ratio adjustment are commonly used techniques for improving the precision and reducing the bias of estimators. Generally speaking, poststratification refers to any method of data analysis which involves forming units into homogeneous groups after observation of the sample, especially for those cases where additional information, external to the sample, is available for the subgroups. While the ordinary post-stratified estimator (or ratio-adjusted estimator) is a special case of regression estimator, raking ratio adjustment can be extended to loglinear models for weighting. One disadvantage is that no simple formula for its variance is available (Bethlehem and Keller, 1987). The regression estimator and raking ratio adjusted estimator, however, are both special cases of a more general class of estimators—the calibration estimator (Deville and Särndal , 1992). More importantly, any other member of the calibration estimator class is asymptotically equivalent to the regression estimator and, as a consequence, all members of the calibration estimator class share the same asymptotic variance (Deville and Särndal , 1992).

In this study, we first present the Horvitz-Thompson estimator in matrix form (section 1) in order to compare it with the regression estimator (section 2). In section 3 we discuss the unconditional variance of the regression estimator and compare it to the unconditional variance of the Horvitz-Thompson estimator. Our intention in discussing the regression estimator here is to throw some light on a more complicated estimator—the raking ratio adjusted estimator. The raking ratio adjusted estimator, although its variance formula is hard to find, shares the same asymptotic variance with the regression estimator (section 4). Since conditional variance estimates are preferred, we reviewed a recent study conducted by Yung and Rao (1996) (section 5). Raking ratio adjustment was performed on the

estimates of 1993 National Household Education Survey (NHES:93) School Readiness component (section 6). In section 7, we compare variance estimates which incorporated the raking ratio adjustment to variance estimates which did not incorporate the adjustment.

1. The Horvitz-Thompson Estimator

Let $\boldsymbol{Y} = (y_1, y_2, \ldots, y_N)'$ denote the $N \times 1$ vector of values of the target variable for all elements in the population *U*. A sample *s* of size *n* from the population can be represented by an *N´N*-diagonal matrix $\boldsymbol{T}(s)$, where $t_{ii} = 1$ if element *i* is in sample *s* and 0 otherwise. The inclusion probability matrix is denoted by $P = diag(p_i)_{N\times N}$, where $p_1, p_2, \ldots, p_N$ are the inclusion probabilities for all elements. Also let $\boldsymbol{1}_N$ be a *N´1* vector of all ones. Our objective is to estimate the population total of y defined by

$$Y = \sum_{i=1}^{N} y_i = \boldsymbol{1}_N' \boldsymbol{Y} .$$

To this end, the commonly used Horvitz-Thompson estimator of *Y* is

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{p_i} = \boldsymbol{1}_N' \Pi^{-1} \boldsymbol{T}(s)\, \boldsymbol{Y} = \boldsymbol{W}_{HT} \boldsymbol{Y} .$$

Here $\boldsymbol{W}_{HT} = \boldsymbol{1}_N' \Pi^{-1} \boldsymbol{T}(s)$ is the design weight variable for Horvitz-Thompson estimator. The variance of $\hat{Y}_{HT}$ is

$$V(\hat{Y}_{HT}) = \boldsymbol{Y}' \Delta \boldsymbol{Y} = \sum_{l=1}^{N} \sum_{k=1}^{N} (p_{kl} - p_k p_l) \frac{y_k}{p_k} \frac{y_l}{p_l} .$$

Here $\Delta = (\Delta_{kl})_{N\times N}$ with $\Delta_{kl} = (p_{kl} - p_k p_l)/p_k p_l$ and p_{kl} is the joint inclusion probability of element *k* and *l* selected in the sample. The corresponding variance estimator is

$$\hat{V}(\hat{Y}_{HT}) = \sum_s \sum_s \frac{(p_{kl} - p_k p_l)}{p_{kl}} \frac{y_k}{p_k} \frac{y_l}{p_l} .$$

2. The Regression Estimator

The Horvitz-Thompson estimator, although unbiased, is not efficient when relevant auxiliary variables are present. In practice, information external to the sample is often available in addition to the inclusion probabilities. This information can be used to increase precision and reduce bias. Let $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x}_N)'$ be the $N \times p$-matrix of values of the auxiliary variables for all elements. Here $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ is the $p \times 1$ vector of values of the p variates for element i. It is natural to chose a vector $\boldsymbol{B} = (b_1, b_2, \ldots, b_p)'$ to regress $\boldsymbol{Y}$ on $\boldsymbol{X}$ such that

$$\boldsymbol{E'E} = \sum_{i=1}^{N} E_i^2 = (\boldsymbol{Y} - \boldsymbol{XB})'(\boldsymbol{Y} - \boldsymbol{XB}) = \sum_{i=1}^{N} (y_i - \boldsymbol{x}_i'\boldsymbol{B})^2$$

is minimized; here $\boldsymbol{E} = (E_1, E_2, \ldots, E_N)'$. Without an assumption of any model, the ordinary least squares method results in

$$\boldsymbol{B} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y} = \left(\sum_{i=1}^{N} \boldsymbol{x_i x_i'}\right)^{-1}\left(\sum_{i=1}^{N} \boldsymbol{x_i} y_i\right) = \boldsymbol{U}^{-1}\boldsymbol{V}$$

with $\boldsymbol{U} = \sum_{i=1}^{N} \boldsymbol{x_i x_i'}$ and $\boldsymbol{V} = \sum_{i=1}^{N} \boldsymbol{x_i} y_i$. Since $\sum_{i=1}^{N} \boldsymbol{x_i x_i'} = (\sum_{i=1}^{N} x_{ik} x_{il})_{p \times p}$, apply Horvitz-Thompson estimator to estimate $\sum_{i=1}^{N} x_{ik} x_{il}$ for fixed k and l results in $\sum_{i=1}^{n} x_{ik} x_{il} / p_i$. Therefore, the Horvitz-Thompson estimator of $\boldsymbol{U} = \sum_{i=1}^{N} \boldsymbol{x_i x_i'}$ can be written as

$$\hat{\boldsymbol{U}} = \sum_{i=1}^{n} \left(\frac{x_{ik} x_{il}}{p_i}\right)_{p \times p} = \sum_{i=1}^{n} \boldsymbol{x_i} p_i^{-1} \boldsymbol{x_i'} = \boldsymbol{X'}\Pi^{-1}\boldsymbol{T}(s)\boldsymbol{X}.$$

Similarly, the Horvitz-Thompson estimator of $\boldsymbol{V} = \sum_{i=1}^{N} \boldsymbol{x_i} y_i$ can be written as

$$\hat{\boldsymbol{V}} = \sum_{i=1}^{n} \boldsymbol{x}_i p_i^{-1} y_i = \boldsymbol{X'}\Pi^{-1}\boldsymbol{T}(\mathrm{s})\boldsymbol{Y}.$$

A customarily used estimator of $\boldsymbol{B}$ is:

$$\hat{\boldsymbol{B}} = \hat{\boldsymbol{U}}^{-1}\hat{\boldsymbol{V}} = (\sum_{i=1}^{n} \boldsymbol{x_i} p_i^{-1} \boldsymbol{x_i'})^{-1} \sum_{i=1}^{n} \boldsymbol{x_i} p_i^{-1} y_i = \left(\boldsymbol{X'}\Pi^{-1}\boldsymbol{T}(s)\boldsymbol{X}\right)^{-1}\boldsymbol{X'}\Pi^{-1}\boldsymbol{T}(s)\boldsymbol{Y}.$$

$\hat{\boldsymbol{B}}$ is asymptotically design unbiased (see for example, Bethlehem & Keller, 1987). Based on $\hat{\boldsymbol{B}}$, the regression estimator of Y is defined as

$$\hat{Y}_R = \hat{Y}_{HT} + (\boldsymbol{t_X} - \hat{\boldsymbol{t}}_{\boldsymbol{X},HT})'\hat{\boldsymbol{B}}.$$

Here $\boldsymbol{t_X} = \boldsymbol{X}'\boldsymbol{1_N} = \sum_{i=1}^{N} \boldsymbol{x_i} = (\sum_{i=1}^{N} x_{i1},\ \sum_{i=1}^{N} x_{i2},\ \ldots,\ \sum_{i=1}^{N} x_{ip})'$ are the population totals of the auxiliary variables, and $\hat{\boldsymbol{t}}_{X,HT} = \boldsymbol{X}'\Pi^{-1}\boldsymbol{T}(s)\boldsymbol{1_N} = (\sum_s x_{i1}p_i^{-1}, \sum_s x_{i1}p_i^{-1},\ \ldots,\ \sum_s x_{ip}p_i^{-1})'$ is the Horvitz-Thompson estimator of the auxiliary variable totals based on the sample. $\hat{Y}_R$ is asymptotically design unbiased for Y (Bethlehem & Keller 1987). Also notice

$$\hat{Y}_R = [\boldsymbol{1}_N'\Pi^{-1}\boldsymbol{T}(s) + (\boldsymbol{t_X} - \hat{\boldsymbol{t}}_{X,HT})'\left(\boldsymbol{X}'\Pi^{-1}\boldsymbol{T}(s)\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\Pi^{-1}\boldsymbol{T}(s)]\,\boldsymbol{Y} = \boldsymbol{W}_R\,\boldsymbol{Y}\ .$$

Here $\boldsymbol{W}_R = \boldsymbol{1}_N'\Pi^{-1}\boldsymbol{T}(s) + (\boldsymbol{t_X} - \hat{\boldsymbol{t}}_{X,HT})'\left(\boldsymbol{X}'\Pi^{-1}\boldsymbol{T}(s)\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\Pi^{-1}\boldsymbol{T}(s)$ is the regression weight variable for the regression estimator. Another important property of $\boldsymbol{W}_R$ is that the regression estimates of auxiliary variables are always equal to the population total:

$$\hat{\boldsymbol{X}}_{\boldsymbol{R}} = \boldsymbol{W}_R\boldsymbol{X} = \boldsymbol{t}_X'\ ,$$

which is termed as calibration equation (Deville and Särndal , 1992). A potential problem is that some of the regression weights can be negative. Huang (1978) designed a computer program to produce nonnegative regression weights.

3. The Mean Square Error of Regression Estimator

We discuss two estimators of MSE($\hat{Y}_R$), the mean square error of $\hat{Y}_R$. The first estimator starts from an alternative expression for the regression estimator:

$$\hat{Y}_R = \hat{Y}_{HT} + (\boldsymbol{t}_X - \hat{\boldsymbol{t}}_{X,HT})'\hat{\boldsymbol{B}} = \sum_s p_i^{-1}y_i + \sum_s(\boldsymbol{t}_X - \hat{\boldsymbol{t}}_{X,HT})'(\sum_s \boldsymbol{x}_i p_i^{-1}\boldsymbol{x}_i')^{-1}\boldsymbol{x}_i p_i^{-1}y_i$$
$$= \sum_s\left[1 + (\boldsymbol{t}_X - \hat{\boldsymbol{t}}_{X,HT})'(\sum_s \boldsymbol{x}_i p_i^{-1}\boldsymbol{x}_i')^{-1}\boldsymbol{x}_i\right]p_i^{-1}y_i\ .$$

Let $g_{i,s} = 1 + (\boldsymbol{t}_X - \hat{\boldsymbol{t}}_{X,HT})'(\sum_s \boldsymbol{x}_i p_i^{-1}\boldsymbol{x}_i')^{-1}\boldsymbol{x}_i$ and notice by definition $y_i = E_i + \boldsymbol{x}_i'\boldsymbol{B}$. We have $\hat{Y}_R = \sum_s g_{i,s}y_i / p_i = \sum_s g_{i,s}\boldsymbol{x}_i'\boldsymbol{B}/p_i + \sum_s g_{i,s}E_i/p_i$. Here $\sum_{\boldsymbol{s}} g_{i,s}\boldsymbol{x}_i'\boldsymbol{B}/p_{\boldsymbol{i}}$ can be also written as

$$\sum_{\boldsymbol{s}} g_{i,s}\boldsymbol{x}_i'\boldsymbol{B}/p_{\boldsymbol{i}} = \left[\sum_s \boldsymbol{x}_i'p_i^{-1} + (\boldsymbol{t}_X - \hat{\boldsymbol{t}}_{X,HT})'(\sum_s \boldsymbol{x}_i p_i^{-1}\boldsymbol{x}_i')^{-1}\sum_s \boldsymbol{x}_i p_i^{-1}\boldsymbol{x}_i'\right]\boldsymbol{B}$$
$$= \left[\ \hat{\boldsymbol{t}}_{X,HT}' + (\boldsymbol{t}_X - \hat{\boldsymbol{t}}_{X,HT})'\ \right]\ \boldsymbol{B} = \boldsymbol{t}_X'\boldsymbol{B}\ ,$$

which is a constant. Therefore

$$V(\hat{Y}_R) = V(\sum_s g_{i,s} E_i / p_i) .$$

Since $g_{i,s}$ depends on the sample *s*, the variance estimator for the Horvitz-Thompson estimator can not be applied directly here. Disregard this and use $e_{i,s} = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{B}}$ to substitute $E_i = y_i - \boldsymbol{x}_i'\boldsymbol{B}$, Särndal (1982) proposed variance estimator

$$\hat{V}_1(\hat{Y}_R) = \sum_{k\in s}\sum_{l\in s}(\Delta_{kl} / p_{kl})(g_{k,s}e_{k,s} / p_k)(g_{l,s}e_{l,s} / p_l) .$$

We shall see in section 5 that $\hat{V}_1(\hat{Y}_R)$ might perform better as a conditional variance estimator of $\hat{Y}_R$.

The second estimator of MSE($\hat{Y}_R$) starts from the Taylor linearizatioin substitution of $\hat{Y}_R$ (Särndal , Swensson and Wretman, 1992):

$$\hat{Y}_R \cong \hat{Y}_{LR} = \hat{Y}_{HT} + (\boldsymbol{t}_X - \hat{\boldsymbol{t}}_{X,HT})'\boldsymbol{B} = \sum_s p_i^{-1} y_i + \boldsymbol{t}_X'\boldsymbol{B} - \sum_s p_i^{-1}\boldsymbol{x}_i'\boldsymbol{B}$$

$$= \boldsymbol{t}_X'\boldsymbol{B} + \sum_s p_i^{-1}(y_i - \boldsymbol{x}_i'\boldsymbol{B}) = \boldsymbol{t}_X'\boldsymbol{B} + \sum_s p_i^{-1} E_i .$$

Here $\hat{Y}_{LR}$ is the linearized regression estimator of Y. Since $\boldsymbol{B}$ and $\boldsymbol{t}_X'$ are population parameters and $\hat{Y}_{LR}$ is unbiased for population mean Y —that is, $E(\hat{Y}_{LR}) = Y$ —therefore

$$MSE(\hat{Y}_R) \cong V(\hat{Y}_{LR}) = \sum_{k=1}^N \sum_{l=1}^N (p_{kl} - p_k p_l)\frac{E_k}{p_k}\frac{E_l}{p_l}$$

which is thereafter estimated by

$$\hat{V}_2(\hat{Y}_R) = \sum_s \sum_s \frac{(p_{kl} - p_k p_l)}{p_{kl}}\frac{e_k}{p_k}\frac{e_l}{p_l}$$

with $e_{i,s} = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{B}}$.

$V(\hat{Y}_{LR})$ provides a heuristic explanation of why the regression estimator has smaller unconditional variance (over all possible samples) compared to Horvitz-Thompson. If $\boldsymbol{x}_i'\boldsymbol{B}$ is a perfect substitute of y_i, that is $y_i = \boldsymbol{x}_i'\boldsymbol{B}$, then $E_i = 0$ and therefore, $MSE(\hat{Y}_R) \cong V(\hat{Y}_{LR}) = 0$. If $\boldsymbol{x}_i$ is not related to y_i at all, then $\boldsymbol{B} \cong \boldsymbol{0}$ and $E_i \cong y_i$. Then

$$MSE(\hat{Y}_R) \cong V(\hat{Y}_{LR}) = \sum_{k=1}^N \sum_{l=1}^N (p_{kl} - p_k p_l)\frac{y_k}{p_k}\frac{y_l}{p_l} = V(\hat{Y}_{HT})$$

which indicates $MSE(\hat{Y}_R) \cong V(\hat{Y}_{HT})$. When $\boldsymbol{x_i}$ is partially related to y_i, E_i has smaller variation than y_i.

4. The Role of Regression Estimator

The auxiliary variables used in the regression estimator can be both quantitative variables and qualitative variables. Actually, the poststratified estimator is a special case of the regression estimator when the auxiliary variables are the indicator variables for the poststrata. Suppose the population is partitioned into C post-strata with known population counts M_c, $c = 1, ..., C$. Let $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iC})'$ be the post-strata indicator vector so that $x_{ic} = 1$ if element i belongs to that post-stratum c and 0 otherwise. The Horvitz-Thompson estimator of M_c is given by

$$\hat{M}_{c,HT} = \sum_{i \in s} x_{ic} p_i^{-1} = \sum_{i \in s_c} p_i^{-1}$$

where s_c is the sample of elements belonging to the c-th post-stratum. And the Horvitz-Thompson estimator of the post-stratum total Y_c is given by

$$\hat{Y}_{c,HT} = \sum_{i \in s} x_{ic} p_i^{-1} y_i = \sum_{i \in s_c} p_i^{-1} y_i \ .$$

The post-stratified estimator is therefore defined as

$$\hat{Y}_{ps} = \sum_c \frac{M_c}{\hat{M}_{c,HT}} \hat{Y}_{c,HT} \ .$$

Notice $\boldsymbol{t}_X = \sum_{i=1}^{N} \boldsymbol{x}_i = (M_1, M_2, \ldots, M_C)'$, $\hat{\boldsymbol{t}}_{X,HT} = (\hat{M}_{1,HT}, \hat{M}_{2,HT}, \ldots, \hat{M}_{C,HT})'$, and

$$\hat{\boldsymbol{B}} = diag(\sum_{i \in s_c} p_i^{-1})_{C \times C}^{-1} (\sum_{i \in s_1} p_i^{-1} y_i, \sum_{i \in s_2} p_i^{-1} y_i, \ldots, \sum_{i \in s_C} p_i^{-1} y_i)'$$

$$= (\hat{R}_1, \hat{R}_2, \ldots, \hat{R}_C)'$$

where $\hat{R}_c = \hat{Y}_{c,HT} \big/ \hat{M}_{c,HT}$. Therefore, the regression estimator reduces to

$$\hat{Y}_R = \hat{Y}_{HT} + \sum_{c=1}^{C} (M_c - \hat{M}_{c,HT}) \hat{R}_c = \hat{Y}_{ps} .$$

The ratio adjusted post-stratified estimator $\hat{Y}_{ps}$ discussed above requires population counts at cell level. However, these cell counts are not always available, especially when several auxiliary

variables are used. For instance, age group counts are available from one file and region group counts are available from another file. Here the population marginal counts are known, but the cross-classification is lacking and, therefore, it is described as incomplete poststratification.

Two techniques are often applied to handle incomplete poststratification. The first approach uses regression estimator by introducing multiple poststrata indicator variables (Bethlehem and Keller, 1987). The second approach uses raking ratio adjustment (Deming and Stephan, 1940). Raking estimation uses iterative proportional fitting and can be extended to loglinear models for weighting. One disadvantage is that no simple formula for its variance is available (Bethlehem and Keller, 1987).

The importance of the regression estimator was revealed by Deville and Särndal (1992). Deville and Särndal introduced the calibration estimator, which includes often used estimators such as the ratio estimator, the regression estimator, and the raking ratio estimator as special cases. They proved that any other member of the calibration estimator class is asymptotically equivalent to the regression estimator and, as a consequence, all members of the calibration estimator class share the same asymptotic variance. Hence the variance estimators for the regression estimator discussed in section 3 and the conditional variance estimator in the next section can be used to estimate the variance of any estimator in the calibration class.

5. Estimation of Conditional Variance of Regression Estimator

In section 3 we considered the unconditional variance of the regression estimator which is calculated over all possible samples under the complex survey design. The unconditional variance can be used when comparing sampling strategies before the sample is drawn. There is a growing belief, however, that inference should be made conditional on the known attributes of the sample. Holt and Smith (1979) gave compelling arguments in favor of conditional inference for the poststratification of a simple random sample. Rao (1985) emphasized the need for conditioning the inference on recognizable subsets of the population by using a number of real examples involving random sample sizes. Valliant (1993) studied the standard linearization variance estimator, BRR, and the jackknife variance estimator

to determine whether they estimate the conditional variance of the poststratified estimator of a finite population total under a super-population model. Yung and Rao (1996) studied the standard linearization variance estimator, jackknife, and the jackknife linearization variance estimators for both the poststratified estimator and the regression estimator.

Following Yung and Rao (1996), under a stratified multistage design with large numbers of strata, L, and relatively few primary sampling units (clusters), n_h (≥ 2), sampled within each stratum, the clusters are treated as if they are selected with replacement to simplify the variance estimation. The standard linearization variance estimator for the ratio adjusted post-stratification estimator $\hat{Y}_{ps}$ is

$$\hat{V}_L(\hat{Y}_{ps}) = \sum_{h=1}^{L} \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} (\tilde{e}_{hi,s} - \bar{\tilde{e}}_{h,s})^2 .$$

Here $\tilde{e}_{hi,s} = \sum_c \sum_{k \in s_c} n_h p_{hik}^{-1} (y_{hik} - \hat{Y}_{c,HT} / \hat{M}_{c,HT})$ and $\bar{\tilde{e}}_{h,s} = \sum_{i=1}^{n_h} \tilde{e}_{hi,s} / n_h$. The jackknife variance estimator of $\hat{Y}_{ps}$ is defined as

$$\hat{V}_J(\hat{Y}_{ps}) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{ps(gj)} - \hat{Y}_{ps})^2 .$$

Here $\hat{Y}_{ps(gj)}$ is obtained from the sample after omitting the data from the j-th sampled cluster in the g-th stratum (j = 1, ..., n_g; g = 1, ..., L) and the reweighting is done each time a cluster is deleted. By linearalizing the jackknife variance estimator $\hat{V}_J(\hat{Y}_{ps})$, the jackknife linearization variance estimator of $\hat{Y}_{ps}$ is then obtained as

$$\hat{V}_{JL}(\hat{Y}_{ps}) = \sum_{h=1}^{L} \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} (\tilde{e}^*_{hi,s} - \bar{\tilde{e}}^*_{h,s})^2 .$$

Here $\tilde{e}^*_{hi,s} = \sum_c \sum_{k \in s_c} n_h p_{hik}^{-1} (M_c / \hat{M}_{c,HT})(y_{hik} - \hat{Y}_{c,HT} / \hat{M}_{c,HT})$ and $\bar{\tilde{e}}^*_{h,s} = \sum_{i=1}^{n_h} \tilde{e}^*_{hi,s} / n_h$.

$\hat{V}_{JL}(\hat{Y}_{ps})$ and $\hat{V}_J(\hat{Y}_{ps})$ are asymptotically equal to higher order terms in the special case of n_h=2 (Yung and Rao, 1996). $\hat{V}_{JL}(\hat{Y}_{ps})$ also reduces to a conditionally valid variance estimator for

simple random sampling given the poststratum sample sizes while $\hat{V}_L(\hat{Y}_{ps})$ does not (Rao, 1985). Therefore, $\hat{V}_{JL}(\hat{Y}_{ps})$ might perform better as a conditional variance estimator of $\hat{Y}_{ps}$.

When quantitative auxiliary variables are used in the regression estimator, the meaning of the conditional variance is not clear. But still $\hat{V}_L(\hat{Y}_R)$, $\hat{V}_J(\hat{Y}_R)$, and $\hat{V}_{JL}(\hat{Y}_R)$ have similar forms as $\hat{V}_L(\hat{Y}_{ps})$, $\hat{V}_J(\hat{Y}_{ps})$, and $\hat{V}_{JL}(\hat{Y}_{ps})$, except now

$$\tilde{e}_{hi,s} = \sum_k n_h p_{hik}^{-1}(y_{hik} - \boldsymbol{x}'_{hik}\hat{\boldsymbol{B}})$$

$$\tilde{e}^*_{hi,s} = \sum_k n_h p_{hik}^{-1}[1+(\boldsymbol{t}_{\boldsymbol{X}} - \hat{\boldsymbol{t}}_{X,HT})'(\sum_{(hik)\in s}\boldsymbol{x}_{hik}p_{hik}^{-1}\boldsymbol{x}'_{hik})^{-1}\boldsymbol{x}_{hik}](y_{hik} - \boldsymbol{x}'_{hik}\hat{\boldsymbol{B}}).$$

Yung and Rao's (1996) simulation study suggests that the three variance estimators, $\hat{V}_L(\hat{Y}_R)$, $\hat{V}_J(\hat{Y}_R)$, and $\hat{V}_{JL}(\hat{Y}_R)$ perform similarly under well balanced samples, while an incorrect jackknife procedure which does not recalculate the regression weights each time a cluster is deleted perform poorly.

When the sample size is not very large and the number of auxiliary variables is not small, Fuller et al. (1994) used

$$\hat{V}_L(\hat{Y}_R) = \frac{n}{n-p}\sum_{h=1}^{L}\frac{1}{n_h(n_h-1)}\sum_{i=1}^{n_h}(\tilde{e}_{hi,s} - \bar{\tilde{e}}_{h,s})^2$$

to compensate for the lost degrees of freedom due to estimating the regression coefficients. It is also interesting to notice that

$$\hat{V}_{JL}(\hat{Y}_R) = \sum_{h=1}^{L}\frac{1}{n_h(n_h-1)}\sum_{i=1}^{n_h}(\tilde{e}^*_{hi,s} - \bar{\tilde{e}}^*_{h,s})^2$$

is actually estimating

$$\sum_{h=1}^{L}V(\bar{\tilde{e}}^*_{h,s}) = V\left(\sum_{h=1}^{L}\bar{\tilde{e}}^*_{h,s}\right) = V\left(\sum_{h=1}^{L}\frac{1}{n_h}\sum_{i=1}^{n_h}\tilde{e}^*_{hi,s}\right)$$

$$= V(\sum_{(hik)\in s}p_{hik}^{-1}[1+(\boldsymbol{t}_X - \hat{\boldsymbol{t}}_{X,HT})'(\sum_{(hik)\in s}\boldsymbol{x}_{hik}p_{hik}^{-1}\boldsymbol{x}'_{hik})^{-1}\boldsymbol{x}_{hik}](y_{hik} - \boldsymbol{x}'_{hik}\hat{\boldsymbol{B}})$$

$$= V(\sum_{(hik)\in s}p_{hik}^{-1}g_{hik,s}e_{hik,s}).$$

Disregarding the fact that $g_{hik,s}$ and $e_{hik,s}$ depend on sample *s*, we can reproduce $\hat{V}_1(\hat{Y}_R)$ of section 3 by estimating $V(\sum_{(hik)\in s}p_{hik}^{-1}g_{hik,s}e_{hik,s})$:

$$\sum_{(hik)\in s}\sum_{(h'i'k')\in s}(\Delta_{hik,h'i'k'}/p_{hik,h'i'k'})(g_{hik,s}e_{hik,s}/p_{hik})(g_{h'i'k',s}e_{h'i'k',s}/p_{h'i'k'}).$$

6. An Overview of NHES Sample Design and Weighting Procedure

We choose the National Household Education Survey (NHES:93) School Readiness (SR) component data for this study since both ratio adjustment and raking adjustment were performed in the weighting procedure. The jackknife variance estimation replicate weights were provided. In addition, the strata identification variable and the PSU identification variable are also included in the data file so that linearization method can be applied to calculate the variance. A clear description of the survey was given by Brick et al. (1994) and is paraphrased here.

The target population of the NHES:93 survey was children aged 3 through 7, or in second grade or below but at least age 3. The method of sampling used in NHES:93 is a variant of the random digit dialing method, which can be viewed as stratified multistage sampling.

The sampling procedure starts with stratifying a list of PSUs (a list of all possible first 8 digits of 10-digit phone numbers) into low and high minority concentration strata. A random selection of PSUs was then made with an unequal sampling rate from each stratum. With each selected PSU, telephone numbers were generated by adding random two-digit numbers to the eight-digit PSU number. A sample of 129,813 telephone numbers was generated from 4,577 PSUs. Because of nonresidence and nonresponse, 63,844 households actually completed screening.

Based on data from the 63,844 Screener interviews, every household with children in the eligible age and grade ranges was sampled. Within each sampled household, if there were one or two eligible children in a household, each was selected with certainty. About 96.4 percent of households with any eligible children met this condition. If there were more than two eligible children in the household, two were randomly sampled from the household. The number of completed School Readiness (SR) interviews was 10,888.

The first step of the weighting procedure was to create a household weight which accounted for the unequal PSU sampling rates, because some households had more than one telephone number and hence had more than one chance of being included in the sample. Then the household weights were adjusted for those children who were not chosen with certainty. This adjusted base weight was the inverse of inclusion probability for the children in the SR component.

Then the weights were adjusted for nonresponse to the extended interview. Six age categories from 3 to 8 and older were used to define the nonresponse adjustment cells. The nonresponse adjustment was the sum of the adjusted base weights for all sampled children in the cell divided by the sum of the adjusted base weights for the respondents in the same cell. The adjustment factors varied from 1.09 to 1.14 across the six cells.

The last stage of weighting was to rake the nonresponse-adjusted person weights to known totals computed from the October 1992 Current Population Survey (CPS). The marginal totals are given in table 1 from Brick et al. (1994). Three dimensions were used in the raking. The first dimension is defined by the cross-classification of home type (owned or not) and Census region. The second dimension is the cross of race/ethnicity and household income. The last dimension is defined by age and grade.

In order to help users to estimate standard errors, 60 jackknife replicate weights were created based on the sampling of clusters of telephone numbers. All 60 replicate weights were created using the same estimation procedures used for the full sample. Also included in the data file are stratum and PSU variables required by software using Taylor series approximation.

Table 1. NHES:93 control totals for School Readiness raking

Control characteristics		Control totals
Home type	**Census region**	
Owned or other	Northeast	2,400,545
Owned or other	Midwest	3,202,557
Owned or other	South	4,116,866
Owned or other	West	2,589,938
Rented	Northeast	1,448,553
Rented	Midwest	1,651,182
Rented	South	2,764,945
Rented	West	1,938,053
Race/ethnicity	**Household income**	
Hispanic	Less than $10,000	818,994
Hispanic	$10,000-$24,999	904,880
Hispanic	$25,000 or more	685,193
Black, non-Hispanic	Less than $10,000	1,360,091
Black, non-Hispanic	$10,000-$24,999	997,013
Black, non-Hispanic	$25,000 or more	792,487
Other	Less than $10,000	1,514,364
Other	$10,000-$24,999	3,610,969
Other	$25,000 or more	9,428,649
Age	**Grade**	
3	All grades	3,905,387
4	All grades	3,806,845
5	All grades	3,832,330
6	All grades	3,763,999
7	All grades	3,809,885
8 and older	Second grade or less	994,193

NOTE: Details do not add to the same total due to rounding.
SOURCE: U.S. Bureau of the Census, *Current Population Survey*, October 1992.

7. Variance Estimates Comparison

Rust (1987) investigated the effect of nonresponse and ratio weight adjustments on sampling error estimates by using the Title IV Quality Control Study survey data for two continuous variables. In his study, the differences between the variances estimated via the two approaches are small, which indicates the relationship between the variable of interest and the auxiliary variable was not a strong one. He also noticed, in another study undertaken by Lago et al. (1987), that when variables of interest (weight, height, and level of cholesterol) are highly correlated with the poststratification variables (age and sex), the use of poststratification gave rise to considerable reduction in sampling variance.

In this section, we compare variance estimates which incorporate the raking ratio adjustments and nonresponse adjustment with the variance estimates which ignore these adjustments for the 1993 NHES School Readiness component.

We first used the jackknife replicate weights which incorporated the adjustments to calculate standard errors for two kinds of estimators—total and mean estimators. The replicate weights were created by Westat, Inc., and were provided with the public use data set. The calculation is implemented by WesVar PC; the standard errors calculated by this approach are denoted as ste_T for total estimator, and ste_R for ratio type estimator (this includes estimators of percentage, mean, and the ratio of two variables).

Then we calculated the standard errors for the same estimators but ignored the adjustments. This was implemented in two ways. The first approach was to let WesVar PC generate the jackknife replicate weights and then use these replicate weights to calculate the standard errors with WesVar PC. In this approach, neither nonresponse adjustment nor raking ratio adjustment are performed when a replicate weight is created; therefore these adjustments were not incorporated. The second way was to use the stratum identification variable and PSU identification variable provided with the public use data file to calculate the standard errors with SUDAAN. This approach actually treats the adjusted full sample final weight (FWGT0—Final Raked Weight which incorporates the nonresponse adjustment

and the raking ratio adjustment) as a design weight (inverse of inclusion probability). And the variance estimator of the Horvitz-Thompson estimator was used. Also notice that the mean estimator in this study is actually a ratio of two raking ratio adjusted estimators. Although SUDAAN is used here, the underlying variance estimator is actually the variance estimator for the ratio of two Horvitz-Thompson estimators, not a genuine linearized variance estimator for the ratio of two raking ratio adjusted estimators. Therefore the adjustments were also ignored in this approach. The variance estimates calculated from these two approaches (from WesVar PC generated replicate weights and from SUDAAN) are identical. They are denoted by ste_T^* for the standard error of the total estimator and ste_R^* for the standard error for the ratio type estimator.

Table 2 shows standard errors for categorical variables. As we can see, in general, ste_T is much smaller than ste_T^* while ste_R is close to ste_R^* except for the last two variables (which were used as auxiliary variables in the raking ratio adjustment). It seems like the adjustments and the gain in precision cancel out for the ratio type estimator.

For the standard error of the total estimate for dichotomous variables (Hastory, Hncare, Birthord, Hlive, Gender), when the adjustments are incorporated in the calculation, the marginal total counts are a constant C = 20,112,639. So the estimated total number children in category one equals C minus the estimated total number of children in category two for each replicate weight. Therefore the estimated standard errors for both categories are the same. When the adjustments are ignored, however, the estimated marginal total varies from one replicate weight to another. The relationship does not hold anymore. This explains why we observe unstable estimates for the standard errors of total estimates. Hncare, for example, has standard errors 92,717 and 370,645 for “Yes” and “No” categories.

For the standard errors of the percentage and mean estimators, when the adjustments are incorporated, the denominator again becomes the constant C for all replicates. Therefore, the standard error equals ste_T / C. When the adjustments are ignored, the denominator varies. But since the

numerator is positively correlated with the denominator, the actual standard error is smaller than ste_T^*/C.

Hincmrng (household income) is one of the auxiliary variables used for the raking ratio adjustment (table 1) where it has three categories ("Less than $10,000", "$10,000-$24,999", "$25,000 or more"). In the public use data file, two categories, "Less than $10,000" and "$10,000-$24,999", were collapsed into one category, "Up To $25,000". The marginal totals for all replicates are still the same. Therefore the standard errors are null.

Raceethn (race/ethnicity) was also used for the raking ratio adjustment where it was collapsed into three categories ("Hispanic", "Black, non-Hispanic", "Other") but in the public data file it has the customary four categories ("White/Nonhisp", "Black/Nonhisp", "Hispanic", "All O/Races"). Now the marginal totals for category "White/Nonhisp" and "All O/Races" are not constant anymore, so we observe standard errors for these two categories but no standard error for the other two.

Table 3 shows standard errors for continuous variables. The gain in precision to the total estimator is obvious. Age92 (Age) is an auxiliary variable used for raking ratio adjustment but was treated as a continuous variable here. Ratio Hbedrms/Hhtotal (Number of Bedrooms in Home/Total Number of Household Members) and Hhundr18/Hhtotal (Number of Household Members Under 18/Total Number of Household Members) are ratios of two raking ratio adjusted estimators. Incorporating the adjustment results in standard error estimates of about 14 and 7 percent less.

Table 4 shows standard errors calculated within the nonresponse adjustment and raking ratio adjustment cells (Home type × Census region × Race/ethnicity × Household income × Age × Grade). Only two cells with comparatively large sample sizes were chosen. Within these cells, the adjustments are the same for all units, so the adjustment factors were canceled out for the ratio type estimator and hence ste_R is about the same as ste_R^*. But still, a gain in precision due to the raking ratio adjustment to the total estimator is present.

Table 2. Standard errors for categorical variables

Categorical Variables	ste_T	ste_T^*	ste_T / ste_T^*	ste_R	ste_R^*	ste_R / ste_R^*
Hastory						
Yes ……	79375	217683	0.3646	0.395	0.507	0.7791
No ……	79374	230654	0.3441	0.395	0.507	0.7791
Hncare						
Yes ……	81658	92717	0.8807	0.406	0.413	0.9831
No ……	81658	370645	0.2203	0.406	0.413	0.9831
Birthord						
Only/Oldest Kid …….	109700	200995	0.5458	0.545	0.535	1.0187
Later Born ……..	109700	255680	0.4291	0.545	0.535	1.0187
Hlive						
Yes ……	152523	257797	0.5916	0.758	0.788	0.9619
No ……	152523	252258	0.6046	0.758	0.788	0.9619
Gender						
Female ……..	104303	222735	0.4683	0.519	0.524	0.9905
Male ……..	104303	231969	0.4496	0.519	0.524	0.9905
Habooks						
None ……..	23347	25110	0.9298	0.116	0.124	0.9355
1 Or 2 Books ……..	35046	38619	0.9075	0.174	0.191	0.9110
3 To 9 Books ……..	73626	90597	0.8127	0.366	0.422	0.8673
10 To 25 Books ……	94273	134211	0.7024	0.469	0.465	1.0086
26 To 50 Books …….	91039	126309	0.7208	0.453	0.469	0.9659
More Than 50 ……	124337	222669	0.5584	0.618	0.667	0.9265
Hincome						
$5,000 Or Less …….	58528	94562	0.6189	0.291	0.416	0.6995
$5,001 - $10,000 ……..	58528	101152	0.5786	0.291	0.434	0.6705
$10,001 - $15,000 ……	58980	79911	0.7381	0.293	0.383	0.7650
$15,001 - $20,000 ……	77404	98786	0.7835	0.385	0.456	0.8443
$20,001 - $25,000 ……	75325	99576	0.7565	0.375	0.455	0.8242
$25,001 - $30,000 ……	69972	80165	0.8729	0.348	0.379	0.9182
$30,001 - $35,000 ……	53173	63908	0.8320	0.264	0.295	0.8949
$35,001 - $40,000 ……	61437	70068	0.8768	0.305	0.319	0.9561
$40,001 - $50,000 ……	81543	96797	0.8424	0.405	0.422	0.9597
$50,001 - $75,000 ……	65695	89348	0.7353	0.327	0.375	0.8720
Over $75,000 ……	76787	87698	0.8756	0.382	0.407	0.9386
Hincmrng						
Up To $25,000 ……..	2	255420	0.0000	0	0.804	0.0000
More Than $25,000 ……	0	260352	0.0000	0	0.804	0.0000
Raceethn						
White/Nonhisp ……..	52425	319287	0.1642	0.261	0.802	0.3254
Black/Nonhisp ……	1	123945	0.0000	0	0.518	0.0000
Hispanic ……..	0	110665	0.0000	0	0.522	0.0000
All O/Races ……	52425	59301	0.8840	0.261	0.273	0.9560

Table 3. Standard errors for continuous variables

Continuous Variables	ste_T	ste_T^*	ste_T / ste_T^*	ste_R	ste_R^*	ste_R / ste_R^*
Hbedrms ……..	231137	1292940	0.1788	0.011	0.014	0.803
Hhtotal …….	415720	1953781	0.2128	0.021	0.021	1.024
Hhundr18 ……	369884	1161715	0.3184	0.018	0.019	0.952
Numsibs ……	351823	747261	0.4708	0.017	0.018	0.944
Tv8to3 ……..	249661	426974	0.5847	0.012	0.014	0.889
Tvafdin …….	250867	493058	0.5088	0.012	0.012	0.984
Tvsat …….	520567	1516009	0.3434	0.026	0.027	0.974
Tvsun ……	500809	1201840	0.4167	0.025	0.025	0.988
Age92 ……	8698	2125447	0.0041	0.000	0.015	0.000
Hbedrms/Hhtotal …				0.003022	0.003515	0.8597
Hhundr18/Hhtotal ..				0.001987	0.002138	0.9294

Table 4. Standard errors calculated within the nonresponse adjustment and raking ratio adjustment cells

		ste_T	ste_T^*	ste_T / ste_T^*	ste_R	ste_R^*	ste_R / ste_R^*
CELL	BIRTHORD						
1	Only/Oldest Kid ……	24014.32	26749.49	0.8977	4.637	4.599	1.0083
1	Later Born ……..	22812.91	24063.16	0.9480	4.637	4.599	1.0083
2	Only/Oldest Kid ……	18091.32	22370.59	0.8087	2.594	2.617	0.9912
2	Later Born ……..	21688.37	24680.15	0.8788	2.594	2.617	0.9912
CELL	HASTORY						
1	Yes ……	5826.182	6005.819	0.9701	1.408	1.421	0.9909
1	No ……..	27764.57	33962.51	0.8175	1.408	1.421	0.9909
2	Yes ……	26773.59	36412.67	0.7353	0.866	0.869	0.9965
2	No ……...	5006.48	4970.37	1.0073	0.866	0.869	0.9965
CELL	HLIVE						
1	Yes ……	20932.59	23983.75	0.8728	4.007	4.003	1.0010
1	No ……..	22133.66	24012.56	0.9218	4.007	4.003	1.0010
2	Yes ……	19193.92	22665.75	0.8468	2.503	2.523	0.9921
2	No ……..	20255.97	23877.39	0.8483	2.503	2.523	0.9921
CELL	HINCMRNG						
1	Up To $25,000 ……...	15329.57	16214.36	0.9454	3.46	3.49	0.9914
1	More Than $25,000 ..	26211.87	30703.29	0.8537	3.46	3.49	0.9914
2	Up To $25,000 ……...	18989.11	20553.96	0.9239	2.924	2.844	1.0281
2	More Than $25,000 ..	24678.83	28751.11	0.8584	2.924	2.844	1.0281
CELL	STATISTIC ……...						
1	HHUNDR18 ……..	82555.12	92617.02	0.8914	0.104	0.103	1.0097
2	HHUNDR18 ……..	70281.32	90818.96	0.7739	0.053	0.053	1.0000
1	TVAFDIN ……..	34909.93	40125.09	0.8700	0.056	0.056	1.0000
2	TVAFDIN ……..	41062.69	46779.49	0.8778	0.046	0.046	1.0000

References

Bethlehem, J.G. and Keller, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

Brick, J.M., Collins, M.A., Nolin, M.J., Ha, P. C., Levinsohn, M. and Chandler, K. (1994). *School Readiness Data File User's Manual*. NCES 94-193. U.S. Department of Education, Office of Educational Research and Improvement. Washington, DC: National Center for Education Statistics.

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

Deville, J-C. and Särndal ,C-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Fuller, W.A., Loughin, M.M. and Baker, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey. *Survey Methodology*, 20, 75-85.

Holt, D. and Smith, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society*, Ser. A, 142, 33-46.

Huang, E.T. (1978). *Nonnegative regression estimation for sample survey data.* Unpublished Ph. D. thesis. Ames, Iowa: Iowa State University.

Lago, J., Massey, J. T., Ezzati, T., Johnson, C. and Fulwood, R. (1987). Evaluation of the design effects for the Hispanic Health and Nutrition Examination Survey. *Proceedings of the Section on Survey Research Methods*. Washington, DC: American Statistical Association.

Rao, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology*, 11, 15-31.

Rust, K. (1987). *Practical problems in sampling error estimation*. Invited paper 10.3 of the Proceedings of the 46th session of the International Statistical Institute, Tokyo, Japan.

Särndal ,C-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*,7, 155-170.

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

Valliant, R., (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*. 88, 89-96.

Yung, W. and Rao, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, 22, 23-31.

BRR Variance Estimation Using VPLX Hadamard Procedure

Stanley Weng

1. Study Purpose

This study attempts to provide information on the use and performance of VPLX's balanced repeated replicates (BRR) capability, the *Hadamard* procedure, by comparing it with variance estimation procedures using existing BRR replicates and those using a jackknife procedure.

Until now, variance estimation for NCES complex surveys using the BRR method has usually been performed when a set of BRR replicates has been created and included in the survey sample datafile. The application of BRR variance estimating has been limited because the creation of BRR replicates requires advanced statistical knowledge. However, when the replicates are created, calculating BRR variance estimates is a simple matter which can be performed using any statistical software.

VPLX (Fay, 1995) and WesVar (Westat, 1996) are two widely used statistical software packages which can create BRR replicates and then perform BRR estimation. However, these capabilities have not been in extensive use, perhaps due their limitations (e.g., WesVar cannot handle large numbers of strata) or lack of instruction (e.g., VPLX has not documented its BRR capability). We chose VPLX, not WesVar, for this study because VPLX's Hadamard procedure has a more general design and greater capabilities.

2. VPLX Hadamard Procedure

Documentation for the VPLX Hadamard procedure was not available when this study was conducted. The author provided an example for the Hadamard command (Fay, 1996). Since it was

made for a very small sample, it did not have complete syntax information, but we were able to figure out the syntax for a large dataset.

3. VPLX Capability of Creating BRR Replicates: Grouped BRR Method

Originally, the BRR method applied to stratified multistage surveys for which each stratum contains two PSUs. The VPLX Hadamard procedure also applies only to such types of survey data. For handling sample with more than two PSUs in a stratum, the usual way is to randomly group the PSUs in each stratum into two groups—pseudo-PSUs—and then apply the BRR procedure to the pseudo-PSUs. This is the so-called *grouped BRR* (GBRR) or *grouped balanced half-sample* (GBHS) procedure. We wrote a SAS macro to perform the random grouping of PSUs within stratum.

Our study used the 1990 SASS Teacher Survey Public School sample. It was used in an earlier study (Weng, Zhang, & Cohen, 1995) which had found the jackknife variance estimates reliable. The 1990 SASS Teacher Survey Public School sample has about 250 strata. We collapsed some small strata according to the stratification structure, making the total number of strata below 240, and a Hadamard matrix of dimension 240 was used.

4. Analysis and Results

The following table lists the standard errors estimated by BRR using VPLX Hadamard procedure and using the existing BRR replicates in the data file. A column of jackknife (JK) estimates is added for reference. The same variables as used in the Weng et al. (1995) study were used in this study.

Table 1. Standard errors by BRR variance estimation

Survey statistics	Variable	Estimate	Standard error		
			BRR		JK
			VPLX Hadamard	Existing replicates	
Percent	Master degree				
	1. YES	46.980	.3499	.326	.393
	2: NO	53.020	.3499	.326	.393
	Look forward to day				
	1: ST AGREE	51.37	.4537	.341	.385
	2: AGREE	40.39	.4366	.313	.363
	3: DISAGREE	6.23	.1435	.163	.180
	4: ST DISAGREE	2.01	.1022	.121	.107
Mean	Salary	30,751	115.32	93.494	102.849
	Age	42.576	.0811	.0751	.0732
Ratio	School hours extra/hours required	0.0886	.0010	.001	.001
	Other hours extra/hours required	0.223	.0011	.0013	.0014

5. Discussion and Future Steps

It was generally expected that the BRR procedure performed in this study would deliver better accuracy for the BRR variance estimates than using the existing BRR replicators, because a larger number of replicates were used. However, the results, as listed in table 1, do not show clear evidence of such improvement (if the jackknife variance estimates used as a reference are considered reliable). Of course, one application of the grouped BRR procedure might not reveal sufficient information on its behavior. Further investigation may be needed. Methodologically, the grouped BRR produces an inconsistent estimator. However, as described below, improvements can be made by repeating the procedure.

Rao and Shao (1996) explored the *repeatedly grouped balanced half-sample* (RGBHS) method as an improvement to the grouped balanced half-sample (GBHS) method. In GBHS, the sample in each stratum is first randomly divided into two groups, and then the balanced half-sample method is applied to the groups. A repeatedly grouped balanced half-sample method involves

independently repeating the random grouping T times and then taking the average of the resulting T GBHS variance estimators, say, $v_G^t(\hat{q})$, **t** = 1, 2, ..., T:

$$v_{RG}(\hat{q}) = \frac{1}{T}\sum_{t=1}^{T} v_G^t(\hat{q}),$$

where $v_{RG}(\hat{q})$ denotes the RGBHS variance estimator.

The RGBHS variance estimator retains the simplicity of the GBHS variance estimator, since the same Hadamard matrix is applied to the random groups generated at each repetition. Rao and Shao (1996) established the asymptotic consistency of the RGBHS estimator, that is,

$$v_{RG}(\hat{q}) / V_a(\hat{q}) \rightarrow_p 1$$

where $V_a(\hat{q})$ is the asymptotic variance of $\hat{q}$. Their simulation study indicated that the RGBHS performs well for T as small as 15, thus providing flexibility in terms of the number of half-samples used. Intuitively, it is understandable since the RGBHS estimator is based on RT half-samples, instead of R half-samples as in GBHS.

Computationally, the RGBHS method is easy to implement.

References

Fay, R. E. (1995). *VPLX*. Washington DC: U.S. Bureau of the Census.

Fay, R. E. (1996). An example with half-sample replication. Personal correspondence.

Rao, J. N. K. and Shao, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association*, 91, 343-348.

Weng, S. S., Zhang, F. and Cohen, M.P. (1995). Variance estimates comparison by statistical software, *ASA 1995 Proceedings of the Section on Survey Research Methods*, vol. 1, 333-338. Alexandria, VA: American Statistical Association.

Westat, Inc. (1996). *A User's Guide to WesVarPC, Version 2.0.* Rockville, MD: the author.

An Alternative Jackknife Variance Estimation for NAEP

Stanley Weng, Sameena Salvucci

1. Study Purpose

This empirical study explores an alternative method for performing jackknife variance estimation which makes better use of the sampling variation than the procedure currently used for the National Assessment of Educational Progress (NAEP), a periodic survey conducted by the National Center for Education Statistics (NCES). Better use of the sampling variation should improve the accuracy of the NAEP variance estimates. The alternative method should also make it possible to implement systematic computational procedures to conduct NAEP jackknife variance estimation.

2. NAEP Sample Design

The basic primary sampling unit (PSU) sample design for the main NAEP assessment is a stratified probability sample with one PSU selected per stratum with probability proportional to the population. The sampling unit within the PSU is the individual school. Schools are selected systematically with probability proportionate to the assigned measure of size. The sample of students within sampled schools is systematically drawn from school-prepared lists of eligible students.

3. Assignment of Sessions to Schools

All sampled students within a school are assigned to assessment sessions based on the following three age/grade eligiblity classes:

Age Class 1: Age 9/Grade 4

Age Class 2: Age 13/Grade 8

Age Class 3: Age 17/Grade 12

Print administered reading, writing, and mathematics sessions and tape administered mathematics sessions were conducted at all age classes. The method of determining the number and type of sessions to be administered in a given school varied by age class.

Our study was limited to examining standard errors for grade 8 reading proficiency estimates in the 1992 NAEP main assessment.

4. NAEP Jackknife Variance Estimation

The NAEP variance estimation procedure, as used for the 1992 and 1994 NAEP, uses a jackknife variance estimator. This method will be referred to as the original "paired" jackknife procedure.

For the purposes of variance estimation, pairs of first-stage sampling units (FSSUs) or of appropriate aggregates of them are defined in a manner that models the design as one in which two first-stage units are drawn with replacement per stratum. The definition and pairing of the FSSUs are different for the certainty and noncertainty PSUs. Each noncertainty PSU constitutes a single FSSU while each certainty PSU contains two or more sampled FSSUs, each consisting of one or more schools. The $2N$ noncertainty PSUs are formed into N pairs of FSSUs, where the pairs are composed of PSUs from adjacent strata and are thus relatively similar on the sample stratification characteristics. Whereas, as described in section 2 above, the actual sample design was to select one FSSU with probability proportional to size from each of $2N$ strata, for variance estimation purposes the design is regarded as calling for the selection of two FSSUs with probability proportional to size with replacement from each of N strata. This alteration probably produces a positive bias to estimates of sampling error.

Although the two-PSU-per-stratum jackknife is a simple procedure**,** it may not perform satisfactorily. The formation of the jackknife replicates greatly changed the original sampling design, and

it ignored much of the sampling variation contained in the sample, with a considerable reduction of the degrees of freedom for the estimation space.

5. NAEP Student Jackknife Replicates

The NAEP variances are bases on a set of student jackknife replicates (replicate weights) contained in each sample. Each main NAEP sample dataset contains a set of 56 jackknife replicates: 30 replicates reflect the amount of sampling variance contributed by the noncertainty strata of PSUs, and 26 reflect the variance contribution of the certainty PSU samples. The replicates were formed in the following way. The 60 noncertainty PSUs, drawn from 60 strata, were formed into 30 pairs, each pair composed of PSUs from adjacent strata within each subuniverse of sampling (thus the strata were relatively similar on the characteristics of stratification). The 26 replicates from the 34 certainty PSUs were created in a more complex way: the seven largest PSUs were assigned to ten replicates, the next five largest PSUs were assigned to one replicate each, and the remaining 22 were paired and assigned to 11 replicates.

6. Alternative Jackknife Variance Estimation

We propose an alternative jackknife procedure to better incorporate the data sampling structure into jackknifing and hence to catch more of the sample variation, and to be able to implement systematic computational procedures. NAEP's sample design has one PSU selected per stratum; therefore, there is no direct way to estimate sampling variance at the PSU level without collapsing strata. The alternative jackknife procedure performs jackknifing at the next sampling level, the school level; that is, the alternative procedure is a general stratified jackknife performed to schools within PSU. Since the sampling fraction of schools within PSU is small we assume they are independent. We expected the alternative to provide improved accuracy for the variance estimates.

In proposing the alternative jackknife procedure, we reviewed the jackknife variance estimation methodology (Shao and Tu, 1995, Shao and Wu, 1989).

7. Analysis and Results

Data

The 1992 NAEP Main Assessment Reading Test Age 13/Grade 8 data were used to conduct the alternative jackknife variance estimation. A SAS data set was created from the raw data in the 1992 NAEP National Assessment CD-ROM. The five composite variables for reading proficiency ("Plausible NAEP reading value") were used as response variables to estimate average reading proficiency for the nation and for the domains defined by Region (Northeast, Southeast, Central, West) and Type of School (Public, Private, Catholic), respectively. Missing cases for the response variables were deleted.

Estimation

We performed jackknife variance estimation using (1) our alternative jackknife procedure and (2) the original "paired" jackknife procedure. Since the our alternative jackknife variance estimation does not include nonresponse, trimming, and poststratification adjustments, we calculated comparable "unadjusted" variances using the original "paired" jackknife procedure. Therefore, in implementing the original "paired" jackknife procedure we used WesVar PC to develop a set of jackknife replicate weights based on the NAEP final student weight instead of using the student jackknife replicate weights available on the NAEP file because these weights already included nonresponse, trimming, and poststratification adjustments. We used the VPLX software (Fay, 1995) for implementing our alternative procedure and as stated above WesVar PC for the original procedure. VPLX has been shown to produce reliable jackknife estimates in a previous study (Weng, Zhang, & Cohen, 1995).

The grade 8 national and domain average reading proficiency estimates and their associated standard errors from the two jackknife procedures in comparison are presented in tables 1, 2, and 3, respectively.

For reference, table 4 lists the grade 8 average reading proficiency and associated standard errors provided by Mullis, Campbell, & Farstrup (1993). However, note that these standard errors

were based on the NAEP student replicate weights which were created to include nonresponse, trimming, and poststratification adjustments. Thus, these standard errors are not directly comparable to the standard errors that we calculated in our analyses.

Discussion

It can be seen from tables 1 and 3 that the standard error for average reading proficiency using our alternative jackknife procedure is just a little greater than that from the original jackknife procedure (except in Catholic schools). In addition, in table 2, the variance for the Central region using our alternative method is almost one third higher than when using the original method. This result conforms with our belief that the alternative jackknife would catch sampling variation ignored by the original jackknife. In comparing variances across the other domains, it can be seen that the variances are very similar. Also, since the alternative method has more degrees of freedom than the original method, the variance estimate precision is improved. Also, Shao and Tu (1995) discuss that the jackknife has some robustness properties against the violation of the school independence assumption.

Note, however, that the alternative jackknife cannot estimate the sampling variation at the NAEP PSU level within strata: the variance estimates provided by this procedure would generally be underestimated.

The two-PSU-per-stratum "paired" version of the jackknife procedure, as implemented in the WesVar software (Westat, 1996) now available on the Internet, has almost been adopted as a standard version of jackknife. It is in wide use for NCES survey variance estimation. This study provides useful information on the performance of such a jackknife procedure. The results of this analysis may be interesting as NCES considers how to improve jackknife variance estimation practice.

8. Further Steps

The alternative jackknife procedure for NAEP variance estimation seems promising. This study is only the first step in exploring how to improve jackknife variance estimation for NAEP. Further steps may be taken according to the following methodological consideration: Shao and Wu (1989) and Wu (1990) discussed the more general delete-*d* version of jackknife procedure, which, with appropriately chosen *d*, can be used to improve the performance of the variance estimation and make the jackknife variance estimator more robust.

Table 1. National grade 8 average reading proficiency and jackknife variance estimates

		Standard error calculated by		
Variable	**Average proficiency**	**Alternative method**	**Original method**	**Alternative s.e./ Original s.e.**
Reading proficiency 1	254.465	0.952	0.853	1.116
Reading proficiency 2	253.995	0.976	0.912	1.070
Reading proficiency 3	254.975	0.948	0.916	1.035
Reading proficiency 4	254.383	0.938	0.902	1.040
Reading proficiency 5	255.011	0.978	0.933	1.048
Average	254.566	0.958	0.903	1.062

Table 2. Domain grade 8 average reading proficiency and jackknife variance estimates, by region

Domain	Average proficiency	Standard error calculated by: Alternative method	Standard error calculated by: Original method	Alternative s.e./ Original s.e.
Northeast				
Reading proficiency 1	257.226	2.341	2.013	1.163
Reading proficiency 2	256.939	2.176	2.050	1.061
Reading proficiency 3	257.660	2.142	1.985	1.079
Reading proficiency 4	257.285	2.246	1.930	1.164
Reading proficiency 5	258.033	2.273	2.108	1.078
Average	257.429	2.236	2.017	1.109
Southeast				
Reading proficiency 1	247.418	2.111	2.265	0.932
Reading proficiency 2	246.601	2.109	2.421	0.871
Reading proficiency 3	247.707	2.059	2.458	0.838
Reading proficiency 4	247.526	2.012	2.434	0.827
Reading proficiency 5	247.524	2.178	2.331	0.934
Average	247.355	2.094	2.382	0.880
Central				
Reading proficiency 1	259.105	1.605	1.195	1.343
Reading proficiency 2	259.283	1.728	1.369	1.262
Reading proficiency 3	260.425	1.543	1.261	1.224
Reading proficiency 4	259.249	1.611	1.329	1.212
Reading proficiency 5	260.392	1.651	1.459	1.132
Average	259.691	1.628	1.323	1.235
West				
Reading proficiency 1	254.250	1.511	1.629	0.928
Reading proficiency 2	253.350	1.681	1.715	0.980
Reading proficiency 3	254.263	1.683	1.742	0.966
Reading proficiency 4	253.691	1.575	1.754	0.898
Reading proficiency 5	254.302	1.637	1.809	0.905
Average	253.971	1.617	1.730	0.935

Table 3. Domain grade 8 average reading proficiency and jackknife variance estimates, by type of school

Domain	Average proficiency	Standard error calculated by: Alternative method	Standard error calculated by: Original method	Alternative s.e./ Original s.e.
Public				
Reading proficiency 1	252.219	1.042	0.937	1.112
Reading proficiency 2	251.813	1.074	0.981	1.095
Reading proficiency 3	252.783	1.037	0.986	1.052
Reading proficiency 4	252.185	1.034	0.972	1.064
Reading proficiency 5	252.800	1.075	1.036	1.038
Average	252.360	1.052	0.982	1.072
Private				
Reading proficiency 1	280.323	2.853	2.817	1.013
Reading proficiency 2	279.919	2.627	2.421	1.085
Reading proficiency 3	280.862	2.812	2.538	1.108
Reading proficiency 4	279.618	2.457	2.497	0.984
Reading proficiency 5	281.336	3.037	2.800	1.085
Average	280.412	2.757	2.615	1.055
Catholic				
Reading proficiency 1	272.527	1.683	1.723	0.977
Reading proficiency 2	271.064	1.683	1.869	0.900
Reading proficiency 3	272.209	1.742	1.846	0.944
Reading proficiency 4	272.098	1.631	1.773	0.920
Reading proficiency 5	272.262	1.635	1.633	1.001
Average	272.032	1.675	1.769	0.948

Table 4. Grade 8 average reading proficiency and standard error

Domain	Average proficiency	Standard error
Nation[1]	260	0.9
Region[2]		
Northeast	263	1.8
Southeast	254	1.7
Central	264	2.2
West	260	1.2
Type of school[3]		
Public	258	1
Private	283	3
Catholic	275	1.9

SOURCE: Mullis et al. (1993), [1]table 1, [2]table 3, [3]table 2.

References

Campbell, J. R., Donahue, P. L., Reese, C. M. and Phillips, G. W. (1996). *NAEP 1994 Reading Report Card for the Nation and the States*. Washington, DC: National Center for Education Statistics.

Fay, R. E. (1995). *VPLX*. Washington DC: U.S. Bureau of the Census.

Johnson, E. G. (1989). Considerations and techniques for the analysis of NAEP data. *Journal of Educational Statistics*, 14, 303-334.

Johnson, E. G. and Allen, N. L. (1992). *The NAEP 1990 Technical Report.* NCES 92-067. Washington, DC: National Center for Education Statistics.

Johnson, E. G. and Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics*, 17, 175-190.

Johnson, E. G. and Carlson, J. E. (1994). *The NAEP 1992 Technical Report.* NCES 94-490. Washington, DC: National Center for Education Statistics.

Mullis, I. V. S., Campbell, J. R. and Farstrup, A. E. (1993). *NAEP 1992 Reading Report Card for the Nation and the States*. Washington, DC: National Center for Education Statistics.

Mullis, I. V. S., Jenkins, F. and Johnson, E. G. (1994). *Effective Schools in Mathematics*. NCES 94-701. Washington, DC: National Center for Education Statistics.

Rust, K., Burke, J., Fahimi, M. and Wallace, L. (1992). *1990 National Assessment of Educational Progress, Sampling and Weighting Procedures,* Part 2 - National Assessment. Rockville, MD: Westat, Inc.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.

Shao, J. and Wu, C. F. J. (1989). A general theory for jackknife variance estimation. *The Annals of Statistics*, 17, 1176-1197.

Wallace, L. and Rust, K. (1994). *1992 National Assessment of Educational Progress, Sampling and Weighting Procedures, Final Report.* Rockville, MD: Westat, Inc.

Wallace, L. and Rust, K.F. (1996). A comparison of raking and poststratification using 1994 NAEP data. Presented at the 1996 Joint Statistical Meeting, Chicago.

Weng, S. S., Zhang, F. and Cohen, M.P. (1995). Variance estimates comparison by statistical software, *ASA 1995 Proceedings of the Section on Survey Research Methods*, vol. 1, 333-338. Alexandria, VA: American Statistical Association.

Westat, Inc. (1996). *A User's Guide to WesVarPC, Version 2.0.* Rockville, MD: the author.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Wu, C. J. F. (1990). On the asymptotic properties of the jackknife histogram. *The Annals of Statistics*, 18, 1438-1452.

On the Performance of Replication-based Variance Estimation Methods with Small Numbers of PSUs

Ming-xiu Hu

Most surveys conducted by the National Center for Education Statistics (NCES) apply complex designs. For a complex survey, there is often no easy way to find unbiased and design-consistent variance estimates analytically. The standard statistical software packages, such as SAS and SPSS, provide inappropriate and usually too small variance estimates for survey statistics including totals, means, proportions. One solution to this difficulty is to use so-called *replication-based variance estimation approaches*, sometimes also called *resampling variance estimation approaches*. A number of replication methods have been proposed over years. Among them, the *simple and stratified jackknife, bootstrap, balanced repeated replication, Fay's method,* and *random group* method have received broad attention. The basic idea behind the replication methods is to select subsamples repeatedly from the whole sample, to calculate the statistic of interest for each of these subsamples, and then use the variability among these subsample or replicate statistics to estimate the variance of the full sample statistics.

This project is to evaluate the six replication-based variance estimation approaches mentioned above when only small numbers of primary sample units (PSU) are available. The problem of variance estimation with small numbers of PSUs happens most often with stratified multistage sampling, which is often adopted by NCES surveys. For example, in the 1993-94 Schools and Staffing Survey (SASS), private schools, which are considered the primary sample units (PSUs) in the private school teacher and student surveys, are stratified by association membership (19 groups), then by school levels (3 levels), and then by Census regions (4 regions), making a total of 228 strata in the private schools and staffing survey. Within each stratum, schools are further sorted by variables such as State, Highest grade in the school, Urbanicity, etc. After schools (PSUs) have been chosen, further sampling takes place to select the secondary units of teachers within each PSU. With this type of sampling design, although the total number of PSUs is very large, some strata (explicit and /or implicit) may only have small numbers of

PSUs but may contribute substantial numbers of secondary units to the sample. If we are interested in inferences on some subpopulation parameters, then we may encounter the problems of variance estimation with small numbers of PSUs since many subpopulations will only have small numbers of PSUs.

In case when a large sample of secondary units are drawn from only a few PSUs, it may be able to provide a pretty close point estimator, but the unreliability of the estimated sampling variance makes it difficult to construct confidence intervals with the desired levels of coverage. This is because direct variance estimators must, explicitly or implicitly, estimate the between PSU component of variance. The precision of this between-PSU variance estimator will be low due to the small number of PSUs. Burke and Rust (1995) conduct a simulation study to examine the performance of two Jackknife variance estimation methods, the usual Jackknife method and a paired Jackknife method, for systematic samples with small numbers of PSUs. Their simulation population consist of 105 private schools (a subset) of 1994 National Assessment of Educational Progress (NAEP) sample.

In this project, we conducted a simulation study on a subset of 1993-94 Schools and Staffing Survey (SASS) to examine the performance of the six replication-based variance estimation approaches stated earlier. Our simulation population consists of 182 private schools of SASS sample. It differs from Burke and Rust (1995) in five aspects: (1) different variance estimation methods. We compared six replication-based methods, while they only compared two Jackknife methods; (2) different evaluation criteria (see section 3); (3) different software used. Burke and Rust used WesVar but we use VPLX (Fay, 1994) and Resampling Stat (Version 4.04) to calculate variance estimates; (4) different statistics. Burke and Rust only considered non-linear statistics (average reading proficiency in a school), whereas we considered both linear statistics (totals of full-time equivalent teachers) and non-linear statistics (student-teacher ratios); (5) different simulation populations (as stated earlier).

In section 1, we will first briefly describe the six replication-based variance estimation methods under study and available software packages for implementing these methods. Section 2 will present the criteria used in our evaluation. The simulation population and the sample design will be described in

section 3. The simulation results and some statistical arguments will be given in section 4. Section 5 includes a summary of findings and our conclusions.

1. Replication-based Variance Estimation Approaches

Complex survey designs which combine sampling techniques such as sampling without replacement, stratification, multistage sampling, or unequal probability of selection, etc., induce a non-independently identical distribution structure to the data. Conventional techniques for variance estimation are often difficult to extend to these complex survey data structures or are cumbersome to implement. It is desirable to have replication-based variance estimation approaches that reuse the existing estimation system repeatedly, using computing power to avoid theoretical work. In recognition of this need, various replication-based methods have been proposed in the literature. These include the method of random group, the Jackknife method, the balanced repeated replication method (half-sample replication method), the modified half-sample replication method (Fay's method), and the bootstrap method. These methods have been implemented in a number of software packages, including WesVarPC (version 2.02, Westat) and VPLX (version 94.06, Fay).

We include a brief description of the six replication-based variance estimation approaches under study below. Details on these methods may be found in Wolter (1985), Fay (1989), Efron (1979, 1982), Sitter (1992) and the references cited therein.

1.1 Random Group Method

In this method, the total sample is randomly divided into K parts, called *random groups*, in a manner designed to represent the major sources of variation arising from the sample design. Suppose the estimator of the statistic of interest for the r-th group is $\hat{q}_r$ (r=1, 2, ..., K), and the estimator based on the overall sample is $\hat{q}$. The design-based estimators $\hat{q}_r$ and $\hat{q}$ are obtained through standard estimating approaches. Then the random group variance estimator is given by

$$\hat{v}_{rg}(\hat{q}) = \frac{1}{K(K-1)} \sum_{r=1}^{K} (\hat{q}_r - \hat{q})^2 \qquad (1)$$

or

$$\hat{v}_{rg}(\hat{\bar{q}}) = \frac{1}{K(K-1)} \sum_{r=1}^{K} (\hat{q}_r - \hat{\bar{q}})^2 , \qquad (2)$$

where $\hat{\bar{q}} = \sum_{r=1}^{K} \hat{q}_r \Big/ K$ is the average of the K estimators. It is apparent that (1) and (2) are identical for linear estimators. For non-linear estimators (1) is more conservative than (2) because

$$\hat{v}_{rg}(\hat{q}) = \sum_{r=1}^{K} (\hat{q}_r - \hat{q})^2 = \sum_{r=1}^{K} (\hat{q}_r - \hat{\bar{q}})^2 + K(\hat{q} - \hat{\bar{q}})^2 \geq \sum_{r=1}^{K} (\hat{q}_r - \hat{\bar{q}})^2 = \hat{v}_{rg}(\hat{\bar{q}}) .$$

Actually, (2) is an estimator for the variance of $\hat{\bar{q}}$ instead of $\hat{q}$, which is obtained based on the whole sample. However, in many complex surveys, the expectation of the squared difference $(\hat{q} - \hat{\bar{q}})^2$ will be unimportant and therefore there should be little difference between (1) and (2). The software package VPLX (Fay, 1994) uses estimator (1). Wolter (1985), however, in his discussion on the properties of the random group estimators, focuses on estimator (2), which is easier to discuss theoretically.

The random group method is perhaps the simplest replication method to understand, but its statistical properties make it one of the least attractive replication-based variance estimation methods (Fay, 1994). The random group method has been implemented in the following statistical software packages:

(1) VPLX V94.06 of Fay, U. S. Bureau of the Census (1994, public domain);

(2) OSIRIS IV of Kish et al., University of Michigan;

(3) CLUSTERS of Verma, University of Essex;

(4) PASS of Finch et al., U. S. Social Security Administration.

1.2 Jackknife Methods (Simple and Stratified)

Here we consider both the simple jackknife method and the stratified jackknife method.

The *simple jackknife method* creates replicate estimates based on all but one cluster in succession; that is, each replicate estimate omits one cluster while re-weighting the remaining K-1 clusters by the factor K/(K-1), where K is the total number of the clusters in the sample. Suppose the *r*-th replicate estimator of the interest parameter based on the sample which leaves *r*-th cluster out is $\hat{q}_r$ (r=1, 2, ..., K), and the estimator based on the overall sample is $\hat{q}$. Then the simple jackknife variance estimator used in our simulation is given by

$$\hat{v}_{jk}(\hat{q}) = \frac{K-1}{K}\sum_{r=1}^{K}(\hat{q}_r - \hat{q})^2 . \qquad (3)$$

Similarly to (2), we may use $\hat{\bar{q}} = \sum_{r=1}^{K}\hat{q}_r \Big/ K$ instead of $\hat{q}$ in (3), which will lead to smaller or equal jackknife variance estimates. For the jackknife approach, Efron and Stein (1981) show that even the later smaller jackknife estimates of variance tend to overestimate the variance of non-linear statistics on average. This implies that (3) will be worse in terms of positive bias. But VPLX implemented this form and we did not change it in our simulation.

For linear statistics, the simple jackknife variance estimator (3) is identical to the random group variance estimator (1) if the same clusters (groups) are used in the variance computation. However, for non-linear statistics, the two estimators are different.

Many complex designs employ stratification in which the universe is divided into distinct subpopulations and one subsample is independently drawn from each subpopulation. In these cases the *stratified jackknife method* generally has advantages over the simple jackknife procedure. To apply the stratified jackknife method, each stratum must have at least two clusters.

Suppose that S strata have been formed in a survey, and the *s*-th stratum has K_s (s=1, 2, …, S) clusters. Within *s*-th stratum, one cluster is omitted in turn and the remaining K_s-1 clusters in that cluster

are re-weighted by the factor $K_s/(K_s-1)$. Therefore, the stratified jackknife assumes that a given cluster represents the stratum from which it was drawn, not the population as a whole. Let $\hat{q}_{rs}$ (r=1, 2, …, K_s, s=1, 2, …, S) denote the estimator obtained from the re-weighted sample which consists of all the clusters but the *r*-th cluster in the *s*-th stratum, while $\hat{q}$ be the estimator based on the parent sample. Then, in our simulation, we will use

$$\hat{v}_{sjk}(\hat{q}) = \sum_{s=1}^{S} \frac{K_s - 1}{K_s} \sum_{r=1}^{K_s} (\hat{q}_{rs} - \hat{q})^2 \qquad (4)$$

as the stratified jackknife variance estimate.

Further details on the jackknife methods may be found in Wolter (1985).

The jackknife method has been implemented in the following software packages:

(1) PLX V94.06 of Fay, U. S. Bureau of the Census (1994, public domain);
(2) WesVarPC V 2.1 of Westat (1997, public domain);
(3) OSIRIS IV of Kish et al., University of Michigan;
(4) GES V4.0 of Statistics Canada (1997, commercial);
(5) BOJA of Boomsma, The Netherlands (1991, commercial).

1.3 Balanced Repeated Replication (BRR) Method

The half-sample replication method forms replicates using half of the sample each time. It is usually applied to stratified sample designs in which the sample consists of two clusters from each stratum (to apply it to non-stratified samples, we may create artificial strata). If some strata have more than two clusters, we may either group them into two superclusters or divide those strata into smaller (artificial) strata such that each stratum consists of two and only two clusters. After the desired strata have been created, one cluster from each stratum will be selected to form one replicate. There is a total of 2^S possible half-sample replicates, where S is the number of strata. The number of all possible half-

sample replicates becomes enormous quickly as S increases. We may choose K half-sample replicates randomly from all 2^S possible replicates with equal probabilities to calculate the variance estimates.

The balanced repeated replication method is a special half-sample replication method in which orthogonal balanced half-sample replicates are chosen to obtain variance estimates through Hadamard matrix (Wolter, 1985). The information contained in the 2^S replicates can be captured using K balanced replications. The minimum number of replicates needed to have full information is the smallest integer greater than or equal to S which is divisible by 4. For example, if there are 12 strata in the sample, then K=12 replicates are needed; if there are 15 strata, then 16 replicates are necessary. The BRR method is the most popular half-sample replication method. It gives the same variance estimates as that of the analytical procedure under simple random sampling design with replacement.

Suppose that a total of K half-sample replicates are used in the BRR variance estimation method. $\hat{q}_r$ (r=1, 2, ..., K) is the estimator based on the *r*-th half sample replicate, and $\hat{q}$ is the estimator based on the overall sample. Then the BRR variance estimator used in our simulation is given by

$$\hat{v}_{brr}(\hat{q}) = \frac{1}{K}\sum_{r=1}^{K}(\hat{q}_r - \hat{q})^2 \ . \qquad (5)$$

Again, the estimates of the statistics of interest, $\hat{q}_r$ and $\hat{q}$, are design-based and obtained through standard survey estimating approaches. Similarly, we may use $\hat{\bar{q}} = \sum_{r=1}^{K}\hat{q}_r \Big/ K$ instead of $\hat{q}$ in (5), which will lead to smaller (or equal) BRR variance estimates. Fay (1989) shows that (5) generally tends to produce overestimates of variance on average although there exist some exceptions to this rule.

More details on the BRR method can be found in Wolter (1985).

The BRR method has be implemented in the following software packages:

(1) VPLX V94.06 of Fay, U. S. Bureau of the Census (1994, public domain);

(2) WesVarPC V2.1 of Westat (1997, public domain);

(3) OSIRIS IV of Kish at el., University of Michigan;

(4) HESBRR of Jones, U. S. National Center for Health Statistics.

1.4 Fay's Method

Fay's method is a modified version of the BRR method. In the BRR method, half of the sample is zero-weighted while the other half is double-weighted. Fay's method assigns weight ρ ($0 \le \rho \le 1$) to one half sample and $2-\rho$ to the other half. If we use the same notations as in section 1.3, the variance estimator of Fay's method is given by

$$\hat{v}_{Fay}(\hat{q}) = \frac{1}{K(1-r)^2}\sum_{r=1}^{K}(\hat{q}_r - \hat{q})^2 \ . \qquad (6)$$

Similarly, $\hat{q}$ may be replaced by $\hat{\bar{q}}$ in (6), which will lead to less conservative variance estimates.

By choosing a value of ρ around 0.7, it is possible that Fay's method may do better for medians than the jackknife, while still doing well for statistics like ratios that are often better estimated by the jackknife (Westat, 1997). More information on this method may be found in Judkins (1990).

Fay's method has been implemented in the following software:

(1) VPLX V94.06 of Fay, U. S. Bureau of the Census (1994, public domain);

(2) WesVarPC V 2.1 of Westat (1997, public domain).

1.5 Bootstrap Method

Efron (1979, 1982) originated the bootstrap method. Suppose a sample S is drawn from a population U with some certain sampling design. The population parameter θ is estimated by $\hat{q}$, and our objective is to seek an estimator for the variance Var($\hat{q}$) through the bootstrap method. The bootstrap method consists of the following three steps:

(1) Using the sample data, construct an artificial population U*, assumed to mimic the real but unknown population U.

(2) Draw K independent samples, called resamples or bootstrap samples, from U* using a design identical to the one by which S was drawn from U. Independence implies that each sample must be replaced into U* before the next one is drawn. For each resample, calculate an estimate $\hat{q}_r$ (r=1, 2, …, K) in the same way as $\hat{q}$ is calculated.

(3) The observed distribution of $\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_K$ is considered an estimate of the sampling distribution of $\hat{q}$, and the bootstrap method estimated V($\hat{q}$) by

$$\hat{v}_{bs}(\hat{q}) = \frac{1}{K}\sum_{r=1}^{K}(\hat{q}_r - \hat{q})^2 \qquad (7)$$

or

$$\hat{v}_{bs}(\hat{q}) = \frac{1}{K-1}\sum_{r=1}^{K}(\hat{q}_r - \bar{\hat{q}})^2 \ . \qquad (8)$$

Here (8) is more like the usual sample variance estimate, while (7) is more like an MSE. In our simulation, we use (7) instead of (8) as bootstrap variance estimates since all the other replication methods implemented through VPLX software use the more conservative form. More information about the bootstrap method may be found in Efron and Tibshirani (1993).

No software product has yet been developed for the general bootstrap method. Such a product would not only be required to simulate bootstrap samples using different types of complex sampling designs, but also required to cooperate with different types of estimates for different types of statistics. So far, *BOJA* which is written by Boomsma (1991) and reviewed by Dalgleish (1995) may be the best software for the bootstrap method. The built-in S-PLUS function “sample” in S-PLUS for Windows (Version 3.3) may be used to generate bootstrap samples for simple random sampling or PPS random sampling schemes with or without replacement, but extra effort is needed to do data manipulation and variance estimation after the resamples are obtained. Another S-PLUS function, written by Tibshirani and available in STATLIB, may be used for some confidence interval variance estimates with the bootstrap method. *Resampling Stat for Windows* (Version 4.0) can only be used for the simple random sampling design. This student-level software is not very convenient for programming and its capacity is severely limited.

1.6 Summary

Replication variance estimates (1), (3), (5), (6), and (7) all take the form $c\sum_{r=1}^{K}(\hat{q}_r - \hat{q})^2$, where c is an adjusting constant which depends on the replication methods used. In the random group method, because only one cluster (or a supercluster) is used to estimate $\hat{q}_r$ for each replication, we should expect more variation among the replicated estimates. Hence $\sum_{r=1}^{K}(\hat{q}_r - \hat{q})^2$ should be the largest among these methods, which implies the smallest adjusting constant $c = 1/K(K-1)$ should be used in (1). On the other hand, since the jackknife uses all but one cluster for each replication, the variation among $\hat{q}_r$ (r=1, 2, …, K) should be the smallest and therefore the largest adjusting constant $c = (K-1)/K$ should be used in the jackknife variance estimate (3). The BRR method uses half of the sample in each replication; its adjusting constant $c = 1/K$ is between the 1/K(K-1) used for the random group and the (K-1)/K used for the jackknife. Fay's method uses more clusters (in fraction) than the BRR method and therefore it has a larger adjusting constant $c = 1/K(1-r)^2$ than the BRR. The bootstrap method has the same adjusting constant as the BRR method.

A very generalized replication variance estimation approach has also been proposed:

$$\hat{v}_g(\hat{q}) = \sum_{r=1}^{K} b_r(\hat{q}_r - \hat{q})^2 , \qquad (9)$$

where b_r is an adjusting coefficient, which will depend on the selection of replicate weights used for the estimates $\hat{q}_r$. This method has been implemented in VPLX V94.06 of Fay, U. S. Bureau of the Census. With this method, the user has to determine the replicate weights and the coefficients b_r for each replication.

2. Simulation Population, Sampling Scheme, and Implementation

To study the behavior of the six replication-based variance estimates, we chose two estimates—the student-teacher ratio (a non-linear statistic) and the total number of full-time equivalent teachers (a linear statistic)—from the 1993-94 Schools and Staffing Survey (SASS) private school data. In the 1993-94 SASS, private schools were stratified by Affiliation (19 affiliations), School Level (3 levels), and Census Region (4 regions). Within each stratum, the schools were further sorted by six variables: State, Highest Grade, Urbanicity, First Two Digits of Zip Code, 1991-92 Enrollment, and PIN number. Then the schools were systematically selected with probabilities proportionate to their sizes (systematic PPS sampling) from each stratum. The measure of size used was the square root of the number of teachers obtained in the 1991-92 Private School Survey (PSS). In the SASS survey, schools serve as the primary sample units (PSU) for the SASS teacher and student surveys (Abramson et al., 1996).

Our artificial simulation population consists of 182 private schools from the four smallest affiliations in the 1993-94 SASS: 26 schools from the Association of American Military Colleges and Schools, 60 from the Friends Council on Education, 44 from the Solomon Schechter Day Schools, and 50 from Other Lutheran affiliation. The original SASS design was projected to include all the schools from these affiliations, but not all of them responded. We included all the respondents of these four affiliations in our simulation population.

The 182 private schools in the artificial population were first divided into three strata by the school level variable: elementary, secondary, and combined. Within each stratum, the schools were further sorted by the same six sorting variables used in the original SASS design. Then the systematic PPS sampling algorithm was used to select the schools. The measure of size for each school was the same as in the original SASS sampling design. We studied the performance of the six replication variance estimation methods for sample sizes (number of PSUs) 2, 4, 6, …, 30.

In our simulation, we employed the systematic PPS sampling scheme used in the original SASS, but we did not exactly apply its stratification strategies. A stratified sampling scheme first allocates a sample size to each stratum, then draws a subsample from each stratum, and then combines all the subsamples into one overall sample. In our simulation, we needed to compute variance estimates for all possible samples. If we had applied the stratification strategy, the number of all possible samples would have become too large to implement. Therefore we decided not to pre-allocate the sample size to each stratum before performing systematic PPS sampling.

Although we did not pre-allocate the sample sizes to the strata, the subsample sizes of the strata obtained through the non-stratified systematic PPS sampling scheme was almost identical to what a stratified sampling scheme would have allocated to the strata if we had employed a stratification strategy. For example, for sample size 20, the samples obtained via the non-stratified systematic PPS sampling scheme have 12 elementary schools, 3 secondary schools, and 5 combined schools, which is exactly the same allocation a stratified sampling scheme would produce. Therefore, we applied the stratified jackknife method anyway for sample sizes over 12 although we did not use the stratified sampling design to obtain our samples.

For each sample size n (n=2, 4, …, or 30), there is a total of 182 possible systematic PPS samples, the same number as the artificial population size. This is the case for most systematic PPS sampling designs. An Excel spreadsheet was used to assist the implementation of the systematic PPS sample selection.

We only chose even numbers as sample sizes to make it easier to implement the BRR and Fay's method. For the BRR and Fay's methods, every two adjacent PSUs were grouped into an artificial stratum. Full orthogonal balanced replicates were generated for the BRR method through the Hadamard matrix.

For the bootstrap method, we used a non-systematic PPS sampling scheme to draw re-samples from the artificial population constructed by each possible sample. Suppose y_k (k=1, 2, …, n) is a

sample S with size n, and π_k is the inclusion probability of unit k under the systematic PPS sampling design. The artificial population U* for this sample may be formed by creating replicates of each element in the sample. For unit k (k=1, 2, …, n), $1/\pi_k$ artificial elements (pretending that $1/\pi_k$ is an integer) will be created for U*, all of which share the same value of y_k. Then n+1 re-samples of size n will be drawn using the PPS sampling scheme from U*. Actually, this is equivalent to drawing n+1 simple random samples with replacement directly from the sample S instead of the artificial population U*. The re-sample selection for the bootstrap was implemented by Resample Stat for Windows (Version 4.0).

The random group and jackknife methods needed no special treatment to generate replicates. After all the possible systematic PPS samples had been selected for each sample size, we only needed to run VPLX once for each sample size to obtain variance estimates for all possible samples with that size. In order to use one run of VPLX to calculate the variance estimates for all samples, a sample indicator variable had to be created to distinguish different samples in the data set. This was true for all the replication methods except the bootstrap method for which we used Resampling Stat for Windows instead of VPLX for variance estimation.

3. Evaluation Criteria

We employed the following criteria in our evaluation of the six replication-based variance estimation methods.

(1) Bias: As usual, bias of the variance estimates is defined as the difference between the expected variance and the true variance of $\hat{q}$

$$Bias = E\hat{v}(\hat{q}) - Var(\hat{q}). \qquad (10)$$

Under our design, the true variance of $\hat{q}$ is given by

$$Var(\hat{q}) = E(\hat{q} - E(\hat{q}))^2 = \sum_{i=1}^{182} p_i \left(\hat{q}_{0i} - E(\hat{q})\right)^2, \qquad (11)$$

where $\hat{q}_{0i}$ is the estimator of θ based on the *i*-th sample (i=1, 2, …, 182), p_i is the inclusion probability of the *i*-th sample, and $E(\hat{q}) = \sum p_i \hat{q}_{0i}$ is the expectation of $\hat{q}$ over all possible samples. While the expectation of the variance estimates is given by

$$E\hat{v}(\hat{q}) = \sum_{i=1}^{182} p_i \hat{v}_i(\hat{q}_{0i}), \qquad (12)$$

where $\hat{v}_i(\hat{q}_{0i})$ is the variance estimate for the *i*-th sample obtained through some replication method, which may be denoted by v_i below for simplicity.

(2) MSE, variance, CV of the variance estimates: Under our design, the variance of the variance estimates is given by

$$Var(\hat{v}) = E(\hat{v} - E\hat{v})^2 = \sum_{i=1}^{182} p_i (\hat{v}_i - E\hat{v})^2, \qquad (13)$$

where $E\hat{v}$ is given by (12). MSE of the variance estimates is

$$MSE = E\left(\hat{v} - Var(\hat{q})\right)^2 = Var(\hat{v}) + Bias^2, \qquad (14)$$

and the CV of the variance estimates is defined as

$$CV = \sqrt{Var(\hat{v})} \Big/ E\hat{v}. \qquad (15)$$

(3) Coverage probability of covering the true value of q**:** The primary interest in Burke and Rust (1995) is the coverage probabilities of the 95 percent confidence intervals.

$$\hat{q}_{0i} \pm 1.96\sqrt{\hat{v}_i}$$

and

$$\hat{q}_{0i} \pm t(0.975, df)\sqrt{\hat{v}_i}$$

covering the true value of θ, where *t(0.975, df)* is the 97.5 percentile of the t-distribution with a degree of freedom of *df*. $\hat{q}_{0i}$ is the estimator based on the *i*-th parent sample and does not depend on the replication methods, while $\hat{v}_i$ varies from one replication method to another; that is, the above intervals have the same center but different widths for different replication methods. Larger variance estimates will lead to higher coverage probabilities. In our situation, this further implies that higher coverage

probabilities are almost equivalent to larger positive biases of variance estimates because all the replication variance estimation methods tend to overestimate the true variance. Therefore, a worse replication method will have higher coverage probabilities in most cases, which contradicts the usual sense of coverage probabilities. We do not think that this is an appropriate criterion for evaluating replication-based variance estimation methods, but we include it since Burke and Rust used it as the criterion of primary interest.

We only considered intervals with t-coefficient; that is, $\hat{q}_{0i} \pm t(0.975, df)\sqrt{\hat{v}_i}$, since our sample sizes were small. In this type of confidence interval, we used K-1 as the degrees of freedom for all the replication methods except the stratified jackknife, where K is the number of replicates. For the stratified jackknife, the degrees of freedom is $n_1 + n_2 + n_3 - 3$, where n_s (s=1, 2, 3) is the number of observations in the *s*-th stratum.

(4) Coverage probabilities of covering the true variance: We also compared the six replication methods in terms of the coverage probabilities that the intervals

$$\hat{v}_i \pm 1.96\sqrt{Var(\hat{v}_i)}$$

cover the true value of variance, where $Var(\hat{v}_i)$ is given by (13). For different replication methods, not only the width $2 \times 1.96\sqrt{Var(\hat{v}_i)}$ but also the center $\hat{v}_i$ of the interval vary. A method with higher coverage rates and shorter confidence intervals will be considered a better method.

(5) 95 percent confidence interval estimates of the true variances: 95 percent confidence interval estimates for the variances were obtained directly from the distribution of the replication variance estimates based on all 182 possible PPS systematic samples. They did not depend on the standard deviation of the variance estimates. A better method is the one that provides shorter confidence interval estimates and covers the true variance.

4. Analysis of Simulation Results

In this section, we present our simulation results and compare the six replication variance estimation methods using the criteria presented above. As stated earlier, our simulation population consists of 182 private schools, and even-numbered sample sizes (number of PSUs) from 2 to 30 are considered. Three school levels, elementary, secondary, and combined, are used in the stratified jackknife method. VPLX was used to perform the variance estimation for the random group, both jackknife, BRR, and Fay's methods, while Resampling Stat was used to carry out the calculation of variance estimates for the bootstrap method. In Fay's method, $\rho=0.5$ was used; that is, one half sample was weighted by 0.5, and the other half by 1.5.

4.1 Comparison of Bias

Tables 1 and 2 present the biases of the variance estimates for the student-teacher ratio and the total number of the full-time equivalent teachers, respectively, for all the replication methods. The corresponding plots are given by figures 1 and 2.

The first column of the two tables gives the true variances for all the sample sizes under study. Generally, we would expect the variance to decrease as sample size increases, but we have some cases which obviously violate this trend. For the student-teacher ratio, the true variance for sample sizes 18, 22, and 24 are much smaller than we expected. This is probably because the systematic sampling scheme hits some pattern in the population so that the average variation among all possible systematic samples are much smaller than the average variation among all possible random samples. On the other hand, for sample size 26, the true variance is larger than we expect, which is probably because the average variation among all possible systematic samples is larger than the average variation among all possible random samples. We should keep in mind that we are trying to estimate the design-based variance; that is, the variance among all possible systematic samples, and have no interest in the variance among all possible random samples since our estimates of the student-teacher ratio and the total of the full-time equivalent teachers are based on systematic samples.

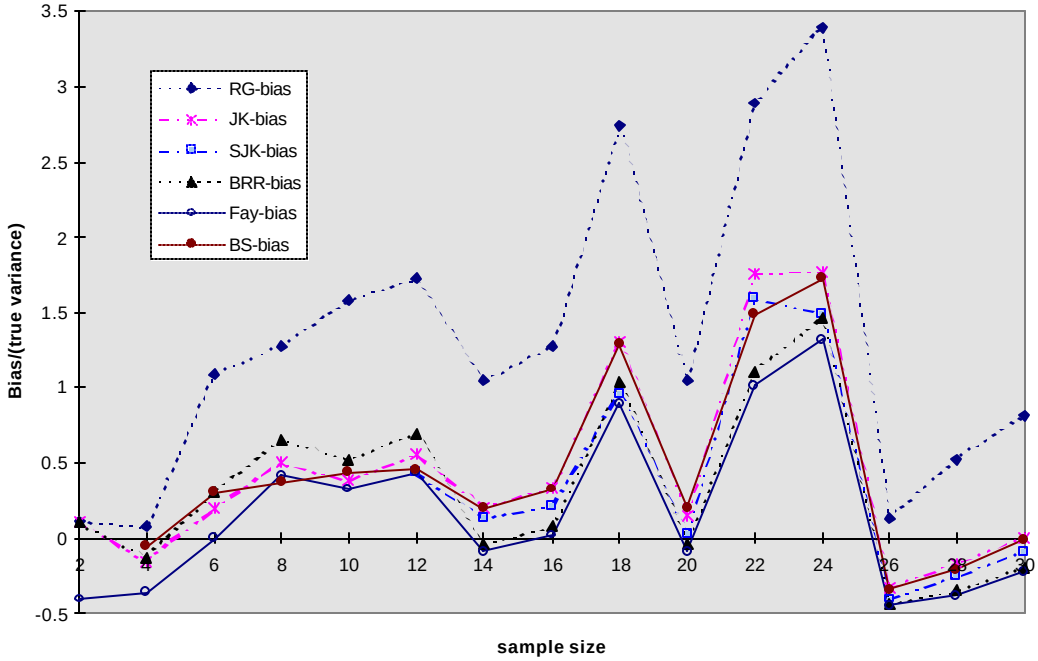
For the total of full-time equivalent teachers, the true variance for sample sizes 18, 22, 24 are again much smaller than we expected. For sample size 26, the true variance for the student-teacher ratio is too large, as we noticed earlier, but it is now too small for the total of full-time equivalent teachers. Similar reasons are responsible for the results. We should not be surprised if the replication methods encounter some problems with these four cases.

From figure 1 and table 1, it is evident that all of the six replication methods on average tend to overestimate the variance of the student-teacher ratio. One reason for this phenomenon is that our samples are drawn without replacement (hereafter we call them *WOR samples*), while the replication methods assume that all the samples are drawn with replacement (hereafter we call them *WR samples*). A WOR sample generally has larger within-sample variation. If we treat a WOR sample as a WR sample, we will overestimate the true variance.

Table 1. Bias of the variance estimates for the student-teacher ratio

Sample size	True variance	Random group	Simple jackknife	Stratified jackknife	BRR	Fay's method	Bootstrap
2	9.8274	1.0471	1.0471		1.0471	-3.9026	
4	5.0131	0.3858	-0.7350		-0.6642	-1.7992	-0.2499
6	1.9082	2.0730	0.3682		0.5764	0.0081	0.5910
8	1.2428	1.5924	0.6212		0.8209	0.5182	0.4587
10	0.8926	1.4078	0.3443		0.4665	0.2888	0.3898
12	0.7122	1.2238	0.3985	0.3123	0.5015	0.3138	0.3280
14	0.7858	0.8275	0.1678	0.1014	-0.0369	-0.0704	0.1575
16	0.6202	0.7896	0.2112	0.1341	0.0510	0.0112	0.2042
18	0.3367	0.9215	0.4415	0.3249	0.3482	0.3009	0.4331
20	0.5485	0.5757	0.0824	0.0206	-0.0199	-0.0489	0.1133
22	0.2622	0.7571	0.4612	0.4185	0.2891	0.2657	0.3893
24	0.2117	0.7186	0.3740	0.3165	0.3087	0.2785	0.3658
26	0.7385	0.1009	-0.2443	-0.2978	-0.3197	-0.3304	-0.2518
28	0.5227	0.2715	-0.0875	-0.1282	-0.1837	-0.2001	-0.1065
30	0.4070	0.3329	0.0021	-0.0343	-0.0812	-0.0870	-0.0019

Figure 1. Bias of the variance estimates for the student-teacher ratio (in the scale of the true variance)



Actually, as discussed by Efron and Stein (1981), and Fay (1989), even if the samples are drawn with replacement, the jackknife, random group, and half-sample methods still tend to overestimate the variance in most cases.

For the student-teacher ratio, the random group method always has the highest positive bias, so is obviously the worst in terms of bias, while Fay's method always has the lowest negative bias. Since all the replication methods tend to overestimate the variance, Fay's method appears to be the best in terms of bias except for the sample sizes 2, 4, 26, 28, and 30. Actually, Fay's method is good except when sample size equals 2 and 4, while for the other three cases all the methods except the random group are close in terms of bias. This probably means that Fay's method breaks down for non-linear statistics when the sample size is too small (≤4). But it becomes the best or close to the best thereafter.

In terms of bias, both the simple and stratified jackknife, BRR, and bootstrap are all comparable for non-linear statistics. All six methods have very large positive biases when sample size equals 18, 22, and 24. As we stated earlier, these cases have very small true variance. True variance

actually measures the variation among all possible parent samples, while each replication variance estimate is based on resamples from one parent sample. If the resamples mimic the parent samples well, we expect the replication variance estimate to be close to the true variance. However, if the within-parent-sample variation is much larger than the between-parent-sample variation (which may be considered variation in the population), then the variation between the resamples will be much larger than the variation between the parent samples, and therefore the replication method will overestimate the true variance. This is what happens for sample sizes 18, 22, and 24. On the other hand, most methods have the largest negative biases when the sample size equals 26, which implies that the within-parent-sample variation is smaller than the between-parent-sample variation for this case.

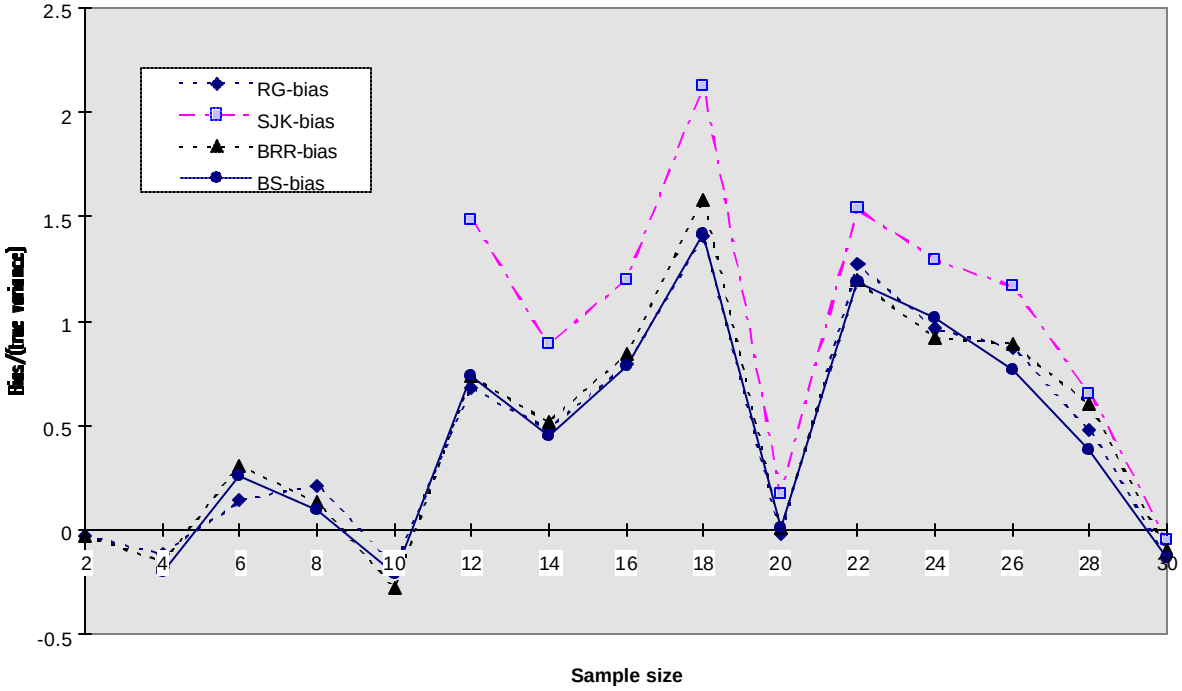
Table 2. Bias of the variance estimates for the total of full-time equivalent teachers (in millions)

Sample size	True variance	Random group/ Simple jackknife	Stratified jackknife	BRR/ Fay's method	Bootstrap
2	2.4807	-0.0694		-0.0694	
4	1.3399	-0.1559		-0.2031	-0.2745
6	0.7288	0.1038		0.2236	0.1885
8	0.5151	0.1102		0.0679	0.0496
10	0.5776	-0.0982		-0.1594	-0.1210
12	0.2512	0.1707	0.3725	0.1845	0.1858
14	0.2417	0.1160	0.2146	0.1241	0.1084
16	0.1756	0.1388	0.2108	0.1483	0.1383
18	0.1168	0.1641	0.2481	0.1847	0.1655
20	0.2493	-0.0049	0.0437	0.0024	0.0023
22	0.1004	0.1278	0.1547	0.1197	0.1194
24	0.1060	0.1021	0.1372	0.0976	0.1074
26	0.1023	0.0893	0.1192	0.0913	0.0787
28	0.1197	0.0571	0.0783	0.0718	0.0456
30	0.1863	-0.0240	-0.0088	-0.0186	-0.0256

For the total of full-time equivalent teachers, figure 2 and table 2 show that all methods except the stratified jackknife are comparable. The stratified jackknife always has the largest positive biases. Two reasons may be responsible for this phenomenon: (1) we did not actually use stratification in our sampling design, and therefore, when we used the stratified jackknife method to estimate the variance, we probably introduced extra variance; (2) the overall sample size is not large enough, and consequently some strata have too few clusters, which leads to large variance within those strata.

In summary, when the sample size equals 18, 22, and 24, all the methods have very large positive bias compared to the true variance, which implies that the within-sample variation is much larger than the population variation. This is very likely caused by the systematic sampling design since both linear and non-linear statistics have the largest positive biases. For the total of full-time equivalent teachers, we do not have any very large negative biases, but we have more cases with large positive biases such as the cases when the sample size equals 12, 16, and 26. As we mentioned earlier, most methods showed their largest negative bias for the non-linear statistic, the student-teacher ratio, for sample size 26.

Figure 2. Bias of the variance estimates for the total of full-time equivalent teachers (in the scale of true variance)



NOTE: The simple jackknife and Fay's method have not been plotted in figure 2 since, for the linear estimator the total of full-time equivalent teachers, the simple jackknife is equivalent to the random group and Fay's method is equivalent to the BRR.

4.2 Comparison of MSE of the Variance Estimates

Table 3 and table 4 give the MSEs of the variance estimates for the student-teacher ratio and the total of full-time equivalent teachers, respectively.

For the student-teacher ratio, table 3 shows that the random group provides much less accurate variance estimates than any other replication methods in terms of MSE of the variance estimates. In many cases, the MSEs of the variance estimates obtained from the random group are more than ten times larger than those from the other replication methods. The large biases of the variance estimates of the random group account for a major part of its large MSEs.

Table 3. MSE of variance estimates for the student-teacher ratio

Sample size	Random group	Simple jackknife	Stratified jackknife	BRR	Fay's method	Bootstrap
2	1699.344	1699.344		1699.344	163.612	
4	107.552	84.243		52.741	22.743	184.062
6	43.312	5.530		13.912	4.092	10.013
8	16.097	1.716		4.084	2.289	2.040
10	5.732	0.429		1.668	0.763	0.902
12	5.408	0.539	0.458	1.432	0.458	0.668
14	2.985	0.358	0.340	0.291	0.245	0.305
16	2.235	0.158	0.124	0.138	0.120	0.172
18	1.949	0.329	0.191	0.207	0.169	0.378
20	1.138	0.068	0.052	0.098	0.085	0.134
22	1.142	0.660	0.541	0.166	0.139	0.498
24	0.950	0.233	0.182	0.191	0.144	0.245
26	0.480	0.113	0.130	0.133	0.136	0.140
28	0.325	0.024	0.031	0.048	0.052	0.034
30	0.371	0.024	0.018	0.026	0.024	0.033

When the sample size is less than or equal to 12, the BRR behaves very poorly in terms of MSEs of variance estimates. However, when the number of PSUs is greater than or equal to 14, the BRR catches up with the other methods and sometimes does even better, which means there is a sample size breakdown point for the BRR method. For non-linear statistics, the BRR method should not be used if the number of PSUs is very small.

Overall, Fay's method is the best in terms of MSE of the variance estimates. It almost always has smaller MSEs than the BRR method. Sample size 22 seems to be a breakdown point for all other methods except the BRR and Fay's method. The stratified jackknife is among the best except for sample size 22. The simple jackknife is a little worse than the stratified jackknife but a little better than the bootstrap. The bootstrap catches up gradually with the other methods as the sample size increases. When the sample size is greater than or equal to 24, all the methods except the random group are comparable.

Table 4 presents the MSEs of the variance estimates for the linear statistic, the total of the full-time equivalent teachers. Here, the random group/simple jackknife has better overall performance than the other four replication methods in terms of MSE. The stratified jackknife method has the largest MSEs except the last case when the sample size is 30, in which it has the smallest MSE. This implies that, for linear statistics, it is not a good idea to use stratification in the replication variance estimation approaches if the sample size is not large enough. Based purely on this simulation, we believe that, in order to apply the stratification strategy to obtain more precise variance estimates, each stratum should have at least five clusters although the method requires only two or more clusters per stratum.

Table 4. MSE of the variance estimates for the total of full-time equivalent teachers ($\times 10^{10}$)

Sample size	Random group	STR-jackknife	BRR	Bootstrap
2	2236.48		2236.48	
4	253.43		202.33	243.84
6	83.18		131.80	141.65
8	33.91		28.36	45.65
10	14.56		9.04	15.56
12	13.51	30.08	21.70	18.55
14	8.32	13.72	10.61	10.35
16	5.59	7.56	5.76	6.98
18	4.32	8.52	5.56	5.74
20	2.58	3.57	2.92	3.47
22	2.45	3.20	2.33	2.98
24	2.78	3.77	3.58	3.61
26	1.28	1.99	1.62	1.43
28	1.14	1.31	1.29	1.03
30	0.67	0.42	0.79	0.72

NOTE: For the total of full-time equivalent teachers, the simple jackknife is identical to the random group, and Fay's method is indistinguishable from the BRR.

For linear statistics, no obvious advantages or disadvantages have been found between the BRR/ Fay's method and Bootstrap in terms of MSE. Overall, these two are a little worse than the random group/simple jackknife, but always better than the stratified jackknife except for sample size 30. As the sample size increases, the differences between these methods become smaller and smaller. As the sample size becomes large enough (s=30), we should expect that the stratified jackknife will have better performance and may be better than the other methods.

4.3 Comparison of Coverage Probabilities of Covering the True Value of q

Table 5 presents the coverage rates of the intervals $\hat{q}_{0i} \pm t(0.975, df)\sqrt{\hat{v}_i}$ covering the true value of the student-teacher ratio, which is 10.454 in our simulation population.

Most of the coverage rates in table 5 can be explained through our examination of biases earlier in section 4.1: (1) for sample sizes equal to 18, 22, and 24, all the methods overestimate the true variance by quite a large amount, and therefore the intervals $\hat{q}_{0i} \pm t(0.975)\sqrt{\hat{v}_i}$ are too wide, which implies too high coverage rates for those cases (almost always 100%); (2) the random group always has the largest positive biases, which implies that it has wider intervals and higher coverage rates than any other method in most cases; (3) Fay's method has the lowest bias, which implies that it has narrower intervals and lower coverage rates than any other method in most cases; (4) all the replication methods tend to overestimate the variance, and therefore most of the coverage rates are very high.

Similarly, for the total of the full-time equivalent teachers, most of the coverage rates in table 6 can be explained by the bias analysis presented in section 4.1: (1) Since the stratified jackknife method has the largest positive biases, it has the widest intervals, which (almost always) leads to the highest coverage rates; (2) for sample size 16, 18, 22, and 24, all the coverage rates are very large (over 96%) because the positive biases are very large at these points for all the methods; (3) since all the replication methods tend to overestimate the true variance, the coverage rates are always high. The coverage rates are all over 90 percent except for sample size 4. But even for sample size 4—the worst case, the coverage rate is still around 85 percent.

Table 5. Coverage rates of covering the true value of the student-teacher ratio

Sample size	Random group	Simple jackknife	Stratified jackknife	BRR	Fay's method	Bootstrap
2	0.9602	0.9602		0.9602	0.9553	
4	0.9914	0.9335		0.9242	0.9034	0.9048
6	0.9866	0.9554		0.9468	0.9370	0.9485
8	1	0.9962		0.9784	0.9784	0.9872
10	0.9768	0.9372		0.9588	0.9492	0.9525
12	1	1	0.9982	0.9875	0.9817	0.9801
14	0.9808	0.9861	0.9657	0.9475	0.9237	0.9749
16	0.9871	1	0.9974	0.9943	0.9943	0.9935
18	1	1	1	1	1	1
20	1	0.9891	0.9891	0.9450	0.9445	0.9545
22	1	1	1	1	1	0.9931
24	1	1	1	1	1	1
26	0.9677	0.9477	0.9428	0.9295	0.9365	0.9093
28	0.9616	0.9483	0.9482	0.8928	0.8928	0.9492
30	1	1	0.9851	0.9558	0.9517	0.9838

Table 6. Coverage rates of covering the true value of the total of full-time equivalent teachers

Sample size	Random Group/ Simple Jackknife	Stratified Jackknife	BRR/ Fay's Method	Bootstrap
2	0.9275		0.9275	
4	0.8630		0.8455	0.8643
6	0.9075		0.9132	0.8997
8	0.9640		0.9663	0.9371
10	0.9803		0.9207	0.9441
12	0.9776	1	0.9521	0.9613
14	0.9363	0.9638	0.9494	0.9193
16	1	1	1	0.9984
18	1	1	1	0.9949
20	0.9272	0.9710	0.9584	0.9308
22	0.9887	1	0.9887	0.9716
24	0.9710	0.9902	0.9686	0.9662
26	1	1	1	0.9957
28	0.9822	0.9828	0.9822	0.9768
30	0.9724	0.9708	0.9350	0.9578

NOTE: For the total of full-time equivalent teachers, the simple jackknife is identical to the random group, and Fay's method is indistinguishable from the BRR.

This type of coverage rate is the primary interest in Burke and Rust (1995) when they compare the two jackknife methods. We doubt this is an appropriate criterion for the evaluation of the replication-based variance estimation approaches due to three reasons: (1) the replication methods tend to overestimate variance, and, therefore, this type of coverage rate is high and not worrisome as seen in

their simulations and our simulations; (2) in most cases, higher coverage rates imply worse variance estimation approaches, which contradicts the usual sense of coverage probabilities; (3) if the normality assumption of the estimates does not hold, it is not appropriate either to compare the coverage rates to 95 percent, the nominal level.

4.4 Coverage Rates of Covering the True Variance

In this section, we discuss the coverage rates of the intervals $\hat{v}_i \pm 1.96\sqrt{Var(\hat{v}_i)}$ covering the true variance. For different replication methods, both the widths and the centers of the intervals may be different. A method with higher coverage rates and narrower widths is considered better. To compare the widths of the intervals, we present the standard deviation of the variance estimates here.

Table 7 shows that the standard deviations of the variance estimates for the random group method are often three times larger than those for other methods in most cases, which implies that the intervals corresponding to the random group will be 6 times wider than those corresponding to the other methods. With much wider intervals, the random group still does not show any sign of higher coverage rates, which means that the centers $\hat{v}_i$ of the intervals are much farther away from the true variance. This again shows that the random group method provides very inaccurate variance estimates for the student-teacher ratio.

In table 7, all non-highlighted coverage rates are over or close to 90 percent. The bootstrap has no alarmed values of coverage rates, while the simple jackknife only has one at sample size 26, which still has a coverage rate close to 80 percent. However, for sample sizes 26 and 28, Fay's method, BRR, and stratified jackknife methods all break down in terms of coverage rate of covering the true variance. This is because, for these two cases, the three methods underestimate the true variance by considerable amounts (as shown by the largest negative biases in table 1) and the variation among the variance estimates is very small, which leads to too short confident intervals. For sample size 18, these three methods also have pretty low coverage rates, especially the BRR and Fay's method. We can not blame inaccurate variance estimates this time because the bias analyses and MSE analyses both show

Table 7. Coverage rates of covering the true variance and standard deviation of variance estimates for the student-teacher ratio (upper entries are coverage rates and lower entries are standard deviations)

Sample size	Random group	Simple jackknife	Stratified jackknife	BRR	Fay's method	Bootstrap
2	0.9916 41.21	0.9916 41.21		0.9916 41.21	0.9832 12.18	
4	0.9833 10.36	0.9758 9.15		0.9694 7.23	0.9516 4.44	0.9784 13.56
6	0.9749 6.25	0.9667 2.32		0.9730 3.69	0.9464 2.02	0.9483 3.11
8	0.9666 3.68	0.9123 1.15		0.9178 1.85	0.9300 1.42	0.9239 1.35
10	0.9446 1.936	0.9058 0.557		0.9544 1.204	0.9512 0.825	0.9128 0.866
12	0.9499 1.977	0.9443 0.616	0.9253 0.602	0.9474 1.086	0.8941 0.600	0.9477 0.748
14	0.9283 1.517	0.9387 0.574	0.9267 0.578	0.9332 0.539	0.9573 0.490	0.9433 0.529
16	0.9332 1.270	0.9070 0.336	0.9279 0.326	0.9596 0.367	0.9596 0.346	0.9384 0.361
18	0.9248 1.049	*0.8819* *0.366*	*0.8508* *0.292*	**0.7658** 0.293	**0.7768** 0.279	*0.8916* *0.436*
20	0.9165 0.898	*0.8966* *0.247*	0.9182 0.228	0.9450 0.313	0.9818 0.288	0.9253 0.348
22	*0.8951* *0.754*	0.9037 0.669	0.9037 0.605	0.9037 0.286	0.9037 0.261	0.9158 0.588
24	*0.8998* *0.659*	*0.8949* *0.305*	*0.8949* *0.286*	*0.8949* *0.310*	*0.8949* *0.257*	*0.8894* *0.333*
26	*0.8914* *0.686*	**0.7996** 0.230	**0.5552** 0.202	**0.4788** 0.176	**0.4502** 0.164	0.9230 0.277
28	*0.8831* *0.501*	0.9077 0.126	**0.7134** 0.122	**0.6950** 0.118	**0.5659** 0.110	*0.8935* *0.152*
30	*0.8747* *0.510*	0.9241 0.155	0.9698 0.130	*0.8867* *0.130*	*0.8809* *0.126*	0.9353 0.182

that these methods have smaller biases and smaller MSEs than the other methods. Therefore, for sample size 18, the BRR, Fay's method, and the stratified jackknife have low coverage probabilities simply because the coverage intervals are too narrow. In this case, we have no reason to reject these three methods except that our primary interest is to construct confidence interval estimates for the true variance.

The bootstrap does not have any low coverage rates, but never has very high coverage rates either (less than 95% for all the cases except for sample size 4 due to the widest interval). Based purely on this criterion, the bootstrap and the simple jackknife are among the best, which are mostly because they have moderately larger standard deviations at the points where Fay's method, the BRR, and the stratified jackknife break down according to this criterion. The bootstrap and the simple jackknife are recommended over Fay's method, the BRR, and the stratified jackknife only if we have more interest in the variance estimate than the estimate of the parameter itself.

Table 8. Coverage rates of covering the true variance and standard deviation of the variance estimates (in millions) for the total of full-time equivalent teachers

Sample	Random group		STR-Jackknife		BRR		Bootstrap	
size	C-rate	SD-VE	C-rate	SD-VE	C-rate	SD-VE	C-rate	SD-VE
2	0.9553	4.729			0.9553	4.729		
4	0.9630	1.584			0.9354	1.408	0.9469	1.537
6	0.9737	0.906			0.9441	1.126	0.9460	1.175
8	0.9650	0.572			0.9682	0.528	0.9731	0.674
10	0.9562	0.369			0.9447	0.255	0.9701	0.376
12	0.9474	0.326	0.9474	0.402	0.9474	0.428	0.9404	0.389
14	0.9387	0.264	0.8492	0.302	0.8730	0.301	0.9269	0.303
16	0.9188	0.191	**0.8261**	0.177	**0.8299**	0.189	0.8985	0.225
18	**0.7501**	0.128	**0.6177**	0.154	**0.7016**	0.147	0.8757	0.173
20	0.9124	0.161	0.9124	0.184	0.9124	0.171	0.9484	0.186
22	0.8495	0.091	**0.7104**	0.090	**0.7824**	0.095	0.8874	0.124
24	0.8949	0.132	0.8949	0.137	0.8949	0.162	0.9135	0.157
26	**0.7336**	0.069	**0.6091**	0.075	**0.8044**	0.089	0.8879	0.090
28	0.8774	0.090	0.8774	0.084	0.8774	0.088	0.9194	0.091
30	0.9046	0.078	0.9386	0.064	0.9397	0.087	0.9600	0.081

NOTE: For the total of full-time equivalent teachers, the simple jackknife is identical to the random group, and Fay's method is indistinguishable from the BRR.

For the total of full-time equivalent teachers, table 8 shows that the stratified jackknife has very low coverage rates and thus is obviously worse than the other methods. It has only 61 percent coverage

rates for sample sizes 18 and 26, and 71 percent coverage rate for sample size 22, which are not acceptable.

Seven out of 10 cases have lower than 90 percent coverage rates and all of them are lower than 95 percent, the nominal level. But its standard deviations of variance estimates are not significantly smaller, and sometimes even larger, than the others, which implies that the widths of the intervals are not the main reasons for the low coverage rates. The main reason for the low coverage rates is that the stratified jackknife provides very inaccurate variance estimates, which agrees with the findings of the bias analyses and the MSE analyses.

The random group/simple jackknife has two low coverage rates of 75 and 73 percent, respectively, when the sample size equals 18 and 26. But the random group has the smallest MSEs and almost smallest biases for these two cases. Therefore, the coverage rates are low mainly because the coverage intervals are too short.

The BRR/Fay's method has four low coverage rates, 83, 70, 78, and 80 percent, for sample sizes 16, 18, 22, and 26, respectively. Both poor variance estimates and short coverage intervals are responsible for the low coverage rates for these cases.

In terms of coverage rates of covering the true variance, the bootstrap method and the random group/simple jackknife have the best performance. The bootstrap has no breakdown point (all coverage rates are over 87.5) and has more cases with higher coverage rates, while the random group (simple jackknife) almost always has shorter coverage intervals (except for sample size 4, in which they are close).

4.5 95 Percent Confidence Interval Estimates and Their Widths

Table 9 presents 95 percent confidence interval estimates and their widths for the variances of the student-teacher ratio estimates which are obtained through the distribution of the variance estimates based on all possible PPS systematic samples.

In table 9, the highlighted confidence intervals do not cover the true variances. In all of these cases, the true values sneak out of the intervals from the lower limits, which means that at least 97.5 percent of variance estimates are larger than the true variance. They are seriously positively biased. The random group and the simple jackknife both have three such bad cases, with sample sizes 18, 22, and 24, the stratified jackknife has two with sample sizes 22 and 24, and the bootstrap has one with sample size 24. For the three disturbing cases, the BRR and Fay's method cover all the true variances with convincingly shorter intervals. Further, Fay's method is consistently better than the BRR and the difference is considerable.

For the student-teacher ratio, with this criterion, Fay's method is the obvious choice. It provides sharp and robust interval variance estimates for the non-linear statistic. Both jackknife methods sometimes provide very sharp estimates, but they may break down when the variation among the design-based samples is very different from the variation among random samples in the population. The BRR is as robust as Fay's method, but it is not sharp. The confidence interval estimates of the bootstrap are considerably wider than those of Fay's method, but it does not break down as easily as the jackknife. The random group is not worth considering. It not only gives much wider interval estimates, but breaks down easily as well.

For the total of full-time equivalent teachers, table 10 shows that Fay's method/the BRR again has the best performance overall. Its 95 percent confidence intervals always cover the true variances, and it more likely provides shorter interval estimates than any other method, but the degree of dominance is much less overwhelming than it is in the estimation of variances for the student-teacher ratios. The random group/the simple jackknife sometimes provides very short interval estimates for the true variances, but it is not robust, as shown by the two seriously positive cases (sample size 18 and 24)

in which the 95 percent confidence intervals can not cover the true values. All confidence interval estimates of the bootstrap cover the true value, but, again, this method does not seem very sharp.

The stratified jackknife obviously has the worst overall performance for the linear statistic. It has three seriously biased cases (sample sizes 18, 22, and 26) in which the 95 percent confidence interval estimates can not cover the true variances. Its lower confidence limits always have the highest values, but it never gives very short confidence intervals. This implies that it has a greater tendency to overestimate the variance, which agrees with our findings in the bias analyses. The random group (the simple jackknife) always has the second largest lower confidence limits, following the stratified jackknife. This may sometimes imply sharper interval estimates, but other times it may mean that this method more likely overestimates the variance compared to the BRR/Fay's method, although this was not shown in our bias analyses.

Table 9. 95 percent true confidence interval and interval width for the true variance of the student-teacher ratio estimate

Sample size	True variance	Random group	Simple jackknife	Stratified jackknife	BRR	Fay's method	Bootstrap
2	9.8274	.011~38.4 38.39	.011~38.4 38.39		.011~ 38.4 38.39	.011~28.2 28.23	
4	5.0131	.196~21.3 21.03	.158~21.7 21.58		.077~20.0 19.97	.067~16.1 16.02.	.126~20.5 20.39
6	1.9082	.477~13.8 13.33	.366~7.11 6.742		.138~9.55 9.412	.110~6.91 6.796	.283~10.2 9.927
8	1.2428	.368~13.7 13.31	.336~4.66 4.326		.207~6.68 6.475	.205~5.68 5.471	.190~5.18 4.985
10	0.8926	.450~13.1 12.63	.407~2.80 2.394		.245~3.51 3.262	.235~3.02 2.787	.331~3.46 3.132
12	0.7122	.482~10.0 9.548	.365~2.96 2.595	.326~2.64 2.314	.323~3.60 3.277	.308~2.33 2.024	.330~2.58 2.250
14	0.7858	.387~7.16 6.771	.286~2.63 2.342	.258~2.55 2.287	.168~2.01 1.843	.167~1.85 1.679	.297~1.99 1.696
16	0.6202	.275~5.98 5.701	.354~1.82 1.464	.299~1.54 1.238	.257~1.74 1.478	.236~1.68 1.440	.280~1.94 1.662
18	0.3367	**.345~4.71** 4.368	**.384~1.94** 1.551	.236~1.38 1.146	.223~1.33 1.102	.254~1.21 0.958	.269~2.07 1.796
20	0.5485	.494~3.97 3.473	.338~1.31 0.976	.284~1.16 0.877	.163~1.19 1.029	.148~1.08 0.929	.197~1.55 1.353
22	0.2622	**.315~3.12** 2.808	**.322~2.99** 2.664	**.274~2.66** 2.388	.238~1.42 1.186	.229~1.29 1.057	.196~2.73 2.529
24	0.2117	**.399~3.02** 2.624	**.307~1.51** 1.205	**.219~1.46** 1.238	.204~1.53 1.328	.198~1.26 1.065	**.247~1.49** 1.245
26	0.7385	.294~2.91 2.614	.221~1.05 0.826	.225~.876 0.651	.134~.789 0.655	.134~.752 0.618	.137~1.13 0.997
28	0.5227	.299~2.11 1.806	.257~.672 0.415	.217~.616 0.399	.133~.589 0.456	.132~.568 0.436	.182~.743 0.561
30	0.4070	.282~1.70 1.422	.250~.785 0.535	.228~.746 0.518	.096~.669 0.573	.094~.643 0.549	.176~.908 0.732

Table 10. 95 percent true confidence interval and interval width for the variance of the estimate of the total of full-time equivalent teachers (in millions)

Sample size	True variance	Random group	Stratified jackknife	BRR	Bootstrap
2	2.4807	(.003, 14.9) 14.847		(.003, 14.9) 14.847	
4	1.3399	(.034, 4.94) 4.902		(.013, 4.43) 4.415	(.006, 5.17) 5.166
6	0.7288	(.083, 3.50) 3.418		(.034, 4.26) 4.222	(.048, 4.21) 4.165
8	0.5151	(.127, 2.90) 2.776		(.054, 1.97) 1.919	(.058, 1.92) 1.857
10	0.5776	(.136, 1.87) 1.737		(.073, .997) 0.924	(.129, 1.36) 1.235
12	0.2512	(.079, 1.51) 1.432	(.181, 1.99) 1.808	(.069, 2.01) 1.939	(.066, 1.61) 1.547
14	0.2417	(.098, 1.09) 0.996	(.166, 1.37) 1.202	(.051, .920) 0.869	(.064, 1.13) 1.069
16	0.1756	(.071, .877) 0.806	(.141, .925) 0.784	(.058, .743) 0.685	(.064, .931) 0.867
18	0.1168	**(.137, .618)** 0.481	**(.154, .732)** 0.578	(.075, .611) 0.536	(.091, .685) 0.594
20	0.2493	(.086, .708) 0.622	(.128, .823) 0.695	(.055, .726) 0.671	(.073, .781) 0.708
22	0.1004	**(.105, .468)** 0.363	**(.132, .491)** 0.359	(.096, .416) 0.320	(.088, .584) 0.496
24	0.1060	(.069, .549) 0.480	(.105, .601) 0.496	(.061, .684) 0.623	(.051, .658) 0.607
26	0.1023	(.088, .319) 0.231	**(.111, .354)** 0.243	(.073, .412) 0.339	(.061, .412) 0.351
28	0.1197	(.066, .392) 0.326	(.085, .393) 0.308	(.053, .357) 0.304	(.048, .399) 0.351
30	0.1863	(.070, .297) 0.227	(.103, .360) 0.257	(.059, .451) 0.392	(.056, .393) 0.337

NOTE: For the total of full-time equivalent teachers, the simple jackknife is identical to the random group, and Fay's method is indistinguishable from the BRR.

5. Summary and Recommendations

All the replication methods tend to overestimate the true variance on average for both linear and non-linear statistics. When the systematic sampling design hits some underlying pattern in the population so that the average variation among all possible systematic samples is much smaller than the average variation among all possible random samples, the replication methods will produce variance estimates with very serious positive biases. For example, in our simulation population, sample sizes 18, 22, and 24 are bad cases of this kind.

Since the replication methods tend to overestimate the variance, the confidence intervals $\hat{q}_{0i} + t(0.975)\sqrt{\hat{v}_i}$ always have very high coverage rates for covering the true parameter. Since higher coverage rates in this case are almost equivalent to higher positive biases, we do not think that this is a good criterion for evaluating replication variance estimation methods. We included this criterion because Burke and Rust (1995) used it as the key criterion in their simulation to evaluate two jackknife methods.

For non-linear statistics, the random group should not be considered a candidate for variance estimation. It always gives much larger biases, much larger MSEs, and much broader interval estimates for the variances which are sometimes still unable to cover the true values. Although our simulation is for small sample sizes, we do not recommend using this method even for large sample sizes since no evidence shows that the random group gets closer to the other methods. We believe that the random group will not perform so poorly if more PSUs are included in each random group, but it requires a large number of PSUs since each PSU is used only once by the random group method.

For non-linear statistics, Fay's method has the best overall performance for non-linear statistics in terms of bias, MSE, and confidence interval estimates for variance estimation. Although Fay's method has very low coverage rates of the intervals $\hat{v}_i + 1.96\sqrt{Var(\hat{v})}$ covering the true variance for sample sizes 18, 26, and 28, this is mainly because the intervals are too short. Fay's method is always

recommended, except when constructing this type of confidence interval estimates for the true variances.

For non-linear statistics, Fay's method is a modified version of the BRR method. According to the criteria used in our simulation, this kind of modification has considerably improved the BRR. The BRR performed poorly when the sample size is smaller or equal to 12. As the sample size increases, it becomes closer to Fay's method.

For non-linear statistics, the stratified jackknife produces very sharp variance estimates on some occasions, but sometimes it provides seriously positively biased estimates when the average variation among design-based samples is much smaller than the average variation among all possible random samples. On the other hand, the bootstrap method never gives very sharp variance estimates, but it never gives very bad variance estimates either. It has slightly larger MSE, slightly broader interval estimate for the true variance compared to the best method in most cases, but the three types of coverage rates are always high, even for the cases when the other replication methods break down.

For non-linear statistics, the simple jackknife is slightly worse than the stratified jackknife in terms of bias, MSEs, and interval variance estimates, but slightly better in terms of coverage rate of covering the true variance. As the sample size increases, the stratified jackknife may have significant advantages over the simple jackknife.

For linear statistics, the random group and the simple jackknife are identical, while the BRR and Fay's methods are indistinguishable. The random group/simple jackknife have the overall best performance in terms of MSE, but they lose to the BRR/Fay's methods in terms of confidence interval estimates for the true variance.

For linear statistics, the stratified jackknife has the overall worst performance according to all the criteria used in the simulation. The bootstrap again does not have very sharp variance estimates, but has no very bad variance estimates either, which is similar to the behavior the bootstrap demonstrates

with the non-linear statistic. It has slightly larger MSEs and slightly broader interval estimates compared to the best ones, but it always gives pretty high coverage rates of covering the true variances, even for the cases when the other replication methods break down. The BRR and Fay's methods are close to the bootstrap in terms of bias, MSE, and interval variance estimates, but they have two very low coverage rates for covering the true variance for sample sizes 18 and 22 when the average variation among all possible systematic samples is much smaller than the average variation among all possible random samples.

Therefore, based on this simulation, we generally recommend Fay's method for variance estimations for ratio estimates when the number of PSUs are more than 4; the random group should not be considered. For linear statistics, no replication method stands out as significantly better than another. The random group/simple jackknife, the bootstrap, and the BRR/Fay's method all are possible choices. However, when the sample sizes are not large enough, it may not be a good idea to apply the stratified jackknife method in the variance estimation.

References

Abramson, R., Cole, C., Fondelier, S., Jackson, B., Parmer, R. and Kaufman, S. (1996*). 1993-94 Schools and Staffing Survey: Sample Design and Evaluation*. NCES 96-089. U.S. Department of Education. Office of Educational Research and Improvement. Washington, D.C.: National Center for Education Statistics.

Boomsma, A. (1991). *BOJA: A Program for Bootstrap and Jackknife Analysis*. Iec ProGAMMA, Groningen, The Netherlands.

Burke, J. and Rust, K. (1995). On the Performance of Jackknife Variance Estimation for Systematic Samples with Small Numbers of Primary Sampling Units. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 321-326.

Dalgleish, L. I. (1995). Software Review: Bootstrapping and Jackknifing with BOJA. *Statistics and Computing*, 5, 165-174.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist*., 7, 1-26.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: SIAM.

Efron, B. and Stein, C. (1981). The Jackknife Estimates of Variance. *Ann. Statist*., 9, 586-596.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Fay, R. E. (1994). *Replication Variance Estimation Methods for Complex Survey Data.* Unpublished Manuals for software VPLX (Version 94.06).

Fay, R. E. (1989). Theory and Application of Replicate Weighting for Variance Calculations. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 212-217.

Judkins, D. R. (1990). Fay's Method for Variance Estimation. *Journal of Official Statistics*, 6 223-239.

Sitter, R. R. (1992). A Resampling Procedure for Complex Survey Data. *Journal of the American Statistical Association*, 87 (419), 755-765.

Westat (1997). *A User's Guide to WesVarPC*. Unpublished.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

An Empirical Study of the Limitation of Using SUDAAN for Variance Estimation

Fan Zhang

1. Introduction

In most NCES surveys, complex sampling designs are employed to deal with the complexity of the problem and reduce the cost. These designs often combine techniques such as multistage sampling, stratification, clustering, systematic sampling, etc. Therefore, it is not always easy to track the variance estimators. For example, since the Schools and Staffing Survey (SASS) 1993-94 Public School component has a stratified systematic design, it is not possible to get an unbiased, or even consistent, estimator of the design variance. In other words, an analytic form of unbiased variance estimator does not exit for this type of design.

In practice, this problem is overcome by applying replication methods to calculate the variances. In replication methods (e.g., jackknife, BRR, Bootstrap) subsamples are selected repeatedly from the full sample, then the statistics of interest are calculated for each subsample, and the variability among these replicate statistics is used to estimate the variance of the full sample statistics. Therefore, replication methods do not require an analytic form of variance estimator for the complex design. Often replicate weights are created and attached to the data file for users to calculate the variances using replication methods. For example, the Bureau of the Census, as a contractor for the National Center for Education Statistics, included 48 sets of replicate weights corresponding 48 bootstrap subsamples on the SASS 1993-94 Public School sample data file. The subsamples were selected systematically without replacement to mimic the original sampling, so the bootstrap variance estimation should be close the true variance.

It is, however, fairly common for users to treat a complex design as a simpler design and use an analytic variance estimator for the simpler design as an approximation for the variance estimator under

the complex design. This approach is often seen in software applications such as SUDAAN or PC CARP, which apply the Taylor series method for variance estimation. The Taylor series method first substitutes a linear statistic for the non-linear statistic of interest and then uses an analytic textbook variance estimator for this linear statistic to calculate the variance estimate. Unfortunately, the design options available in these software applications are limited. Users who do not find the appropriate underlying complex design may select a similar option, subjecting their variance estimates to bias. Therefore, using SUDAAN, for example, to estimate the variances for the SASS 1993-94 Public School sample may result in greater bias than using the bootstrap variances described above.

This study uses SASS 1993-94 Public School component data to compare three different approaches to developing variance estimates:

- Bootstrap method using the bootstrap replicate weights attached to the data file, performed by WesVar PC®;
- Taylor series method under a stratified with replacement sampling design, with SUDAAN (design option = STRWR); and
- Taylor series method under a stratified without replacement sampling design, with SUDAAN (design option = STRWOR).

Section 2 describes the SASS 1993-94 Public School sampling design. Section 3 discusses the variance estimation methods used in this study. Section 4 is an analysis of the results.

2. SASS 1993-94 Public School Sampling Design

The SASS 1993-94 Public School Survey has a stratified one stage systematic design. The sample was selected with a probability proportionate to size algorithm. (See Abramson et al. 1996 for a detailed description.)

Public schools were first stratified at three levels. The first level of stratification is by school type:

(A) BIA (Bureau of Indian Affairs) schools
(B) Native American schools
(C) Schools in Delaware, Nevada, and West Virginia, and
(D) All other schools.

The second level of stratification was by states within the (B), (C), and (D) strata. The third level of stratification was performed within each second level stratum by grade level (elementary, secondary, and combined schools).

Then the non-BIA schools were sorted by the following variables:

State,
Local education agency (LEA) metro status,
Recoded LEA Zip code,
Common Code of Data (CCD) LEA ID number,
Highest grade in school,
School percent minority,
School enrollment, and
CCD school ID.

All BIA schools were selected into the sample. Within each non-BIA stratum, schools were systematically selected using a probability proportionate to size algorithm. The measure of size that SASS used for the schools was the square root of the number of teachers in the school as reported on the CCD file.

3. Variance Estimation Methods

Bootstrap Method

As mentioned above, Bureau of the Census statisticians included in the SASS Public School data file 48 replicate weights corresponding 48 bootstrap samples selected systematically without replacement to mimic the original sampling. They subsequently reweighted the bootstrap replicate basic weights (inverse of the probability of selection) by processing each set of replicate basic weights through the same weighting procedure used to create the full sample weights (Abramson et al. 1996). This should make the bootstrap variance estimation better reflect the true variance. In our study, we used these 48 bootstrap replicate weights to calculate variance estimates using WesVar PC®.

Let $\hat{q}$ be the estimate of q based on the full sample and $\hat{q}_k$ be the estimate of q based on the k-th bootstrap sample; the bootstrap variance estimator used in this study is (Westat, 1995)

$$v(\hat{q}) = \frac{1}{48}\sum_{k=1}^{48}(\hat{q}_k - \hat{q})^2 .$$

Taylor Series Methods

Six specific design options are available in SUDAAN (Shah et al., 1995):

1) With Replacement: DESIGN=WR
   - Sampling with replacement at the first stage
   - Sampling with or without replacement at subsequent stages
   - With equal or unequal probabilities of selection at both the first and subsequent stages

2) Without Replacement: DESIGN=WOR
   - Sampling without replacement at the first stage
   - Sampling with or without replacement at subsequent stages
   - With equal probabilities of selection at both the first and subsequent stages

3) Unequal Probabilities Without Replacement: DESIGN=UNEQWOR

- Sampling without replacement with unequal probabilities of selection at the first stage
- Sampling with equal probabilities at subsequent stages, with or without replacement

4) Stratified With Replacement: DESIGN=STRWR

- A single-stage design
- Stratified random sampling with replacement
- Equal or unequal probabilities of selection within each stratum

5) Stratified Without Replacement: DESIGN=STRWOR

- A single-stage design
- Stratified random sampling without replacement
- Equal probabilities of selection within each stratum

6) Simple Random Sampling: DESIGN=SRS

- A single-stage design
- Simple random sampling

Options 4 and 5, STRWR and STRWOR, are special cases of single stage WR and WOR, respectively, except they are more computationally efficient. Option 6, SRS, is equivalent to standard statistical software such as SAS. Thus SUDAAN accommodates three basic types of sample designs: WR, WOR, and UNEQWOR. However, Option 3, UNEQWOR, requires users to provide the joint probabilities of selection for each pair of PSUs within each first-stage stratum. As this information is rarely available, UNEQWOR is not often used.

Since there is no unbiased design variance estimator for systematic sampling design, a lot of approximate estimators have been proposed and studied (Wolter, 1985). In practice, two frequently used approaches to handling this problem are to treat the systematic sample as a with replacement

sample from a finite population or a without replacement simple random sample from a finite population, corresponding to the WR (or STRWR for single stage design) and WOR (or STRWOR for single stage design) design options in SUDAAN. Under a simple random sampling with replacement design, the variance estimator of the population total estimator $\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h = \sum_{h=1}^{H} N_h \sum_{k=1}^{n_h} y_{hk} \big/ n_h$ is

$$\hat{V}_{WR}(\hat{Y}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h} s_h^2$$

Here $s_h^2 = \sum_{k=1}^{n_h} (y_{hk} - \hat{\bar{y}}_{hk})^2 \big/ (n_h - 1)$, $\hat{\bar{y}}_{hk} = \sum_{k=1}^{n_h} y_{hk} / n_h$. Under a simple random sampling without replacement design, the variance estimator of the population total estimator $\hat{Y}$ is

$$\hat{V}_{WOR}(\hat{Y}) = \sum_{h=1}^{H} \frac{N_h^2 (1 - f_h)}{n_h} s_h^2$$

Here $f_h = n_h / N_h$.

In general, $E\left[\hat{V}_{WOR}(\hat{Y})\right] < E\left[\hat{V}_{WR}(\hat{Y})\right]$. Let $V_{SY}(\hat{q})$ denote the variance of $\hat{q}$ under systematic sampling design; then we hope $\hat{V}_{WOR}(\hat{q})$ or $\hat{V}_{WR}(\hat{q})$ are about the same or slightly conservative for $V_{SY}(\hat{q})$. But these are not true in general. If we actually consider the situation of no stratification for simplicity, assuming there are total a possible systematic samples represented by $s_r, r = 1,2,\ldots,a$, then the design effect of the total estimator $\hat{Y} = N \sum_{k=1}^{n} y_k \big/ n$ under the systematic sampling design is

$$deff(SY, \hat{Y}) = \frac{V_{SY}(\hat{Y})}{V_{SRSWOR}(\hat{Y})} = 1 + \frac{n-1}{1-f} d$$

Here $d = 1 - (N-1)SSW \big/ [(N-a)SST]$, $SST = \sum_{k=1}^{N} (y_k - \bar{y}_U)^2$ represents the total sum of squares, $SSW = \sum_{r=1}^{a} \sum_{s_r} (y_k - \bar{y}_{s_r})$ represents the within systematic sample sum of squares, $\bar{y}_{s_r} = \sum_{s_r} y_k \big/ n$ is the sample mean of the *r*-th systematic sample, and $\bar{y}_U = \sum_{k=1}^{N} y_k \big/ N$ is the population mean. It can be shown that $SST = SSW + SSB$, where $SSB = \sum_{r=1}^{a} n (\bar{y}_{s_r} - \bar{y}_U)^2$ represents the between systematic sample sum of squares. See, for example, Särndal , Swensson, and Wretman (1992) section 3.4.

Therefore, systematic sampling is more efficient than simple random sampling without replacement if $d < 0$. In other words, the more homogeneous the elements within systematic samples are, the less efficient the systematic sampling is. It can also be shown that systematic sampling is more efficient than simple random sampling with replacement if $d < 1/N$. To create a situation where these conditions hold, users commonly strive for an ordering that entails a low degree of homogeneity among elements within the same systematic sample. However, an ordering which is suitable for one variable may not be good for another variable. Therefore, in large scale surveys, systematic sampling may not consistently more efficient or less efficient than simple random sampling.

To implement the STRWR approach in SUDAAN, we simply need the stratum variable (in this case STRATM, included in the public data file). This is probably the approach most public data users will adopt if they use SUDAAN. To implement the STRWOR approach in SUDAAN, we will also need the population counts for each stratum. We put all certainty PSUs together as a new stratum and recalculated the population counts for each stratum. This certainty PSU stratum does not contribute variance to the variance estimates.

4. Variance Estimate Outputs

Standard errors for the total estimator are listed in table 1. The estimates from STRWOR and STRWR are quite different from the Bootstrap method estimates. In fact, the ratio of STRWOR and STRWR standard error estimates to the Bootstrap standard error estimates ranges from 44.21 percent to 134.58 percent. Variables DCNOST, DCNOTE, S0255, and S0455 are highly correlated with the measure of school size (the square root of total number of teachers), which is proportional to the selection probability. In table 2, the standard error estimates of proportions from STRWOR and STRWR are bigger than the Bootstrap method estimates.

Variables DCNOST, DCNOTE, S0255, and S0455, which show big differences between standard errors in table 1, were also used to construct the ratio estimates in table 3. The ratio estimate

for DCNOST/DCNOTE has 13 percent and 16 percent smaller standard errors when calculated from STRWOR and STRWR than by the Bootstrap method, while S0455/S0255 shows about the same standard error estimates for all methods.

5. Summary

This study demonstrates the limitation of software programs like SUDAAN when applied to more complex designs, such as systematic sampling. Software programs which apply Taylor series method often have limited design options available. When the underlying sample design is different from the design options available in the software, approximation is inevitable, which incurs bias.

Table 1. Standard errors of the totals

Variable	Label	Estimate total	Standard error			Ratio of standard error	
			Bootstrap	WOR	WR	WOR/Boots	WR/Boots
DCNOST	Total Students*	41179175	401044	198855	206963	49.58%	51.61%
DCNOTE	Total Teachers*	2339065	21691	9589	9997	44.21%	46.09%
S0255	Total Students	41621660	393746	208871	217536	53.05%	55.25%
S0455	Male Students	21232672	209225	110167	114663	52.65%	54.80%
S0405	American Indian Students	453042	10604	9931	11243	93.65%	106.03%
S0410	Asian Students	1396638	62953	64242	66257	102.05%	105.25%
S0415	Hispanic Students	4969062	178946	158454	161929	88.55%	90.49%
S0420	Black Students	6781341	117841	153563	158594	130.31%	134.58%
S0425	White Students	28021397	265950	226396	233813	85.13%	87.92%
S1365	Students in Remedial Reading Program	4526677	102326	101468	103896	99.16%	101.53%
S1375	Students in Remedial Math Program	2871518	92492	90021	92211	97.33%	99.70%
S1385	Students in Program for Disabilities	2862212	36281	36009	37527	99.25%	103.43%
S1395	Students in G. T. Program	2675964	57977	60313	62569	104.03%	107.92%

* DCNOST and DCNOST are frame variables known to all units in the frame.

Table 2. Standard errors of proportions

Variable	Label	Estimate proportion	Standard error			Ratio of standard error	
			Bootstrap	WOR	WR	WOR/Boots	WR/Boots
S1360	Remedial Reading Program Available	80.90%	0.535	0.569	0.588	106.4%	109.9%
S1370	Remedial Math Program Available	60.95%	0.725	0.734	0.755	101.2%	104.1%
S1380	Program for Disabilities Available	89.15%	0.463	0.516	0.532	111.4%	114.9%
S1390	GT Program Available	70.73%	0.544	0.67	0.691	123.2%	127.0%
S1435	Med. Health Care Service Available	58.73%	0.719	0.73	0.751	101.5%	104.5%
S1440	Have a Library	95.64%	0.338	0.365	0.378	108.0%	111.8%

Table 3. Standard errors of ratios

Statistics	Label	Estimate ratio	Standard error			Ratio of standard error	
			Bootstrap	WOR	WR	WOR/Boots	WR/Boots
DCNOST/DCNOTE	Student/Teacher	17.6	0.052	0.0438	0.0454	84.2%	87.3%
S0455/S0255	Male/Total	0.51	0.001	0.001	0.001	100.0%	100.0%

References

Abramson, R., Cole, C., Fondelier, S., Jackson, B., Parmer, R. and Kaufman, S. 1996. *1993-94 Schools and Staffing Survey: Sample Design and Estimation*. NCES 96-089. U.S. Department of Education, Office of Educational Research and Improvement. Washington, DC: National Center for Education Statistics.

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

Shah, B.V., Barnwell, B.G. and Bieler, G.S. 1995. *SUDAAN User's Manual*. Release 6.40. Research Triangle Park, NC: Research Triangle Institute.

Westat, Inc. 1995. *A User's Guide for WesVar* PC® . Rockville, MD: the author.

Wolter, K.M. 1985. *Introduction to Variance Estimation*. New York: Springer-Verlag New York, Inc.

Listing of NCES Working Papers to Date

Working papers can be downloaded as pdf files from the NCES Electronic Catalog (http://nces.ed.gov/pubsearch/). You can also contact Sheilah Jupiter at (202) 502–7444 (sheilah_jupiter@ed.gov) if you are interested in any of the following papers.

Listing of NCES Working Papers by Program Area

No.	Title	NCES contact
Baccalaureate and Beyond (B&B)		
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
2001–15	Baccalaureate and Beyond Longitudinal Study: 2000/01 Follow-Up Field Test Methodology Report	Andrew G. Malizio
Beginning Postsecondary Students (BPS) Longitudinal Study		
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
1999–15	Projected Postsecondary Outcomes of 1992 High School Graduates	Aurora D'Amico
2001–04	Beginning Postsecondary Students Longitudinal Study: 1996–2001 (BPS:1996/2001) Field Test Methodology Report	Paula Knepper
Common Core of Data (CCD)		
95–12	Rural Education Data User's Guide	Samuel Peng
96–19	Assessment and Analysis of School-Level Expenditures	William J. Fowler, Jr.
97–15	Customer Service Survey: Common Core of Data Coordinators	Lee Hoffman
97–43	Measuring Inflation in Public School Costs	William J. Fowler, Jr.
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
1999–03	Evaluation of the 1996–97 Nonfiscal Common Core of Data Surveys Data Collection, Processing, and Editing Cycle	Beth Young
2000–12	Coverage Evaluation of the 1994–95 Common Core of Data: Public Elementary/Secondary School Universe Survey	Beth Young
2000–13	Non-professional Staff in the Schools and Staffing Survey (SASS) and Common Core of Data (CCD)	Kerry Gruber
2001–09	An Assessment of the Accuracy of CCD Data: A Comparison of 1988, 1989, and 1990 CCD Data with 1990–91 SASS Data	John Sietsema
2001–14	Evaluation of the Common Core of Data (CCD) Finance Data Imputations	Frank Johnson
Data Development		
2000–16a	Lifelong Learning NCES Task Force: Final Report Volume I	Lisa Hudson
2000–16b	Lifelong Learning NCES Task Force: Final Report Volume II	Lisa Hudson
Decennial Census School District Project		
95–12	Rural Education Data User's Guide	Samuel Peng
96–04	Census Mapping Project/School District Data Book	Tai Phan
98–07	Decennial Census School District Project Planning Report	Tai Phan
2001–12	Customer Feedback on the 1990 Census Mapping Project	Dan Kasprzyk
Early Childhood Longitudinal Study (ECLS)		
96–08	How Accurate are Teacher Judgments of Students' Academic Performance?	Jerry West
96–18	Assessment of Social Competence, Adaptive Behaviors, and Approaches to Learning with Young Children	Jerry West
97–24	Formulating a Design for the ECLS: A Review of Longitudinal Studies	Jerry West
97–36	Measuring the Quality of Program Environments in Head Start and Other Early Childhood Programs: A Review and Recommendations for Future Research	Jerry West
1999–01	A Birth Cohort Study: Conceptual and Design Considerations and Rationale	Jerry West

No.	Title	NCES contact
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
2001–02	Measuring Father Involvement in Young Children's Lives: Recommendations for a Fatherhood Module for the ECLS-B	Jerry West
2001–03	Measures of Socio-Emotional Development in Middle Childhood	Elvira Hausken
2001–06	Papers from the Early Childhood Longitudinal Studies Program: Presented at the 2001 AERA and SRCD Meetings	Jerry West
Education Finance Statistics Center (EDFIN)		
94–05	Cost-of-Education Differentials Across the States	William J. Fowler, Jr.
96–19	Assessment and Analysis of School-Level Expenditures	William J. Fowler, Jr.
97–43	Measuring Inflation in Public School Costs	William J. Fowler, Jr.
98–04	Geographic Variations in Public Schools' Costs	William J. Fowler, Jr.
1999–16	Measuring Resources in Education: From Accounting to the Resource Cost Model Approach	William J. Fowler, Jr.
High School and Beyond (HS&B)		
95–12	Rural Education Data User's Guide	Samuel Peng
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
HS Transcript Studies		
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
International Adult Literacy Survey (IALS)		
97–33	Adult Literacy: An International Perspective	Marilyn Binkley
Integrated Postsecondary Education Data System (IPEDS)		
97–27	Pilot Test of IPEDS Finance Survey	Peter Stowe
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
2000–14	IPEDS Finance Data Comparisons Under the 1997 Financial Accounting Standards for Private, Not-for-Profit Institutes: A Concept Paper	Peter Stowe
National Assessment of Adult Literacy (NAAL)		
98–17	Developing the National Assessment of Adult Literacy: Recommendations from Stakeholders	Sheida White
1999–09a	1992 National Adult Literacy Survey: An Overview	Alex Sedlacek
1999–09b	1992 National Adult Literacy Survey: Sample Design	Alex Sedlacek
1999–09c	1992 National Adult Literacy Survey: Weighting and Population Estimates	Alex Sedlacek
1999–09d	1992 National Adult Literacy Survey: Development of the Survey Instruments	Alex Sedlacek
1999–09e	1992 National Adult Literacy Survey: Scaling and Proficiency Estimates	Alex Sedlacek
1999–09f	1992 National Adult Literacy Survey: Interpreting the Adult Literacy Scales and Literacy Levels	Alex Sedlacek
1999–09g	1992 National Adult Literacy Survey: Literacy Levels and the Response Probability Convention	Alex Sedlacek
2000–05	Secondary Statistical Modeling With the National Assessment of Adult Literacy: Implications for the Design of the Background Questionnaire	Sheida White
2000–06	Using Telephone and Mail Surveys as a Supplement or Alternative to Door-to-Door Surveys in the Assessment of Adult Literacy	Sheida White
2000–07	"How Much Literacy is Enough?" Issues in Defining and Reporting Performance Standards for the National Assessment of Adult Literacy	Sheida White
2000–08	Evaluation of the 1992 NALS Background Survey Questionnaire: An Analysis of Uses with Recommendations for Revisions	Sheida White
2000–09	Demographic Changes and Literacy Development in a Decade	Sheida White
2001–08	Assessing the Lexile Framework: Results of a Panel Meeting	Sheida White

No.	Title	NCES contact
National Assessment of Educational Progress (NAEP)		
95–12	Rural Education Data User's Guide	Samuel Peng
97–29	Can State Assessment Data be Used to Reduce State NAEP Sample Sizes?	Steven Gorman
97–30	ACT's NAEP Redesign Project: Assessment Design is the Key to Useful and Stable Assessment Results	Steven Gorman
97–31	NAEP Reconfigured: An Integrated Redesign of the National Assessment of Educational Progress	Steven Gorman
97–32	Innovative Solutions to Intractable Large Scale Assessment (Problem 2: Background Questionnaires)	Steven Gorman
97–37	Optimal Rating Procedures and Methodology for NAEP Open-ended Items	Steven Gorman
97–44	Development of a SASS 1993–94 School-Level Student Achievement Subfile: Using State Assessments and State NAEP, Feasibility Study	Michael Ross
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
2001–07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein
2001–08	Assessing the Lexile Framework: Results of a Panel Meeting	Sheida White
2001–11	Impact of Selected Background Variables on Students' NAEP Math Performance	Arnold Goldstein
2001–13	The Effects of Accommodations on the Assessment of LEP Students in NAEP	Arnold Goldstein
National Education Longitudinal Study of 1988 (NELS:88)		
95–04	National Education Longitudinal Study of 1988: Second Follow-up Questionnaire Content Areas and Research Issues	Jeffrey Owings
95–05	National Education Longitudinal Study of 1988: Conducting Trend Analyses of NLS-72, HS&B, and NELS:88 Seniors	Jeffrey Owings
95–06	National Education Longitudinal Study of 1988: Conducting Cross-Cohort Comparisons Using HS&B, NAEP, and NELS:88 Academic Transcript Data	Jeffrey Owings
95–07	National Education Longitudinal Study of 1988: Conducting Trend Analyses HS&B and NELS:88 Sophomore Cohort Dropouts	Jeffrey Owings
95–12	Rural Education Data User's Guide	Samuel Peng
95–14	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
96–03	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
98–06	National Education Longitudinal Study of 1988 (NELS:88) Base Year through Second Follow-Up: Final Methodology Report	Ralph Lee
98–09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
1999–15	Projected Postsecondary Outcomes of 1992 High School Graduates	Aurora D'Amico
2001–16	Imputation of Test Scores in the National Education Longitudinal Study of 1988	Ralph Lee
National Household Education Survey (NHES)		
95–12	Rural Education Data User's Guide	Samuel Peng
96–13	Estimation of Response Bias in the NHES:95 Adult Education Survey	Steven Kaufman
96–14	The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component	Steven Kaufman
96–20	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler
96–21	1993 National Household Education Survey (NHES:93) Questionnaires: Screener, School Readiness, and School Safety and Discipline	Kathryn Chandler
96–22	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler

No.	Title	NCES contact
96–29	Undercoverage Bias in Estimates of Characteristics of Adults and 0- to 2-Year-Olds in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
96–30	Comparison of Estimates from the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97–02	Telephone Coverage Bias and Recorded Interviews in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97–03	1991 and 1995 National Household Education Survey Questionnaires: NHES:91 Screener, NHES:91 Adult Education, NHES:95 Basic Screener, and NHES:95 Adult Education	Kathryn Chandler
97–04	Design, Data Collection, Monitoring, Interview Administration Time, and Data Editing in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97–05	Unit and Item Response, Weighting, and Imputation Procedures in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97–06	Unit and Item Response, Weighting, and Imputation Procedures in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97–08	Design, Data Collection, Interview Timing, and Data Editing in the 1995 National Household Education Survey	Kathryn Chandler
97–19	National Household Education Survey of 1995: Adult Education Course Coding Manual	Peter Stowe
97–20	National Household Education Survey of 1995: Adult Education Course Code Merge Files User's Guide	Peter Stowe
97–25	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
97–28	Comparison of Estimates in the 1996 National Household Education Survey	Kathryn Chandler
97–34	Comparison of Estimates from the 1993 National Household Education Survey	Kathryn Chandler
97–35	Design, Data Collection, Interview Administration Time, and Data Editing in the 1996 National Household Education Survey	Kathryn Chandler
97–38	Reinterview Results for the Parent and Youth Components of the 1996 National Household Education Survey	Kathryn Chandler
97–39	Undercoverage Bias in Estimates of Characteristics of Households and Adults in the 1996 National Household Education Survey	Kathryn Chandler
97–40	Unit and Item Response Rates, Weighting, and Imputation Procedures in the 1996 National Household Education Survey	Kathryn Chandler
98–03	Adult Education in the 1990s: A Report on the 1991 National Household Education Survey	Peter Stowe
98–10	Adult Education Participation Decisions and Barriers: Review of Conceptual Frameworks and Empirical Studies	Peter Stowe
National Longitudinal Study of the High School Class of 1972 (NLS-72)		
95–12	Rural Education Data User's Guide	Samuel Peng
National Postsecondary Student Aid Study (NPSAS)		
96–17	National Postsecondary Student Aid Study: 1996 Field Test Methodology Report	Andrew G. Malizio
2000–17	National Postsecondary Student Aid Study:2000 Field Test Methodology Report	Andrew G. Malizio
National Study of Postsecondary Faculty (NSOPF)		
97–26	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
2000–01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
Postsecondary Education Descriptive Analysis Reports (PEDAR)		
2000–11	Financial Aid Profile of Graduate Students in Science and Engineering	Aurora D'Amico
Private School Universe Survey (PSS)		
95–16	Intersurvey Consistency in NCES Private School Surveys	Steven Kaufman
95–17	Estimates of Expenditures for Private K–12 Schools	Stephen Broughman
96–16	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
96–26	Improving the Coverage of Private Elementary-Secondary Schools	Steven Kaufman
96–27	Intersurvey Consistency in NCES Private School Surveys for 1993–94	Steven Kaufman
97–07	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman

No.	Title	NCES contact
97–22	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
2000–15	Feasibility Report: School-Level Finance Pretest, Private School Questionnaire	Stephen Broughman
Recent College Graduates (RCG)		
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
Schools and Staffing Survey (SASS)		
94–01	Schools and Staffing Survey (SASS) Papers Presented at Meetings of the American Statistical Association	Dan Kasprzyk
94–02	Generalized Variance Estimate for Schools and Staffing Survey (SASS)	Dan Kasprzyk
94–03	1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report	Dan Kasprzyk
94–04	The Accuracy of Teachers' Self-reports on their Postsecondary Education: Teacher Transcript Study, Schools and Staffing Survey	Dan Kasprzyk
94–06	Six Papers on Teachers from the 1990–91 Schools and Staffing Survey and Other Related Surveys	Dan Kasprzyk
95–01	Schools and Staffing Survey: 1994 Papers Presented at the 1994 Meeting of the American Statistical Association	Dan Kasprzyk
95–02	QED Estimates of the 1990–91 Schools and Staffing Survey: Deriving and Comparing QED School Estimates with CCD Estimates	Dan Kasprzyk
95–03	Schools and Staffing Survey: 1990–91 SASS Cross-Questionnaire Analysis	Dan Kasprzyk
95–08	CCD Adjustment to the 1990–91 SASS: A Comparison of Estimates	Dan Kasprzyk
95–09	The Results of the 1993 Teacher List Validation Study (TLVS)	Dan Kasprzyk
95–10	The Results of the 1991–92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation	Dan Kasprzyk
95–11	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
95–12	Rural Education Data User's Guide	Samuel Peng
95–14	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
95–15	Classroom Instructional Processes: A Review of Existing Measurement Approaches and Their Applicability for the Teacher Follow-up Survey	Sharon Bobbitt
95–16	Intersurvey Consistency in NCES Private School Surveys	Steven Kaufman
95–18	An Agenda for Research on Teachers and Schools: Revisiting NCES' Schools and Staffing Survey	Dan Kasprzyk
96–01	Methodological Issues in the Study of Teachers' Careers: Critical Features of a Truly Longitudinal Study	Dan Kasprzyk
96–02	Schools and Staffing Survey (SASS): 1995 Selected papers presented at the 1995 Meeting of the American Statistical Association	Dan Kasprzyk
96–05	Cognitive Research on the Teacher Listing Form for the Schools and Staffing Survey	Dan Kasprzyk
96–06	The Schools and Staffing Survey (SASS) for 1998–99: Design Recommendations to Inform Broad Education Policy	Dan Kasprzyk
96–07	Should SASS Measure Instructional Processes and Teacher Effectiveness?	Dan Kasprzyk
96–09	Making Data Relevant for Policy Discussions: Redesigning the School Administrator Questionnaire for the 1998–99 SASS	Dan Kasprzyk
96–10	1998–99 Schools and Staffing Survey: Issues Related to Survey Depth	Dan Kasprzyk
96–11	Towards an Organizational Database on America's Schools: A Proposal for the Future of SASS, with comments on School Reform, Governance, and Finance	Dan Kasprzyk
96–12	Predictors of Retention, Transfer, and Attrition of Special and General Education Teachers: Data from the 1989 Teacher Followup Survey	Dan Kasprzyk
96–15	Nested Structures: District-Level Data in the Schools and Staffing Survey	Dan Kasprzyk
96–23	Linking Student Data to SASS: Why, When, How	Dan Kasprzyk
96–24	National Assessments of Teacher Quality	Dan Kasprzyk
96–25	Measures of Inservice Professional Development: Suggested Items for the 1998–1999 Schools and Staffing Survey	Dan Kasprzyk
96–28	Student Learning, Teaching Quality, and Professional Development: Theoretical Linkages, Current Measurement, and Recommendations for Future Data Collection	Mary Rollefson
97–01	Selected Papers on Education Surveys: Papers Presented at the 1996 Meeting of the American Statistical Association	Dan Kasprzyk

No.	Title	NCES contact
97–07	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97–09	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
97–10	Report of Cognitive Research on the Public and Private School Teacher Questionnaires for the Schools and Staffing Survey 1993–94 School Year	Dan Kasprzyk
97–11	International Comparisons of Inservice Professional Development	Dan Kasprzyk
97–12	Measuring School Reform: Recommendations for Future SASS Data Collection	Mary Rollefson
97–14	Optimal Choice of Periodicities for the Schools and Staffing Survey: Modeling and Analysis	Steven Kaufman
97–18	Improving the Mail Return Rates of SASS Surveys: A Review of the Literature	Steven Kaufman
97–22	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman
97–23	Further Cognitive Research on the Schools and Staffing Survey (SASS) Teacher Listing Form	Dan Kasprzyk
97–41	Selected Papers on the Schools and Staffing Survey: Papers Presented at the 1997 Meeting of the American Statistical Association	Steve Kaufman
97–42	Improving the Measurement of Staffing Resources at the School Level: The Development of Recommendations for NCES for the Schools and Staffing Survey (SASS)	Mary Rollefson
97–44	Development of a SASS 1993–94 School-Level Student Achievement Subfile: Using State Assessments and State NAEP, Feasibility Study	Michael Ross
98–01	Collection of Public School Expenditure Data: Development of a Questionnaire	Stephen Broughman
98–02	Response Variance in the 1993–94 Schools and Staffing Survey: A Reinterview Report	Steven Kaufman
98–04	Geographic Variations in Public Schools' Costs	William J. Fowler, Jr.
98–05	SASS Documentation: 1993–94 SASS Student Sampling Problems; Solutions for Determining the Numerators for the SASS Private School (3B) Second-Stage Factors	Steven Kaufman
98–08	The Redesign of the Schools and Staffing Survey for 1999–2000: A Position Paper	Dan Kasprzyk
98–12	A Bootstrap Variance Estimator for Systematic PPS Sampling	Steven Kaufman
98–13	Response Variance in the 1994–95 Teacher Follow-up Survey	Steven Kaufman
98–14	Variance Estimation of Imputed Survey Data	Steven Kaufman
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
98–16	A Feasibility Study of Longitudinal Design for Schools and Staffing Survey	Stephen Broughman
1999–02	Tracking Secondary Use of the Schools and Staffing Survey Data: Preliminary Results	Dan Kasprzyk
1999–04	Measuring Teacher Qualifications	Dan Kasprzyk
1999–07	Collection of Resource and Expenditure Data on the Schools and Staffing Survey	Stephen Broughman
1999–08	Measuring Classroom Instructional Processes: Using Survey and Case Study Fieldtest Results to Improve Item Construction	Dan Kasprzyk
1999–10	What Users Say About Schools and Staffing Survey Publications	Dan Kasprzyk
1999–12	1993–94 Schools and Staffing Survey: Data File User's Manual, Volume III: Public-Use Codebook	Kerry Gruber
1999–13	1993–94 Schools and Staffing Survey: Data File User's Manual, Volume IV: Bureau of Indian Affairs (BIA) Restricted-Use Codebook	Kerry Gruber
1999–14	1994–95 Teacher Followup Survey: Data File User's Manual, Restricted-Use Codebook	Kerry Gruber
1999–17	Secondary Use of the Schools and Staffing Survey Data	Susan Wiley
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
2000–10	A Research Agenda for the 1999–2000 Schools and Staffing Survey	Dan Kasprzyk
2000–13	Non-professional Staff in the Schools and Staffing Survey (SASS) and Common Core of Data (CCD)	Kerry Gruber
2000–18	Feasibility Report: School-Level Finance Pretest, Public School District Questionnaire	Stephen Broughman
Third International Mathematics and Science Study (TIMSS)		
2001–01	Cross-National Variation in Educational Preparation for Adulthood: From Early Adolescence to Young Adulthood	Elvira Hausken
2001–05	Using TIMSS to Analyze Correlates of Performance Variation in Mathematics	Patrick Gonzales
2001–07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein

Listing of NCES Working Papers by Subject

No.	Title	NCES contact
Achievement (student) - mathematics		
2001–05	Using TIMSS to Analyze Correlates of Performance Variation in Mathematics	Patrick Gonzales
Adult education		
96–14	The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component	Steven Kaufman
96–20	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler
96–22	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler
98–03	Adult Education in the 1990s: A Report on the 1991 National Household Education Survey	Peter Stowe
98–10	Adult Education Participation Decisions and Barriers: Review of Conceptual Frameworks and Empirical Studies	Peter Stowe
1999–11	Data Sources on Lifelong Learning Available from the National Center for Education Statistics	Lisa Hudson
2000–16a	Lifelong Learning NCES Task Force: Final Report Volume I	Lisa Hudson
2000–16b	Lifelong Learning NCES Task Force: Final Report Volume II	Lisa Hudson
Adult literacy—see Literacy of adults		
American Indian – education		
1999–13	1993–94 Schools and Staffing Survey: Data File User's Manual, Volume IV: Bureau of Indian Affairs (BIA) Restricted-Use Codebook	Kerry Gruber
Assessment/achievement		
95–12	Rural Education Data User's Guide	Samuel Peng
95–13	Assessing Students with Disabilities and Limited English Proficiency	James Houser
97–29	Can State Assessment Data be Used to Reduce State NAEP Sample Sizes?	Larry Ogle
97–30	ACT's NAEP Redesign Project: Assessment Design is the Key to Useful and Stable Assessment Results	Larry Ogle
97–31	NAEP Reconfigured: An Integrated Redesign of the National Assessment of Educational Progress	Larry Ogle
97–32	Innovative Solutions to Intractable Large Scale Assessment (Problem 2: Background Questions)	Larry Ogle
97–37	Optimal Rating Procedures and Methodology for NAEP Open-ended Items	Larry Ogle
97–44	Development of a SASS 1993–94 School-Level Student Achievement Subfile: Using State Assessments and State NAEP, Feasibility Study	Michael Ross
98–09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
2001–07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein
2001–11	Impact of Selected Background Variables on Students' NAEP Math Performance	Arnold Goldstein
2001–13	The Effects of Accommodations on the Assessment of LEP Students in NAEP	Arnold Goldstein
Beginning students in postsecondary education		
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
2001–04	Beginning Postsecondary Students Longitudinal Study: 1996–2001 (BPS:1996/2001) Field Test Methodology Report	Paula Knepper

No.	Title	NCES contact
Civic participation		
97–25	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
Climate of schools		
95–14	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
Cost of education indices		
94–05	Cost-of-Education Differentials Across the States	William J. Fowler, Jr.
Course-taking		
95–12	Rural Education Data User's Guide	Samuel Peng
98–09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
Crime		
97–09	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
Curriculum		
95–11	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
98–09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
Customer service		
1999–10	What Users Say About Schools and Staffing Survey Publications	Dan Kasprzyk
2000–02	Coordinating NCES Surveys: Options, Issues, Challenges, and Next Steps	Valena Plisko
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
2001–12	Customer Feedback on the 1990 Census Mapping Project	Dan Kasprzyk
Data quality		
97–13	Improving Data Quality in NCES: Database-to-Report Process	Susan Ahmed
2001–11	Impact of Selected Background Variables on Students' NAEP Math Performance	Arnold Goldstein
2001–13	The Effects of Accommodations on the Assessment of LEP Students in NAEP	Arnold Goldstein
Data warehouse		
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
Design effects		
2000–03	Strengths and Limitations of Using SUDAAN, Stata, and WesVarPC for Computing Variances from NCES Data Sets	Ralph Lee
Dropout rates, high school		
95–07	National Education Longitudinal Study of 1988: Conducting Trend Analyses HS&B and NELS:88 Sophomore Cohort Dropouts	Jeffrey Owings
Early childhood education		
96–20	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler

No.	Title	NCES contact
96–22	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler
97–24	Formulating a Design for the ECLS: A Review of Longitudinal Studies	Jerry West
97–36	Measuring the Quality of Program Environments in Head Start and Other Early Childhood Programs: A Review and Recommendations for Future Research	Jerry West
1999–01	A Birth Cohort Study: Conceptual and Design Considerations and Rationale	Jerry West
2001–02	Measuring Father Involvement in Young Children's Lives: Recommendations for a Fatherhood Module for the ECLS-B	Jerry West
2001–03	Measures of Socio-Emotional Development in Middle School	Elvira Hausken
2001–06	Papers from the Early Childhood Longitudinal Studies Program: Presented at the 2001 AERA and SRCD Meetings	Jerry West
Educational attainment		
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
2001–15	Baccalaureate and Beyond Longitudinal Study: 2000/01 Follow-Up Field Test Methodology Report	Andrew G. Malizio
Educational research		
2000–02	Coordinating NCES Surveys: Options, Issues, Challenges, and Next Steps	Valena Plisko
Eighth-graders		
2001–05	Using TIMSS to Analyze Correlates of Performance Variation in Mathematics	Patrick Gonzales
Employment		
96–03	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
2000–16a	Lifelong Learning NCES Task Force: Final Report Volume I	Lisa Hudson
2000–16b	Lifelong Learning NCES Task Force: Final Report Volume II	Lisa Hudson
2001–01	Cross-National Variation in Educational Preparation for Adulthood: From Early Adolescence to Young Adulthood	Elvira Hausken
Employment – after college		
2001–15	Baccalaureate and Beyond Longitudinal Study: 2000/01 Follow-Up Field Test Methodology Report	Andrew G. Malizio
Engineering		
2000–11	Financial Aid Profile of Graduate Students in Science and Engineering	Aurora D'Amico
Enrollment – after college		
2001–15	Baccalaureate and Beyond Longitudinal Study: 2000/01 Follow-Up Field Test Methodology Report	Andrew G. Malizio
Faculty – higher education		
97–26	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
2000–01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
Fathers – role in education		
2001–02	Measuring Father Involvement in Young Children's Lives: Recommendations for a Fatherhood Module for the ECLS-B	Jerry West
Finance – elementary and secondary schools		
94–05	Cost-of-Education Differentials Across the States	William J. Fowler, Jr.
96–19	Assessment and Analysis of School-Level Expenditures	William J. Fowler, Jr.
98–01	Collection of Public School Expenditure Data: Development of a Questionnaire	Stephen Broughman
1999–07	Collection of Resource and Expenditure Data on the Schools and Staffing Survey	Stephen Broughman

No.	Title	NCES contact
1999–16	Measuring Resources in Education: From Accounting to the Resource Cost Model Approach	William J. Fowler, Jr.
2000–18	Feasibility Report: School-Level Finance Pretest, Public School District Questionnaire	Stephen Broughman
2001–14	Evaluation of the Common Core of Data (CCD) Finance Data Imputations	Frank Johnson
Finance – postsecondary		
97–27	Pilot Test of IPEDS Finance Survey	Peter Stowe
2000–14	IPEDS Finance Data Comparisons Under the 1997 Financial Accounting Standards for Private, Not-for-Profit Institutes: A Concept Paper	Peter Stowe
Finance – private schools		
95–17	Estimates of Expenditures for Private K–12 Schools	Stephen Broughman
96–16	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
97–07	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97–22	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman
1999–07	Collection of Resource and Expenditure Data on the Schools and Staffing Survey	Stephen Broughman
2000–15	Feasibility Report: School-Level Finance Pretest, Private School Questionnaire	Stephen Broughman
Geography		
98–04	Geographic Variations in Public Schools' Costs	William J. Fowler, Jr.
Graduate students		
2000–11	Financial Aid Profile of Graduate Students in Science and Engineering	Aurora D'Amico
Graduates of postsecondary education		
2001–15	Baccalaureate and Beyond Longitudinal Study: 2000/01 Follow-Up Field Test Methodology Report	Andrew G. Malizio
Imputation		
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meeting	Dan Kasprzyk
2001–10	Comparison of Proc Impute and Schafer's Multiple Imputation Software	Sam Peng
2001–14	Evaluation of the Common Core of Data (CCD) Finance Data Imputations	Frank Johnson
2001–16	Imputation of Test Scores in the National Education Longitudinal Study of 1988	Ralph Lee
2001–17	A Study of Imputation Algorithms	Ralph Lee
2001–18	A Study of Variance Estimation Methods	Ralph Lee
Inflation		
97–43	Measuring Inflation in Public School Costs	William J. Fowler, Jr.
Institution data		
2000–01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
Instructional resources and practices		
95–11	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
1999–08	Measuring Classroom Instructional Processes: Using Survey and Case Study Field Test Results to Improve Item Construction	Dan Kasprzyk
International comparisons		
97–11	International Comparisons of Inservice Professional Development	Dan Kasprzyk
97–16	International Education Expenditure Comparability Study: Final Report, Volume I	Shelley Burns
97–17	International Education Expenditure Comparability Study: Final Report, Volume II, Quantitative Analysis of Expenditure Comparability	Shelley Burns
2001–01	Cross-National Variation in Educational Preparation for Adulthood: From Early Adolescence to Young Adulthood	Elvira Hausken

No.	Title	NCES contact
2001–07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein
International comparisons – math and science achievement		
2001–05	Using TIMSS to Analyze Correlates of Performance Variation in Mathematics	Patrick Gonzales
Libraries		
94–07	Data Comparability and Public Policy: New Interest in Public Library Data Papers Presented at Meetings of the American Statistical Association	Carrol Kindel
97–25	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
Limited English Proficiency		
95–13	Assessing Students with Disabilities and Limited English Proficiency	James Houser
2001–11	Impact of Selected Background Variables on Students' NAEP Math Performance	Arnold Goldstein
2001–13	The Effects of Accommodations on the Assessment of LEP Students in NAEP	Arnold Goldstein
Literacy of adults		
98–17	Developing the National Assessment of Adult Literacy: Recommendations from Stakeholders	Sheida White
1999–09a	1992 National Adult Literacy Survey: An Overview	Alex Sedlacek
1999–09b	1992 National Adult Literacy Survey: Sample Design	Alex Sedlacek
1999–09c	1992 National Adult Literacy Survey: Weighting and Population Estimates	Alex Sedlacek
1999–09d	1992 National Adult Literacy Survey: Development of the Survey Instruments	Alex Sedlacek
1999–09e	1992 National Adult Literacy Survey: Scaling and Proficiency Estimates	Alex Sedlacek
1999–09f	1992 National Adult Literacy Survey: Interpreting the Adult Literacy Scales and Literacy Levels	Alex Sedlacek
1999–09g	1992 National Adult Literacy Survey: Literacy Levels and the Response Probability Convention	Alex Sedlacek
1999–11	Data Sources on Lifelong Learning Available from the National Center for Education Statistics	Lisa Hudson
2000–05	Secondary Statistical Modeling With the National Assessment of Adult Literacy: Implications for the Design of the Background Questionnaire	Sheida White
2000–06	Using Telephone and Mail Surveys as a Supplement or Alternative to Door-to-Door Surveys in the Assessment of Adult Literacy	Sheida White
2000–07	"How Much Literacy is Enough?" Issues in Defining and Reporting Performance Standards for the National Assessment of Adult Literacy	Sheida White
2000–08	Evaluation of the 1992 NALS Background Survey Questionnaire: An Analysis of Uses with Recommendations for Revisions	Sheida White
2000–09	Demographic Changes and Literacy Development in a Decade	Sheida White
2001–08	Assessing the Lexile Framework: Results of a Panel Meeting	Sheida White
Literacy of adults – international		
97–33	Adult Literacy: An International Perspective	Marilyn Binkley
Mathematics		
98–09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
1999–08	Measuring Classroom Instructional Processes: Using Survey and Case Study Field Test Results to Improve Item Construction	Dan Kasprzyk
2001–05	Using TIMSS to Analyze Correlates of Performance Variation in Mathematics	Patrick Gonzales
2001–07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein
2001–11	Impact of Selected Background Variables on Students' NAEP Math Performance	Arnold Goldstein

No.	Title	NCES contact
Parental involvement in education		
96–03	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
97–25	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
1999–01	A Birth Cohort Study: Conceptual and Design Considerations and Rationale	Jerry West
2001–06	Papers from the Early Childhood Longitudinal Studies Program: Presented at the 2001 AERA and SRCD Meetings	Jerry West
Participation rates		
98–10	Adult Education Participation Decisions and Barriers: Review of Conceptual Frameworks and Empirical Studies	Peter Stowe
Postsecondary education		
1999–11	Data Sources on Lifelong Learning Available from the National Center for Education Statistics	Lisa Hudson
2000–16a	Lifelong Learning NCES Task Force: Final Report Volume I	Lisa Hudson
2000–16b	Lifelong Learning NCES Task Force: Final Report Volume II	Lisa Hudson
Postsecondary education – persistence and attainment		
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
1999–15	Projected Postsecondary Outcomes of 1992 High School Graduates	Aurora D'Amico
Postsecondary education – staff		
97–26	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
2000–01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
Principals		
2000–10	A Research Agenda for the 1999–2000 Schools and Staffing Survey	Dan Kasprzyk
Private schools		
96–16	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
97–07	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97–22	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman
2000–13	Non-professional Staff in the Schools and Staffing Survey (SASS) and Common Core of Data (CCD)	Kerry Gruber
2000–15	Feasibility Report: School-Level Finance Pretest, Private School Questionnaire	Stephen Broughman
Projections of education statistics		
1999–15	Projected Postsecondary Outcomes of 1992 High School Graduates	Aurora D'Amico
Public school finance		
1999–16	Measuring Resources in Education: From Accounting to the Resource Cost Model Approach	William J. Fowler, Jr.
2000–18	Feasibility Report: School-Level Finance Pretest, Public School District Questionnaire	Stephen Broughman
Public schools		
97–43	Measuring Inflation in Public School Costs	William J. Fowler, Jr.
98–01	Collection of Public School Expenditure Data: Development of a Questionnaire	Stephen Broughman
98–04	Geographic Variations in Public Schools' Costs	William J. Fowler, Jr.
1999–02	Tracking Secondary Use of the Schools and Staffing Survey Data: Preliminary Results	Dan Kasprzyk
2000–12	Coverage Evaluation of the 1994–95 Public Elementary/Secondary School Universe Survey	Beth Young
2000–13	Non-professional Staff in the Schools and Staffing Survey (SASS) and Common Core of Data (CCD)	Kerry Gruber

No.	Title	NCES contact
State		
1999–03	Evaluation of the 1996–97 Nonfiscal Common Core of Data Surveys Data Collection, Processing, and Editing Cycle	Beth Young
Statistical methodology		
97–21	Statistics for Policymakers or Everything You Wanted to Know About Statistics But Thought You Could Never Understand	Susan Ahmed
Statistical standards and methodology		
2001–05	Using TIMSS to Analyze Correlates of Performance Variation in Mathematics	Patrick Gonzales
Students with disabilities		
95–13	Assessing Students with Disabilities and Limited English Proficiency	James Houser
2001–13	The Effects of Accommodations on the Assessment of LEP Students in NAEP	Arnold Goldstein
Survey methodology		
96–17	National Postsecondary Student Aid Study: 1996 Field Test Methodology Report	Andrew G. Malizio
97–15	Customer Service Survey: Common Core of Data Coordinators	Lee Hoffman
97–35	Design, Data Collection, Interview Administration Time, and Data Editing in the 1996 National Household Education Survey	Kathryn Chandler
98–06	National Education Longitudinal Study of 1988 (NELS:88) Base Year through Second Follow-Up: Final Methodology Report	Ralph Lee
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
98–16	A Feasibility Study of Longitudinal Design for Schools and Staffing Survey	Stephen Broughman
1999–07	Collection of Resource and Expenditure Data on the Schools and Staffing Survey	Stephen Broughman
1999–17	Secondary Use of the Schools and Staffing Survey Data	Susan Wiley
2000–01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
2000–02	Coordinating NCES Surveys: Options, Issues, Challenges, and Next Steps	Valena Plisko
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
2000–12	Coverage Evaluation of the 1994–95 Public Elementary/Secondary School Universe Survey	Beth Young
2000–17	National Postsecondary Student Aid Study:2000 Field Test Methodology Report	Andrew G. Malizio
2001–04	Beginning Postsecondary Students Longitudinal Study: 1996–2001 (BPS:1996/2001) Field Test Methodology Report	Paula Knepper
2001–07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein
2001–09	An Assessment of the Accuracy of CCD Data: A Comparison of 1988, 1989, and 1990 CCD Data with 1990–91 SASS Data	John Sietsema
2001–11	Impact of Selected Background Variables on Students' NAEP Math Performance	Arnold Goldstein
2001–13	The Effects of Accommodations on the Assessment of LEP Students in NAEP	Arnold Goldstein
Teachers		
98–13	Response Variance in the 1994–95 Teacher Follow-up Survey	Steven Kaufman
1999–14	1994–95 Teacher Followup Survey: Data File User's Manual, Restricted-Use Codebook	Kerry Gruber
2000–10	A Research Agenda for the 1999–2000 Schools and Staffing Survey	Dan Kasprzyk
Teachers – instructional practices of		
98–08	The Redesign of the Schools and Staffing Survey for 1999–2000: A Position Paper	Dan Kasprzyk
Teachers – opinions regarding safety		
98–08	The Redesign of the Schools and Staffing Survey for 1999–2000: A Position Paper	Dan Kasprzyk
Teachers – performance evaluations		
1999–04	Measuring Teacher Qualifications	Dan Kasprzyk

No.	Title	NCES contact
Teachers – qualifications of		
1999–04	Measuring Teacher Qualifications	Dan Kasprzyk
Teachers – salaries of		
94–05	Cost-of-Education Differentials Across the States	William J. Fowler, Jr.
Training		
2000–16a	Lifelong Learning NCES Task Force: Final Report Volume I	Lisa Hudson
2000–16b	Lifelong Learning NCES Task Force: Final Report Volume II	Lisa Hudson
Variance estimation		
2000–03	Strengths and Limitations of Using SUDAAN, Stata, and WesVarPC for Computing Variances from NCES Data Sets	Ralph Lee
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
2001–18	A Study of Variance Estimation Methods	Ralph Lee
Violence		
97–09	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
Vocational education		
95–12	Rural Education Data User's Guide	Samuel Peng
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson