Measuring the Power of Learning.®

Research Report
ETS RR–17-24

Applying the Hájek Approach in Formula-Based Variance Estimation

Jiahe Qian

December 2017

ETS Research Report Series

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Applying the Hájek Approach in Formula-Based Variance Estimation

Jiahe Qian

Educational Testing Service, Princeton, NJ

The variance formula derived for a two-stage sampling design without replacement employs the joint inclusion probabilities in the first-stage selection of clusters. One of the difficulties encountered in data analysis is the lack of information about such joint inclusion probabilities. One way to solve this issue is by applying Hájek's approximation of the joint probabilities in variance estimation. To assess the Hájek approach, several estimators of Hájek's *c* and *d* are proposed. The application is illustrated with simulation and real data. A Monte Carlo simulation is employed to compare the results of joint inclusion probabilities yielded from the probability-proportional-to-size sampling methods with the results from Hájek's approximation. Empirically estimated variances from the jackknife procedure are also compared with the formula-based variances with incorporated Hájek's approximation.

Keywords Two-stage PPS sampling without replacement; Horvitz–Thompson estimator; joint inclusion probability; simulation; Brewer's method of cluster sampling; Durbin's method of cluster sampling

doi:10.1002/ets2.12154

In assessment surveys, unequal probability sampling without replacement (Cochran, 1977; Kish, 1965) is used to draw samples. One such type of sampling approach is two-stage probability-proportional-to-size (PPS) sampling,[1] with a PPS selection of schools and a simple random sampling (SRS) of students. For example, two-stage PPS sampling has been applied in the National Assessment of Educational Progress (NAEP; Allen, Donoghue, & Schoeps, 2001; Rust, 1985) and in the Programme for International Student Assessment (PISA; Nohara, 2001; Turner & Adams, 2007).

For a two-stage PPS sampling without replacement, the Horvitz–Thompson (H–T; Horvitz & Thompson, 1952) estimator is often used to estimate the population total, and its variance formula (Cochran, 1977, p. 260; Wolter, 2007) can also be derived. For a nonlinear function of estimates, such as the ratio estimator of two means, the delta method (Cochran, 1977, p. 154; C. R. Rao, 1973, pp. 385–389), based on a Taylor series approximation, can be used to derive the formulae used in variance estimation. For stratified complex sampling, Woodruff (1971) applied the delta method to approximate the variance of a complicated estimate.

One condition for applying the variance formula of the H–T estimators involves having information about the joint probabilities of all the possible pairs of sample units that are included during the first stage of a two-stage PPS sampling process. However, joint inclusion probabilities are usually not available for survey data; therefore it is hard to apply a formula to estimate variances for data drawn by PPS sampling without replacement.

To address this issue, Hájek (1964) explored the properties of joint inclusion probabilities and derived a formula based on rejective sampling, a sampling procedure in which a Poisson sample is rejected unless it contains exactly *n* sample units as required by the sample design (Fuller, 2009; Hájek, 1981, p. 66). Rejective sampling is also called *conditional Poisson sampling*. Accordingly, some researchers (Lohr, 1999; Qian, 2015; Särndal & Lundström, 2005) have recommended using Hájek's approximation of the joint inclusion probability in formula-based variance estimation with data drawn by PPS sampling without replacement and derived corresponding formulae. However, few studies have evaluated such applications based on real and/or simulation data; thus issues still exist in applying Hájek's approximation to formula-based variance estimation.

Moreover, Berger (2003, 2004) proposed several adjusted Hájek-based variance estimators and applied the Hájek approach to weighted least squares regression. Based on Monte Carlo studies, Fuller (2009) discussed some design properties of a rejective sampling procedure. Särndal, Swenson, and Wretman (1992) used the Hájek approximation in

Corresponding author: J. Qian, E-mail: jqian@ets.org

model-assisted survey sampling. Rizzo and Rust (2011) developed an approximated estimation of the joint inclusion probabilities in variance estimation and applied it to analyzing NAEP samples (Kali, Burke, Hicks, Rizzo, & Rust, 2011; Qian, 2015).

The goal of this study is to assess the appropriateness of applying Hájek's approximation to variance estimation with complex data, both simulation and real, drawn by PPS sampling without replacement. The study is focused on two tasks: first, to conduct simulations to check the accuracy of the approximation of the joint inclusion probabilities estimated by Hájek's approximation, and second, based on real survey data, to compare the variances estimated by the formula with Hájek's approximation with those estimated by a grouped jackknife approach. The real data sets used are the NAEP state assessment samples drawn by two-stage systematic PPS sampling.

In the next section, the methodologies applied are reviewed, including the two-stage PPS sampling without replacement, the H−T estimator, the Hájek approximation of joint probability, and the sampling approaches employed in the simulation. In the "Results" section, the simulation results of the joint probabilities are used to examine the goodness of Hájek's approximation. The jackknifed variance estimates are compared with the variances estimated from the formula-based method using Hájek approximation. The final section offers a summary and conclusions.

Methodology

Probability Sampling Without Replacement

In sampling, and for assessment surveys in particular, sample designers prefer using a two-stage design with PPS selection of schools and SRS selection of students, for example, as with the NAEP state assessments.

When drawing a probability sample without replacement, each selection will modify the chances of other cases to be selected; that is, the probability of making a given selection is no longer independent from the others. A properly designed, unequal probability sampling without replacement, of sample size n from a population of size N, will guarantee that each unit in the population has the designated inclusion probability to be drawn in an n -step procedure of selection. Let π_i be the inclusion probability of sample unit i (=1, 2, …, N) and let π_{ij} be the joint inclusion probability of sample units i and j (i and j = 1, 2, …, N), that is, the chance that both units i and j are included in the sample through the drawing of a sample. A properly implemented unequal probability sampling without replacement satisfies the following properties:

$$\sum_{i=1}^{N} \pi_i = n,$$

$$\sum_{j\neq i}^{N} \pi_{ij} = (n-1)\,\pi_i, \text{ and}$$

$$\sum_{i=1}^{N} \sum_{j<i} \pi_{ij} = \frac{1}{2} n\,(n-1)$$

(Cochran, 1977, p. 259).

Horvitz–Thompson Estimators for Survey Samples

Horvitz–Thompson Estimators

For a two-stage PPS sampling design, let y_{ik} be the value of a variable of interest for student k in school i. Assume that the population consists of N schools and, for the first stage of sampling, that the sample size of schools is n. Let M_i be the total number of students in school i; let m_i be the sample size of students drawn from school i. Let $\widetilde{Y}_i = \sum_{j=1}^{M_i} y_{ij}$ be the school total in school i. The statistic of interest is population total, $\widetilde{Y} = \sum_{i=1}^{N} \widetilde{Y}_i$, a sum of all school totals.

Let π_i be the inclusion probability of school i. The school weight for school i (=1, 2, …, n) equals the inverse of π_i; that is, $w_i = \pi_i^{-1}$ (Allen et al., 2001; Rust & Johnson, 1992). Let $\pi_{k|i}$ be the conditional inclusion probability for student k in school i. The conditional case weight for student k within school i is $w_{k|i} = \pi_{k|i}^{-1}$. The case weights for student k in school i equal $w_{ik} = \pi_i^{-1}\pi_{k|i}^{-1}$ (i = 1, 2, …, n and k = 1, 2, …, m_i); after being created, the case weights are often also subject to data adjustments such as poststratification and raking (Allen et al., 2001; Rust, Bethel, Burke, & Hansen, 1990). The total for school i, $\tilde{Y}_i$, can be estimated by

$$\widetilde{y}_i = \sum_{k=1}^{m_i} w_{k|i} y_{ik}.$$

Note that the symbol $\tilde{y}_i$ is not for the sample total $\sum_{k=1}^{m_i} y_{ik}$. The H – T estimators (Cochran, 1977) for the total is de fined as

$$\widetilde{y}_{\mathrm{HT}} = \sum_{i=1}^{n} w_i \widetilde{y}_i, \tag{1}$$

which is an unbiased estimator of the population total $\tilde{Y}$.

Variance of a Horvitz–Thompson Estimator of the Total

Let π_{ij} be the joint inclusion probabilities of schools i and j. The variance of $\tilde{Y}_{\mathrm{HT}}$ is

$$V\left(\widetilde{y}_{\mathrm{HT}}\right) = \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\pi_i\pi_j - \pi_{ij}\right)\left(\frac{\widetilde{Y}_i}{\pi_i} - \frac{\widetilde{Y}_j}{\pi_j}\right)^2 + \sum_{i=1}^{N}\frac{M_i^2\left(1 - f_{2i}\right)}{m_i\pi_i}S_{2i}^2 \tag{2}$$

(Cochran, 1977, p. 301; Lohr, 1999, p. 245), where the squared standard deviation (SD^2) $S_{2i}^2 = \frac{1}{M_i - 1}\sum_{k=1}^{M_i}\left(y_{ik} - \overline{Y}_i.\right)^2$ with $\overline{Y}_i. = M_i^{-1}\sum_{k=1}^{M_i} y_{ik}$ and $f_{2i} = m_i/M_i$. The expression is in the Sen – Yates – Grundy (SYG) form of the variance of $\tilde{Y}_{\mathrm{HT}}$ (Sen, 1953; Yates & Grundy, 1953). The estimate of $V(\tilde{Y}_{\mathrm{HT}})$ is

$$v\left(\widetilde{y}_{\mathrm{HT}}\right) = \sum_{i=1}^{n}\sum_{j>i}^{n}\frac{\left(\pi_i\pi_j - \pi_{ij}\right)}{\pi_{ij}}\left(\frac{\widetilde{y}_i}{\pi_i} - \frac{\widetilde{y}_j}{\pi_j}\right)^2 + \sum_{i=1}^{n}\frac{M_i^2\left(1 - f_{2i}\right)}{m_i\pi_i}s_{2i}^2 \tag{3}$$

(Lohr, 1999, p. 245). In Equation 3, s_{2i}^2 is the estimator of S_{2i}^2; it can be

$$s_{2i}^2 = \frac{m_i}{\left(m_i - 1\right)\sum_{k=1}^{m_i} w_{ik}}\sum_{k=1}^{m_i} w_{ik}\left(y_{ik} - \overline{y}_i.\right)^2,$$

where weighted average $\overline{y}_i. = \sum_{k=1}^{m_i} w_{ik}y_{ik} / \sum_{k=1}^{m_i} w_{ik}$ (Feng, Ni, & Zou, 1998, p. 262). Because of the variability in its values, the term $\pi_i\pi_j - \pi_{ij}$ in the variance estimates of the H – T estimator can be negative. Moreover, the computation of the formula $v(\tilde{y}_{\mathrm{HT}})$ in Equation 3 requires knowledge of the joint inclusion probabilities π_{ij}, that is, an $n \times n$ symmetric matrix of the π_{ij} s. However, these probabilities are often unavailable to data users. This study is intended to solve the issue by applying Hájek's approximation to formula-based variance estimation and then evaluating the approximation via the simulation.

Variance of a Horvitz–Thompson Estimator of the Mean

Let $\widetilde{\mathcal{N}} = \sum_{i=1}^{N} M_i$ be the population size. Thus the population mean is defined as

$$\overline{Y} = \frac{\widetilde{Y}}{\widetilde{\mathcal{N}}}$$

The H – T estimator of the mean,

$$\overline{y}_{\mathrm{HT}} = \frac{\widetilde{y}_{\mathrm{HT}}}{\widetilde{w}} = \frac{\sum_{i=1}^{n} w_i\widetilde{y}_i}{\sum_{i=1}^{n}\widetilde{w}_i},$$

is a ratio estimator of two total estimators, where the H–T estimator of the population size $\widetilde{w} = \sum_{i=1}^{n} \widetilde{w}_i$ and $\widetilde{w}_i = \sum_{k=1}^{m_i} w_{ik}$. Note that the inclusion probability of student k in school i is $w_{ik} = (\pi_i \pi_{k|i})^{-1}$. Although $\overline{y}_{HT}$ is biased (Cochran, 1977, p. 155; Hájek, 1960), the bias vanishes as the sample sizes increase; this bias has order $O(n^{-1})$ and goes to 0 as the sample size n increases.

Because the estimator $\overline{y}_{HT}$ is a nonlinear function, its variance formula can be derived by the delta method (Cochran, 1977; C. R. Rao, 1973, pp. 385–89), an approach based on Taylor approximation. For the H–T estimator of the population size, $E\left(\widetilde{w}\right) = \widetilde{\mathcal{N}}$. When the sample sizes in the two-stage sample are large, the discrepancy between $\widetilde{w}$ and $\widetilde{\mathcal{N}}$ tends to be small, that is, $\widetilde{w} \approx \widetilde{\mathcal{N}}$; the term $\overline{y}_{HT} - \overline{Y}$ approximately equals

$$\overline{y}_{\mathrm{HT}} - \overline{Y} \approx \frac{1}{E\left(\widetilde{w}\right)} \sum_{i=1}^{n} \left(w_i \widetilde{y}_i - \overline{Y} w_i \right) = \frac{1}{E\left(\widetilde{w}\right)} \sum_{i=1}^{n} w_i \left(\widetilde{y}_i - \overline{Y} \right)$$

(Woodruff, 1971). Let $z_i = \widetilde{y}_i - \overline{Y}$ and let $\widetilde{z}_{\mathrm{HT}} = \sum_{i=1}^{n} w_i z_i$. So the variance of $\overline{y}_{\mathrm{HT}}$ can be approximated by

$$V\left(\overline{y}_{\mathrm{HT}}\right) \approx \frac{1}{E\left(\widetilde{w}\right)^2} V\left(\sum_{i=1}^{n} w_i z_i \right) = \frac{V\left(\widetilde{z}_{\mathrm{HT}}\right)}{E\left(\widetilde{w}\right)^2}, \tag{4}$$

where the form of $V\left(\widetilde{z}_{\mathrm{HT}}\right)$ is the same as in Equation 2. The derived formula $V\left(\overline{y}_{\mathrm{HT}}\right)$ in Equation 4 is equivalent to

$$V\left(\overline{y}_{\mathrm{HT}}\right) \approx \overline{Y}^2 \left(\frac{V\left(\widetilde{y}_{\mathrm{HT}}\right)}{\widetilde{Y}^2} - 2 \frac{\mathrm{Cov}\left(\widetilde{y}_{\mathrm{HT}}, \widetilde{w}\right)}{\widetilde{Y}\widetilde{\mathcal{N}}} + \frac{V\left(\widetilde{w}\right)}{\widetilde{\mathcal{N}}^2} \right) \tag{5}$$

(Cochran, 1977, p. 155), which has order $O(n^{-1})$.

The variance $V\left(\overline{y}_{\mathrm{HT}}\right)$ in Equation 4, in the format of a ratio, can be estimated by a ratio estimator of $v\left(\widetilde{z}_{\mathrm{HT}}\right)$ to $\widetilde{w}^2$ (Cochran, 1977, p. 153):

$$v\left(\overline{y}_{\mathrm{HT}}\right) = \frac{1}{\widetilde{w}^2} v\left(\widetilde{z}_{\mathrm{HT}}\right). \tag{6}$$

The term $v\left(\widetilde{z}_{\mathrm{HT}}\right)$ in Equation 6 can be estimated by Equation 3, with $\overline{Y}$ being replaced by $\overline{y}_{\mathrm{HT}}$. Let $\overline{y}_{\mathrm{HT},i} = \overline{y}_{\mathrm{HT}}/m_i$ and let $\widetilde{z}_i = \widetilde{y}_i - \overline{y}_{\mathrm{HT}} = \sum_{k=1}^{m_i} \left(w_{k|i} y_{ik} - \overline{y}_{\mathrm{HT},i} \right)$. Then $v\left(\widetilde{z}_{\mathrm{HT}}\right)$ can be estimated:

$$v\left(\widetilde{z}_{\mathrm{HT}}\right) = \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{\left(\pi_i \pi_j - \pi_{ij}\right)}{\pi_{ij}} \left(\frac{\widetilde{z}_i}{\pi_i} - \frac{\widetilde{z}_j}{\pi_j} \right)^2 + \sum_{i=1}^{n} \frac{M_i^2 \left(1 - f_{2i}\right)}{m_i \pi_i} s_{z,\,2i}^2,$$

where $s_{z,\,2i}^2$ is defined as

$$s_{z,\,2i}^2 = \frac{1}{m_i - 1} \sum_{k=1}^{m_i} \left(w_{k|i} y_{ik} - \overline{y}_{\mathrm{HT},i} \right)^2.$$

Note that the estimator $s_{z,\,2i}^2$ is not unique. Similarly, the variance $V\left(\overline{y}_{\mathrm{HT}}\right)$ in Equation 5 can also be estimated:

$$v\left(\overline{y}_{\mathrm{HT}}\right) = \overline{y}_{\mathrm{HT}}^2 \left(\frac{v\left(\widetilde{y}_{\mathrm{HT}}\right)}{\widetilde{y}_{\mathrm{HT}}^2} - 2 \frac{\mathrm{cov}\left(\widetilde{y}_{\mathrm{HT}},\ \widetilde{w}\right)}{\widetilde{y}_{\mathrm{HT}} \widetilde{w}} + \frac{v\left(\widetilde{w}\right)}{\widetilde{w}^2} \right).$$

Let $s_{w,\,2i}^2 = \frac{1}{m_i - 1} \sum_{k=1}^{m_i} \left(w_{ik} - \overline{w}_i \right)^2$ with $\overline{w}_i = \frac{\widetilde{w}_i}{m_i} = \frac{1}{m_i} \sum_{k=1}^{m_i} w_{ik}$. The variance term $v\left(\widetilde{w}\right)$ can also be estimated by Equation 3:

$$v\left(\widetilde{w}\right) = \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{\left(\pi_i \pi_j - \pi_{ij}\right)}{\pi_{ij}} \left(\frac{\widetilde{w}_i}{\pi_i} - \frac{\widetilde{w}_j}{\pi_j} \right)^2 + \sum_{i=1}^{n} \frac{M_i^2 \left(1 - f_{2i}\right)}{m_i \pi_i} s_{w,\,2i}^2$$

(Kish, 1965, p. 285).

The Hájek Joint Inclusion Probability Approximation

Let $c_i = \pi_i \left(1 - \pi_i\right)$ and $c = \sum_{i=1}^{N} c_i$. Based on rejective sampling, Hájek (1981, p. 75) provided an asymptotically valid approximation of π_{ij}:

$$\widehat{\pi}_{H,ij} \approx \pi_i \pi_j \left(1 - \frac{\left(1 - \pi_i\right)\left(1 - \pi_j\right)}{c} \right). \tag{7}$$

Hájek also showed the following large-sample property:

$$\frac{c\left(\pi_i \pi_j - \pi_{ij}\right)}{\pi_i\left(1 - \pi_i\right)\pi_j\left(1 - \pi_j\right)} \to 1$$

when $n \to \infty$, $(N - n) \to \infty$, and $c \to \infty$ (Hájek, 1964, p. 1496). Under this Hájek setup, the term $\pi_i \pi_j - \widehat{\pi}_{\mathrm{H},ij}$ approximates $\pi_i \pi_j - \pi_{ij}$.

Estimators of Hájek's c

The application of c requires information on all the inclusion probabilities $\pi_1, \pi_2, \ldots$, and π_N; however, most of the assessment data sets only contain $\pi_1, \pi_2, \ldots$, and π_n. Thus the parameter c cannot be computed directly and has to be estimated. One method of estimation uses the H–T estimator:

$$\widehat{c}_1 = \sum_{i=1}^{n} \frac{c_i}{\pi_i} = \sum_{i=1}^{n} \left(1 - \pi_i\right) = n - \sum_{i=1}^{n} \pi_i. \tag{8}$$

For unequal probability sampling without replacement, the estimator $\hat{c}_1$ is an unbiased estimator of c, and the variance of $\hat{c}_1$,

$$V\left(\widehat{c}_1\right) = \sum_{i=1}^{N} \sum_{j>i}^{N} \left(\pi_i \pi_j - \pi_{ij}\right) \left(\frac{c_i}{\pi_i} - \frac{c_j}{\pi_j} \right)^2$$

(Cochran, 1977, p. 260), can be estimated by

$$\begin{aligned} v\left(\widehat{c}_1\right) &= \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{\left(\pi_i \pi_j - \pi_{ij}\right)}{\pi_{ij}} \left(\frac{c_i}{\pi_i} - \frac{c_j}{\pi_j} \right)^2 \\ &= \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{\left(1 - \pi_i\right)\left(1 - \pi_j\right)}{\widehat{c}_1 - \left(1 - \pi_i\right)\left(1 - \pi_j\right)} \left(\pi_i - \pi_j\right)^2 \end{aligned}$$

(Cochran, 1977, p. 261). Based on $\hat{c}_1$, $\widehat{\pi}_{ij}$ (i and $j = 1, 2, \ldots, n$) can be estimated by

$$\widehat{\pi}_{\widehat{c}_1,ij} = \pi_i \pi_j \left[1 - \frac{\left(1 - \pi_i\right)\left(1 - \pi_j\right)}{\widehat{c}_1} \right]. \tag{9}$$

Hájek's c, that is, $\sum_{i=1}^{N} c_i$, can also be estimated by

$$\widehat{c}_2 = \frac{N}{n} \sum_{i=1}^{n} c_i; \tag{10}$$

thus

$$\widehat{\pi}_{\widehat{c}_2,ij} = \pi_i \pi_j \left[1 - \frac{\left(1 - \pi_i\right)\left(1 - \pi_j\right)}{\widehat{c}_2} \right]. \tag{11}$$

Estimators of Hájek's d

Define Hájek's d as

$$d_{\text{Hájek},\ ij} = \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}$$

(i and $j = 1, 2, \ldots, n$). Then, based on $\widehat{\pi}_{\widehat{c}_1, ij}$ and $\widehat{\pi}_{\widehat{c}_2, ij}$ in Equation 9 and Equation 11, Hájek's d can be estimated:

$$\widehat{d}_{\widehat{c}_1,\ ij} = \frac{\pi_i \pi_j - \widehat{\pi}_{\widehat{c}_1, ij}}{\widehat{\pi}_{\widehat{c}_1, ij}} = \frac{\left(1 - \pi_i\right)\left(1 - \pi_j\right)}{\widehat{c}_1 - \left(1 - \pi_i\right)\left(1 - \pi_j\right)},$$

and

$$\widehat{d}_{\widehat{c}_2,\ ij} = \frac{\pi_i \pi_j - \widehat{\pi}_{\widehat{c}_2, ij}}{\widehat{\pi}_{\widehat{c}_2, ij}} = \frac{\left(1 - \pi_i\right)\left(1 - \pi_j\right)}{\widehat{c}_2 - \left(1 - \pi_i\right)\left(1 - \pi_j\right)}.$$

In addition, instead of estimating π_{ij}, we can estimate $\widehat{d}_{\text{Hájek},\ ij}$ directly. For the unequal probability sampling without replacement, one of the large-sample properties of $(\pi_i \pi_j - \pi_{ij})/\pi_i \pi_j$ is that

$$\frac{n\left(\pi_i \pi_j - \pi_{ij}\right)}{\pi_i \pi_j} \rightarrow 1, \qquad (12)$$

when $N \rightarrow \infty$ while n is fixed, one set of the Hartley–Rao conditions (Hájek, 1964, p. 1495; Hartley & Rao, 1962). The property in Equation 12 is false if $\sum_{i=1}^{N} \pi_i \left(1 - \pi_i\right) \rightarrow \infty$; the formulae based on Equation 12 are applicable if N is much larger than n (Hájek, 1964, p. 1496). The large-sample property in Equation 12 implies

$$\frac{\left(\pi_i \pi_j - \pi_{ij}\right)}{\pi_{ij}} \rightarrow \frac{1}{n - 1}.$$

Therefore the form

$$\widehat{d}_{\text{HR},\ ij} = \frac{1 - \eta}{n - 1} \qquad (13)$$

(Rizzo & Rust, 2011) can be treated as the *lower bound* of the estimator of $\widehat{d}_{\text{Hájek},\ ij}$, where η is a small positive number and can be a function of π_i, π_j, and π_{ij}. Although $\widehat{d}_{\text{Hájek},\ ij}$ can also be expressed in the form $1/(\gamma - 1)$ with γ having a value range (n, ∞), it is more straightforward to discuss modified estimators in the form of Equation 13, as follows.

The empirical results in "Results" section show that η can be estimated adequately by the geometric mean and the arithmetic mean of π_i and π_j:

$$\widehat{d}_{\text{Geo},\ ij} = \frac{1 - \sqrt{\pi_i \, \pi_j}}{n - 1}, \qquad (14)$$

and

$$\widehat{d}_{\text{Arith},\ ij} = \frac{1 - \left(\pi_i + \pi_j\right)/2}{n - 1}. \qquad (15)$$

Compared with $\widehat{d}_{\text{Arith},\ ij}$, the estimator $\widehat{d}_{\text{Geo},\ ij}$ is less conservative. The approximation of $\widehat{d}_{\text{HR},\ ij}$ with $\eta = \min(\pi_{i,}\ \pi_j)$ is

$$\widehat{d}_{\text{RR},\ ij} = \frac{1 - \min\left(\pi_{i,}\ \pi_j\right)}{n - 1} \qquad (16)$$

(Rizzo & Rust, 2011). Compared with school inclusion probabilities, if the joint inclusion probabilities are very small, the estimation of $\widehat{d}_{\text{RR},\ ij}$ can be conservative and underestimated.

The empirical results in the "Results" section show that those yielded by the formula-based variance estimation using Hájek's $\hat{c}_1$ and $\hat{c}_2$ are compatible with those using the approximations $\widehat{d}_{\text{Arith},\ ij}$, $\widehat{d}_{\text{Geo},\ ij}$, and $\widehat{d}_{\text{RR},\ ij}$. Note that, in application, Equation 13 is applicable only if N is much larger than n, which may not be true for a sample drawn from a small state.

The Sampling Approaches Used in the Simulation

In the simulation, the Durbin and Brewer – Rao sampling approaches (Cochran, 1977) are used to draw PPS samples of size 2 ($n = 2$) without replacement. The joint inclusion probabilities are known for the samples drawn by the Durbin and Brewer – Rao approaches. Thus the joint inclusion probabilities estimated with Hájek's approximation can be compared with the known probabilities.

Brewer–Rao Approach (Sample Size n = 2)

Let $Z_i = M_i/M$, where $M = \sum_{i=1}^{N} M_i$. Assume $z_i < 0.5$, which guarantees that every case has a positive probability to be selected; define

$$D = \sum_{i=1}^{N} \frac{(Z_i)\left(1 - Z_i\right)}{1 - 2Z_i} = \frac{1}{2}\left(1 + \sum_{i=1}^{N} \frac{Z_i}{1 - 2Z_i}\right).$$

The Brewer – Rao approach consists of two steps:

1. During the first drawing, case i is selected with the probability $\frac{Z_i(1-Z_i)}{D(1-2Z_i)}$.
2. During the second drawing, case j ($j \neq i$) will be selected with the probability $\frac{Z_j}{1-Z_i}$.

The inclusion probability of unit i is then

$$\begin{aligned}
\pi_i &= \frac{Z_i\left(1 - Z_i\right)}{D\left(1 - 2Z_i\right)} + \sum_{j \neq i}^{N} \frac{Z_j\left(1 - Z_j\right)}{D\left(1 - 2Z_j\right)} \frac{Z_i}{1 - Z_j} \\
&= \frac{Z_i\left(1 - Z_i\right)}{D\left(1 - 2Z_i\right)} - \frac{Z_i Z_i}{D\left(1 - 2Z_j\right)} + \frac{Z_i}{D} \sum_{j=1}^{N} \frac{Z_j}{1 - 2Z_j} \\
&= \frac{Z_i}{D}\left(1 + \sum_{j=1}^{N} \frac{Z_j}{1 - 2Z_j}\right) \\
&= 2Z_i
\end{aligned} \tag{17}$$

.

The joint probability can be expressed as

$$\begin{aligned}
\pi_{ij} &= P\{i \text{ drawn 1st}\} P\left\{j \text{ drawn 2nd} | i \text{ drawn}\right\} + P\left\{j \text{ drawn 1st}\right\} P\left\{i \text{ 2nd} | j \text{ drawn}\right\} \\
&= \frac{Z_i\left(1 - Z_i\right)}{D\left(1 - 2Z_i\right)} \frac{Z_j}{1 - Z_i} + \frac{Z_j\left(1 - Z_j\right)}{D\left(1 - 2Z_j\right)} \frac{Z_i}{1 - Z_j} \\
&= \frac{Z_i Z_j}{D}\left(\frac{1}{1 - 2Z_i} + \frac{1}{1 - 2Z_j}\right).
\end{aligned} \tag{18}$$

It is straightforward to verify that $\sum_{i=1}^{N} \pi_i = 2$, $\sum_{j \neq i}^{N} \pi_{ij} = \pi_i$, and $\sum_{i=1}^{N} \sum_{j<i} \pi_{ij} = 1$. For the Brewer – Rao approach, the term $\pi_i \pi_j - \pi_{ij} > 0$ (J. N. K. Rao, 1965), whereas the same term $\pi_i \pi_j - \pi_{ij}$ in the variance estimates of the H – T estimator can be negative for data drawn by other PPS sampling methods.

Durbin Approach (Sample Size n = 2)

Z_i and D are defined in the same way as for Brewer's approach. The two steps for implementing Durbin's approach are as follows:

1. During the first drawing, case i is drawn with the probability Z_i.
2. During the second drawing, case j ($j \neq i$) will be drawn with the probability

$$\frac{Z_j}{2D}\left(\frac{1}{1-2Z_i}+\frac{1}{1-2Z_j}\right).$$

The inclusion probability of case i is

$$\begin{aligned}\pi_i &= Z_i+\sum_{j\neq i}^{N} Z_j\frac{Z_i}{2D}\left(\frac{1}{1-2Z_i}+\frac{1}{1-2Z_j}\right)\\ &= Z_i+\frac{1-Z_i}{2D}\frac{Z_i}{1-2Z_i}+Z_i-\frac{Z_i\left(1-Z_i\right)}{2D\left(1-2Z_i\right)}\\ &= 2Z_i.\end{aligned} \tag{19}$$

Clearly the joint inclusion probability of cases i and j is

$$\pi_{ij}=\frac{Z_iZ_j}{D}\left(\frac{1}{1-2Z_i}+\frac{1}{1-2Z_j}\right), \tag{20}$$

which is the same as in Brewer's approach. In this regard, Durbin's approach is equivalent to Brewer's.

Results

The Results of the Simulation

To evaluate the Hájek approximation of the joint inclusion probabilities, the simulation data are generated by PPS sampling approaches without replacement, that is, the Durbin and Brewer–Rao approaches introduced in the "Methodology" section of this report. The joint inclusion probabilities of any two units in a sample can be (a) computed by the formulae in Equation 18 or Equation 20, (b) estimated by the proportion of unit pairs in a Monte Carlo process, and (c) estimated by the formulae with Hájek's approximation in Equations 7, 9, and 11. Therefore the estimates of the joint probabilities yielded by the formula with Hájek's approximation can be directly compared with the joint inclusion probabilities.

Table 1 presents a summary of the five population frames used in the simulation. Each frame contains 40 aggregates ($N=40$), corresponding to school clusters in a sampling, with different sizes. The sample units, corresponding to students in a sampling, per aggregate range from 100 to 200, and the standard deviations of aggregate sizes range from 32.1 to 68.44, as listed in columns 4 and 5 in Table 1, respectively. Column 3 contains the total size of the 40 aggregates from the five frames, ranging from 4,000 to 8,000. Let $M=\sum_{i=1}^{40} M_i$ be the total size of the 40 aggregates in a sampling frame. For each frame, the proportion of all the aggregate sizes $z_i=M_i/M$ $(i=1, 2, \ldots, 40)$ is known, implying that the aggregate sizes are normalized to 1, that is, $\sum_{i=1}^{40} z_i=1$. Thus $\pi_i=z_i$ is the inclusion probability of aggregate i in the PPS sampling. In the simulation, each sampling frame is formed by a set of normalized z_i s. For each frame, the parameters $c_i=\pi_i(1-\pi_i)$ and $c=\sum_{i=1}^{40} c_i$ in the Hájek approximation can be estimated.

For every pair of aggregates in the frame, following the two steps of Durbin or Brewer–Rao in the "Methodology" section, PPS sampling is employed to draw samples with 16,000 replications. There are two approaches to estimating π_{ij}, as mentioned earlier: (a) compute using the formulae in Equation 18 and/or Equation 20 with known π_i and π_j in each replication or (b) count the proportion of cases when both aggregates i and j are included in the sample in 16,000 replications. To summarize the results, let $\overline{\pi}_{ij}^{M}$ be the average of the π_{ij} s computed using the first approach based on Equation 18 in 16,000 replications; let $\overline{\pi}_{ij}^{M_e}$ be the average of the estimated $\widehat{\pi}_{ij}$ s estimated using the second approach. Moreover, let $\overline{\pi}_{ij}^{H}$ be the averages of the π_{ij}'s estimated based on Equation 7 with true π_i, π_j, c_i, and c obtained from the frames with 16,000 replications; let $\overline{\pi}_{ij}^{H_e}$ be the average of the estimates based on Equation 9 with estimated parameters from the simulation samples.

Table 2 presents the means and standard deviations of $\overline{\pi}_{ij}^{M}$, $\overline{\pi}_{ij}^{M_e}$, $\overline{\pi}_{ij}^{H}$, and $\overline{\pi}_{ij}^{H_e}$ $(i,j=1, 2, \ldots, 40)$ computed from the simulation samples. The average differences between $\overline{\pi}_{ij}^{H}$ and $\overline{\pi}_{ij}^{M}$ are all less than .00004, which implies that the estimates approximated by Equation 7 are very close to those yielded by the theoretical formula in Equation 18 for the PPS sampling approaches used. When comparing the estimates yielded by the simulation with those computed by theoretical models,

Table 1 Summary of Five Sampling Frames Employed in the Simulation

Population	No. of clusters	Size	Average cluster size	*SD*
Frame 1	40	4,000	100	32.10
Frame 2	40	5,000	125	47.61
Frame 3	40	6,000	150	59.74
Frame 4	40	7,000	175	68.44
Frame 5	40	8,000	200	63.72

Table 2 Means and Standard Deviations of the Joint Probabilities, Yielded by Formulae From Two Sampling Methods in the Simulation

	$\overline{\pi}_{ij}^{M}$		$\overline{\pi}_{ij}^{M_e}$		$\overline{\pi}_{ij}^{H}$		$\overline{\pi}_{ij}^{H_e}$	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Durbin								
Frame 1	0.001284	0.000582	0.001342	0.000659	0.001312	0.000610	0.001317	0.000636
Frame 2	0.001283	0.000699	0.001359	0.000790	0.001317	0.000732	0.001320	0.000769
Frame 3	0.001288	0.000732	0.001361	0.000839	0.001322	0.000770	0.001334	0.000818
Frame 4	0.001285	0.000719	0.001366	0.000808	0.001312	0.000754	0.001322	0.000784
Frame 5	0.001284	0.000578	0.001356	0.000656	0.001319	0.000605	0.001320	0.000641
Brewer								
Frame 1	0.001284	0.000582	0.001382	0.000636	0.001312	0.000610	0.001364	0.000693
Frame 2	0.001283	0.000699	0.001411	0.000754	0.001317	0.000732	0.001393	0.000829
Frame 3	0.001288	0.000732	0.001367	0.000843	0.001322	0.000770	0.001368	0.000887
Frame 4	0.001285	0.000719	0.001376	0.000777	0.001312	0.000754	0.001401	0.000850
Frame 5	0.001284	0.000578	0.001335	0.000624	0.001319	0.000605	0.001371	0.000687

Table 3 Relative Absolute Errors of the Means of the Joint Probabilities for Two Sampling Methods in the Simulation, in Percentages

	Δ_{ij}^{HM}	$\Delta_{ij}^{H_eM}$	$\Delta_{ij}^{M_eM}$	$\Delta_{ij}^{H_eH}$
Durbin				
Frame 1	2.218	2.583	4.553	0.358
Frame 2	2.616	2.871	5.902	0.249
Frame 3	2.640	3.563	5.724	0.900
Frame 4	2.101	2.924	6.335	0.806
Frame 5	2.701	2.814	5.565	0.110
Brewer				
Frame 1	2.218	6.254	7.678	3.949
Frame 2	2.616	8.521	9.969	5.755
Frame 3	2.640	6.201	6.113	3.470
Frame 4	2.101	9.048	7.090	6.804
Frame 5	2.701	6.746	3.971	3.938

that is, $\overline{\pi}_{ij}^{M_e}$ versus $\overline{\pi}_{ij}^{M}$ and $\overline{\pi}_{ij}^{H_e}$ versus $\overline{\pi}_{ij}^{H}$, they are all less than .0001 in average absolute error. The average differences between $\overline{\pi}_{ij}^{H_e}$ and $\overline{\pi}_{ij}^{M_e}$ are less than .00005, which implies that the estimates approximated by Equation 9 are very close to those counted by the proportion of aggregate pairs in the simulation samples. The accuracy of the estimates in Table 2 can be better measured by their relative absolute errors of the joint probabilities.

Table 3 presents the relative absolute errors of the means and standard deviations of the joint probabilities. In the table, $\Delta_{ij}^{HM} = \left|\overline{\pi}_{ij}^{H} - \overline{\pi}_{ij}^{M}\right| / \overline{\pi}_{ij}^{M}$, $\Delta_{ij}^{H_eM} = \left|\overline{\pi}_{ij}^{H_e} - \overline{\pi}_{ij}^{M}\right| / \overline{\pi}_{ij}^{M}$, $\Delta_{ij}^{H_eH} = \left|\overline{\pi}_{ij}^{H_e} - \overline{\pi}_{ij}^{H}\right| / \overline{\pi}_{ij}^{H}$, and $\Delta_{ij}^{M_eM} = \left|\overline{\pi}_{ij}^{M_e} - \overline{\pi}_{ij}^{M}\right| / \overline{\pi}_{ij}^{M}$. The first column of Δ_{ij}^{HM} are slightly above 2% in average relative absolute errors. The average relative absolute errors of $\Delta_{ij}^{H_eH}$ for Durbin's approach are much smaller than those for Brewer's. An additional two types of average relative absolute error, that is, $\Delta_{ij}^{H_eM}$ and $\Delta_{ij}^{M_eM}$, have a range from 2.5% to 7.5%, except for three cases that are between 8% and 9.9%. In general, the average relative absolute errors for Durbin's approach are also smaller than those for Brewer's.

Table 4 Sizes of the Schools in Sample and Population, Mean Estimates, Estimates of Hájek's $\hat{c}_1$ and $\hat{c}_2$, Standard Errors of the Mean Scores Yielded by the Formula with Hájek's $\hat{c}$ Estimates, and the Jackknifed Standard Errors for Five State Samples

State	No. of schools sampled	No. of schools in population	Mean score	$\hat{c}_1$	$\hat{c}_2$	$SE_{\hat{c}_1}$	$SE_{\hat{c}_2}$	SE_J
All								
1	114	1,637	138.64	75.24	79.57	0.826	0.828	1.141
2	117	1,102	143.53	52.26	96.05	0.724	0.690	1.303
3	120	1,825	155.66	70.65	89.24	0.821	0.738	0.956
4	155	4,269	145.73	85.00	118.92	2.258	2.183	1.027
5	120	2,387	146.88	71.97	88.44	1.581	1.473	0.992
Male								
1	114	1,637	141.58	75.56	79.35	1.041	1.030	1.422
2	117	1,102	145.15	53.53	95.42	1.214	1.216	1.849
3	120	1,825	157.84	71.37	88.80	1.213	1.090	1.355
4	155	4,269	147.62	85.42	118.68	3.064	2.917	1.312
5	120	2,387	149.93	72.95	87.83	1.983	1.807	1.076
Female								
1	114	1,637	135.59	75.35	79.49	1.035	1.062	1.273
2	117	1,102	141.86	52.52	95.92	0.841	0.804	1.148
3	120	1,825	153.38	71.02	89.02	1.114	1.095	1.114
4	155	4,269	143.82	85.02	118.91	3.000	2.990	1.182
5	120	2,387	143.77	72.08	88.37	2.227	2.167	1.276

Note. Note that unidimensional scaling was applied to item calibration for the 2009 NAEP science assessments.

Several factors can cause the discrepancies in the relative absolute errors. First, the relative absolute error is a ratio estimator that is subject to large errors; second, the joint inclusion probabilities are the estimates of very small proportions with an average value of .12 to .14; third, the size of the simulation is limited to $N = 40$, $n = 2$, and 16,000 replications. Because the large-sample properties of the PPS sampling without replacement (Hájek, 1964) are derived based on Poisson sampling when $n \to \infty$ and $(N - n) \to \infty$, all the simulation parameters, such as N, n, and replication times, need to be expanded for better accuracy. In particular, as n increases, the increase of N must occur at a faster rate, and the number of combinations of possible joint selections will increase exponentially. This is indeed a computationally heavy task, even with contemporary software such as the SAS packages (SAS Institute Inc., 2011, 2014).

Empirical Variance Estimates Yielded by the Formula Incorporating the Hájek Approximation

The empirical data employed in assessing the formula-based variance estimates using the Hájek approximation are the 2009 NAEP state science assessment samples, which were drawn using a two-stage systematic PPS sample design using the systematic PPS selection for schools at the first stage and SRS selection for students at the second stage. Although the inclusion probability of any school is proportional to size for the systematic PPS sampling, it differs somewhat from the regular PPS sampling because some pairs of schools in the sampling frame can be excluded from the samples. Moreover, a sorting algorithm for the sample units in a sampling frame can impose some effects on variance estimates (Cochran, 1977, pp. 212–221), although the stratification and sorting algorithms for the sample units in the NAEP sample design are carefully considered (Allen et al., 2001).

In formula-based variance estimation with empirical data, the estimators of Hájek's $\hat{c}_1$ and $\hat{c}_2$ in Equations 8 and 10 and of $\widehat{d}_{\text{Arith},\, ij}$, $\widehat{d}_{\text{Geo},\, ij}$, and $\widehat{d}_{\text{RR},\, ij}$ in Equations 14–16 are employed. Note that the variance estimates of means yielded by the jackknife replicate resampling (JRR) procedure (Wolter, 2007) are also presented for reference. To implement the jackknife procedure, the jackknifing strata are created by first aggregating a pair of groups (e.g., primary sampling units or schools) in one stratum; then, a replicate sample is formed by randomly dropping one school and doubling the weights of the cases in the remaining school. Moreover, the strata are formed in a way consistent with the sampling mechanism. The details of the NAEP jackknife procedure can be found in the *NAEP 1998 Technical Report* (Allen et al., 2001; Qian, 2005). Note that all the assessment items of the 2009 NAEP science are treated as being of one-dimensional scale in operational analysis.

Table 4 presents the sizes of the schools in the population and the schools sampled and the mean estimates of the five states of the NAEP 2009 science assessment. The table contains the variances of the mean scores obtained from the

Table 5 Correlation Coefficients Between $SE_{\hat{c}_1}$ and Other Standard Errors Estimated Incorporating Hájek's Approximations Across States for Total, Male, and Female Groups

	$SE_{\hat{c}_2}$	$SE_{\hat{d}_{\text{Arith},\,\cdot\cdot}}$	$SE_{\hat{d}_{\text{Geo},\,\cdot\cdot}}$	$SE_{\hat{d}_{\text{RR},\,\cdot\cdot}}$
All	0.998	0.997	0.990	0.986
Male	0.997	0.995	0.990	0.985
Female	0.999	0.999	0.998	0.998

Table 6 Standard Deviations of the Standard Errors Estimated by the Formulae with $\hat{c}_1$ and $\hat{c}_2$

	Total	Male	Female
$\hat{c}_1$	.64	.84	.93
$\hat{c}_2$	.64	.79	.93

formula with Hájek's $\hat{c}_1$ and $\hat{c}_2$ for five state samples. The empirical results incorporating Hájek's $\hat{c}_1$ are very close to those with $\hat{c}_2$.

Table 5 presents the correlation coefficients between $SE_{\hat{c}_1}$ and other standard errors estimated incorporating Hájek's approximations across states for total, male, and female groups. For five state samples, the correlation coefficients between the standard errors estimated with Hájek's $\hat{c}_1$ and the standard errors with $\hat{c}_2$ for total, male, and female groups are .998, .997, and .999, respectively. Moreover, the standard deviations of the standard errors estimated by the formula with $\hat{c}_1$ are also very close to those estimated with $\hat{c}_2$ for total, male, and female groups. The results of the standard deviations of the standard errors estimated using Hájek's $\hat{c}_1$ and $\hat{c}_2$ can be found in Table 6.

Table 7 presents the estimates of $\hat{d}_{\text{Arith},\,ij}$, $\hat{d}_{\text{Geo},\,ij}$, and $\hat{d}_{\text{RR},\,ij}$ and the variances of mean scores estimated by the variance estimator in Equation 6 using these $\hat{d}_{ij}$ estimates. As expected, the standard errors estimated incorporating $\hat{d}_{\text{Arith},\,ij}$ are the smallest, whereas those incorporating $\hat{d}_{\text{RR},\,ij}$ are the largest. Using the $SE_{\hat{d}_{\text{RR},\,\cdot\cdot}}$ as the basis for comparison, the relative absolute errors of the $SE_{\hat{d}_{\text{Geo},\,\cdot\cdot}}$ for total, male, and female groups are approximately 5.5%, 3.5%, and 3.8% smaller, respectively, whereas those $SE_{\hat{d}_{\text{Arith},\,\cdot\cdot}}$ for total, male, and female groups are approximately 13.3%, 9.0%, and 9.5% smaller, respectively. The approximation $\hat{d}_{\text{Arith},\,ij}$ is the least conservative among the three, whereas, by contrast, the approximation $\hat{d}_{\text{RR},\,ij}$ can cause overestimation in variance. In application, the approximation $\hat{d}_{\text{Geo},\,ij}$ is relatively robust, even when there exist large gaps in the numbers of students in school pairs in a sample.

In general, the three sets of standard error estimates incorporating $\hat{d}_{\text{Arith},\,ij}$, $\hat{d}_{\text{Geo},\,ij}$, and $\hat{d}_{\text{RR},\,ij}$, respectively, are compatible with and close to each other. In Table 5, the correlation coefficients across states between $SE_{\hat{c}_1}$ and the standard errors with other Hájek's approximations are all larger than .98 for total, male, and female groups. For example, using $SE_{\hat{c}_1}$ as the basis, the correlations between the standard errors estimated with $\hat{d}_{\text{Arith},\,ij}$ and those estimated with Hájek's $\hat{c}_1$ for total, male, and female groups are .997, .995, and .999, respectively. Evidently, the empirical estimates incorporating $\hat{d}_{\text{Arith},\,ij}$, $\hat{d}_{\text{Geo},\,ij}$, and $\hat{d}_{\text{RR},\,ij}$ are all compatible with those estimated with Hájek's $\hat{c}_1$, in particular, for those based on $\hat{d}_{\text{Geo},\,ij}$.

The correlation coefficients between the standard errors estimated with Hájek's c (c_1 or c_2) and the jackknifed standard errors range from .25 to .45—not as high as those between $SE_{\hat{c}_1}$ and the standard errors with other Hájek's approximations. It appears that the standard errors estimated incorporating the Hájek approximation are more volatile than those from the jackknifing approach. Compared with the five standard errors yielded by the jackknifing procedure, either for total, male, or female groups, formula-based standard errors are smaller in about three out of five cases. A couple of confounding factors could have caused this trend. The NAEP state samples were selected using a two-stage systematic PPS sample design from a sampling frame sorted by demographics, and these samples differ with the data drawn by PPS sampling without replacement. The use of systematic sampling is likely to reduce sampling variance (Burke & Rust, 1995). Although the formula-based variance estimation incorporating the Hájek approximation is appropriate to be applied to the data drawn by PPS sampling, further studies need to be pursued before applying it to the data drawn by systematic PPS sampling. The results of the JRR variance estimates are not the focus of this study and are presented for reference only.

Table 7 Estimates of $\widehat{d}_{\text{Arith},\ ij}$, $\widehat{d}_{\text{Geo},ij}$, and $\widehat{d}_{\text{RR},\ ij}$, the Standard Errors of the Mean Scores Yielded by the Formula With the $\widehat{d}_{..,\ ij}$ Estimates, and Jackknifed Standard Errors for Five State Samples

State	No. of schools sampled	No. of schools in population	$\widehat{d}_{\text{Arith},..}$	$\widehat{d}_{\text{Geo},\ ..}$	$\widehat{d}_{\text{RR},\ ..}$	$SE_{\widehat{d}_{\text{Arith},\ ..}}$	$SE_{\widehat{d}_{\text{Geo},\ ..}}$	$SE_{\widehat{d}_{\text{RR},\ ..}}$	SE_J
All									
1	114	1,637	0.894	1.002	1.085	0.894	1.002	1.085	1.141
2	117	1,102	0.819	1.001	1.095	0.819	1.001	1.095	1.303
3	120	1,825	0.798	0.882	0.941	0.798	0.882	0.941	0.956
4	155	4,269	2.192	2.210	2.232	2.192	2.210	2.232	1.027
5	120	2,387	1.503	1.559	1.606	1.503	1.559	1.606	0.992
Male									
1	114	1,637	1.091	1.198	1.282	1.091	1.198	1.282	1.422
2	117	1,102	1.314	1.468	1.552	1.314	1.468	1.552	1.849
3	120	1,825	1.146	1.228	1.288	1.146	1.228	1.288	1.355
4	155	4,269	2.926	2.943	2.965	2.926	2.943	2.965	1.312
5	120	2,387	1.834	1.882	1.882	1.834	1.882	1.882	1.076
Female									
1	114	1,637	1.114	1.203	1.273	1.114	1.203	1.273	1.273
2	117	1,102	0.913	1.074	1.158	0.913	1.074	1.158	1.148
3	120	1,825	1.141	1.209	1.259	1.141	1.209	1.259	1.114
4	155	4,269	2.996	3.010	3.026	2.996	3.010	3.026	1.182
5	120	2,387	2.190	2.235	2.274	2.190	2.235	2.274	1.276

Summary

The simulation conducted in this study shows that the joint probabilities estimated by the Hájek approximation are very close to those yielded by the formulae provided by the sampling methods. The relative absolute errors of the joint probabilities between the Hájek approximation and those yielded from the simulation are also quite small.

Several estimators of Hájek's c and d_{ij} have been proposed in this study. In analyzing the real samples drawn by a two-stage systematic PPS sample design, the five sets of results, yielded based on five distinct Hájek's approximates (i.e., Hájek's $\hat{c}_1$, $\hat{c}_2$, $\widehat{d}_{\text{Arith},\ ij}$, $\widehat{d}_{\text{Geo},\ ij}$, and $\widehat{d}_{\text{RR},\ ij}$) are compatible with each other. This implies that the formula-based variance estimation incorporating the Hájek approximation is stable and appropriate, in particular for the estimation based on Hájek's $\hat{c}_1$ and $\widehat{d}_{\text{Geo},\ ij}$. The application of the Hájek approximation can certainly be extended to variance estimation in assessment data.

However, there are some discrepancies between the results yielded by the formula-based variance estimation and those from the grouped jackknifing approach. The NAEP state samples were drawn by systematic PPS sampling without replacement. This may imply that it is appropriate to apply formula-based variance estimation incorporating the Hájek approximation to the data drawn by PPS sampling; for systematic PPS sampling, further studies are needed. It was also found that the formula-based variance estimates are slightly more volatile than those from the jackknifing approach. However, the empirical analysis in this study has no power to tell which set of results reflect the true variability of the variance estimates for the NAEP state samples.

Acknowledgments

The author thanks Shelby Haberman, Yue Jia, Bruce Kaplan, Keith Rust, and Rebecca Zwick for their suggestions and comments. The author also thanks Kim Fryer and Shuhong Li for editorial help. Any opinions expressed in this paper are those of the author and not necessarily those of Educational Testing Service.

Note

1 PPS sampling without replacement can be implemented through different specific approaches (Brewer & Hanif, 1983). For example, Durbin and Brewer-Rao PPS sampling approaches (Cochran, 1977, pp. 261–262) can be used to draw a pair of sample units (i.e., sample size $n = 2$). Rejective sampling, based on Poisson sampling, can be used to select a PPS sample of size n (Hájek, 1981, pp. 54–60, 66–72). Moreover, NAEP uses systematic PPS sampling to select schools (sample size n) in the first stage of state

samples (Allen et al., 2001, pp. 61 – 77). The sample design of the systematic PPS approach, which is different from standard PPS sampling, does not assign a nonzero chance of including every pair of sample units in the sampling frame.

References

Allen, N., Donoghue, J., & Schoeps, T. (2001). *The NAEP 1998 technical report* (NCES Report No. 2001-509). Washington, DC: National Center for Education Statistics.

Berger, Y. G. (2003). A modified Hájek variance estimator for systematic sampling, *Statistics in Transition*, *6*, 5 – 21.

Berger, Y. G. (2004). A simple variance estimator for unequal probability sampling without replacement. *Journal of Applied Statistics*, *31*, 305 – 315. https://doi.org/10.1080/0266476042000184046

Brewer, K. R. W., & Hanif, M. (1983). *Sampling with unequal probabilities*. New York, NY: Springer. https://doi.org/10.1007/978-1-4684-9407-5

Burke, J., & Rust, K. (1995). On the performance of jackknife variance estimation for systematic samples with small numbers of primary sampling units. In *JSM proceedings, Survey Research Methods Section* (pp. 321 – 326). Alexandria, VA: American Statistical Association.

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

Feng, S., Ni, J., & Zou, G. (1998). *Theory and methods of survey sampling*. Beijing, China: China Statistical Publishing House.

Fuller, W. A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, *96*, 933 – 944. https://doi.org/10.1093/biomet/asp042

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences, 5*, 361 – 371.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics, 35*, 1491 – 1523. https://doi.org/10.1214/aoms/1177700375

Hájek, J. (1981). *Sampling from a finite population*. New York, NY: Dekker.

Hartley, H. O., & Rao, J. N. K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, *33*, 350 – 374. https://doi.org/10.1214/aoms/1177704564

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*, 663 – 685. https://doi.org/10.1080/01621459.1952.10483446

Kali, J., Burke, J., Hicks, L., Rizzo, L., & Rust, K. (2011). Incorporating a first-stage finite population correction (FPC) in variance estimation for a two-stage design in the National Assessment of Educational Progress (NAEP). In *JSM proceedings, Survey Research Methods Section* (pp. 2576 – 2583). Alexandria, VA: American Statistical Association.

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

Lohr, S. L. (1999). *Sampling: Design and analysis.* Pacific Grove, CA: Duxbury Press.

Nohara, D. (2001). *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)* (NCES Report No. 2001-07). Washington, DC: National Center for Education Statistics.

Qian, J. (2005, April). *Measuring the cumulative linking errors of NAEP trend assessments.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Qian, J. (2015, April). *Applying the Hájek approach in the delta method of variance estimation.* Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Rao, C. R. (1973). *Linear statistical inference and its applications*. New York, NY: John Wiley. https://doi.org/10.1002/9780470316436

Rao, J. N. K. (1965). On two simple schemes of unequal probability without replacement. *Journal of the Indian Statistical Association, 3*, 173 – 180.

Rizzo, L., & Rust, K. (2011). Finite population correction for NAEP variance estimation. In *JSM proceedings, Survey Research Methods Section* (pp. 2502 – 2515). Alexandria, VA: American Statistical Association.

Rust, K. F. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics, 1*, 381 – 397.

Rust, K. F., Bethel, J., Burke, J., & Hansen, M. H. (1990). *The 1988 National Assessment of Educational Progress: Sampling and weighting procedures*. Rockville, MD: Westat.

Rust, K. F., & Johnson, E. G. (1992). Sampling and weighting in the national assessment. *Journal of Educational Statistics, 17*(2), 111 – 129. https://doi.org/10.2307/1165165

Särndal, C. E., & Lundström, S. (2005). *Estimation in surveys with nonresponse*. New York, NY: John Wiley. https://doi.org/10.1002/0470011351

Särndal, C. E., Swenson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York, NY: Springer. https://doi.org/10.1007/978-1-4612-4378-6

SAS Institute Inc. (2011). *SAS/IML 9.3 user's guide* [Software manual]. Cary, NC: Author.

SAS Institute Inc. (2014). *SAS 9.3 macro language: Reference* (3rd ed.) [Software manual]. Cary, NC: Author.

Sen, A. R. (1953). On the estimate of variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics, 5*, 119–127.

Turner, R., & Adams, R. J. (2007). The Program for International Student Assessment: An overview. *Journal of Applied Measurement, 8*, 237–248.

Wolter, K. (2007). *Introduction to variance estimation*. New York, NY: Springer.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association, 66*(334), 411–414. https://doi.org/10.1080/01621459.1971.10482279

Yates, F., & Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society: Series B, 15*, 253–261.

Suggested citation:

Qian, J. (2017). *Applying the Hájek approach in formula-based variance estimation* (Research Report No. RR-17-24). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12154

Action Editor: Rebecca Zwick

Reviewers: Yue Jia and Bruce Kaplan