Research Report

# Variance Estimation With Complex Data and Finite Population Correction — A Paradigm for Comparing Jackknife and Formula-Based Methods for Variance Estimation

ETS RR–20-11

Jiahe Qian

*December 2020*

# ETS Research Report Series



Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

# Variance Estimation With Complex Data and Finite Population Correction—A Paradigm for Comparing Jackknife and Formula-Based Methods for Variance Estimation

Jiahe Qian

Educational Testing Service, Princeton, NJ

The finite population correction (FPC) factor is often used to adjust variance estimators for survey data sampled from a finite population without replacement. As a replicated resampling approach, the jackknife approach is usually implemented without the FPC factor incorporated in its variance estimates. A paradigm is proposed to compare the jackknifed variance estimates with those yielded by the delta method because the delta method has the effect of the FPC factor implicitly integrated. The goal is to examine whether the grouped jackknife approach properly estimates the variance of complex samples without incorporating the FPC effect, in particular for data sampled from a finite population with a high sampling rate. The investigation focuses on the data drawn by two-stage sampling with probability proportional to size of schools and with simple random sampling of students. Moreover, the Hájek approximation of the joint probabilities is used in the delta method for a Horvitz–Thompson (H–T) estimator. Samples of the National Assessment of Educational Progress (NAEP) state science assessments are used in the analysis.

**Keywords** Finite population sampling; two-stage PPS sampling; Horvitz–Thompson estimator; jackknife approach; delta method; joint inclusion probability; Hájek approximation

doi:10.1002/ets2.12294

The finite population correction (FPC) factor is often used to adjust a variance estimator for surveys sampled from a finite population without replacement (Cochran, 1977; Kish, 1965). The variance formula incorporating the FPC factor for simple random sampling without replacement (SRS/wor) is provided in Cochran (1977, p. 24) and in the Methods section of this paper. Because the FPC factor is less than 1, the variance estimates adjusted by multiplying the FPC factor are thus smaller than those without incorporating the factor. In educational surveys, to ensure efficient and practical implementation, most assessment survey data are collected through complex sampling designs, such as two-stage sampling with probability proportional to size (PPS) selection of schools and with SRS of students. Some examples are the National Assessment of Educational Progress (NAEP; Allen et al., 2001; Rust, 1985) and the Programme for International Student Assessment (Neidorf et al., 2006; Nohara, 2001).

For complex data, variance estimation is often based on the delta method and resampling techniques, such as jackknife repeated replication (JRR) and balanced repeated replication (BRR; Cochran, 1977; Hájek, 1981; Kish, 1965; Rust, 1985; Wolter, 2007). Failure to incorporate the effect of the FPC factor, also called the FPC effect, is one issue that should concern statisticians in estimating variances using resampling techniques (Rizzo & Rust, 2011). When the size of a finite population is small, variances can be appreciably overestimated if the FPC effect is omitted (Fay, 1989; Rizzo & Rust, 2011; Wolter, 2007).

The delta method estimates the variance based on a Taylor series expansion (Cochran, 1977). For two-stage PPS sampling, the variance formula consists of two terms: the errors between the first-stage sample units and the errors within the sampled units. The delta method takes into account a primary feature of finite population sampling, namely, incorporating the FPC effect (see the Horvitz–Thompson Estimators and the Delta Method section of this paper and Cochran, 1977, Chapters 9A and 11). Although the delta method is an efficient and robust approach for variance estimation (Fuller, 1996; Wolter, 2007), applying the delta method requires the joint inclusion probabilities for each pair of sample units, whereas most of the data sets for analysis lack such information. Hájek (1964) solved this issue by approximating the joint inclusion probabilities. Hájek (1964, 1981) also derived the properties of the inclusion probabilities and variance formulas

*Corresponding author:* J. Qian, E-mail: jqian58@gmail.com

for the estimates yielded by Poisson sampling. He further refined these results for estimates related to rejective sampling (Hájek, 1964). Särndal et al. (1992) used rejective sampling in model-assisted survey sampling. Berger (2004) applied the Hájek estimator to weighted least squares regressions for complex data.

Many statisticians have discussed the oversight of not incorporating the FPC effect in variance estimation in resampling techniques (Fay, 1989; Kalton, 2002; Lee et al., 1995; Rao & Shao, 1992; Rao & Sitter, 1995; Rizzo & Judkins, 2004; Rizzo & Rust, 2011; Steel & Fay, 1995). For SRS without replacement and stratified sampling, Kalton (1979) proposed an approximate design to solve the issue within first-stage sample units. Fay (1984, 1989) developed an eigenvalue-matching approach to adjust the replicate weights for the first-stage sample units. Rizzo and Judkins (2004) applied Fay's approach to the jackknife procedure in analyzing the National Survey of Parents and Youth, and Fuller (1998) applied this approach to regression estimators in two-stage samples. Rao and Wu (1988) used consistent bootstrap estimators for multistage samples that were subject to nontrivial FPCs. For each replicate at each stage, Wolter (2007) introduced the appropriate scaling factors to resolve the issues arising from JRR or BRR procedures. For a two-stage PPS sampling without replacement, such as the NAEP state samples, Rizzo and Rust (2011) proposed an approximate method to incorporate the FPC effect into the replicate weights of the first-stage sampling units. Kali et al. (2011) evaluated the approximate method proposed by Rizzo and Rust in the grouped jackknifing procedure for the NAEP 2009 reading state samples. Although as a classic approach, the delta method can be adequately applied in variance estimation for cosmplex data including the two-stage PPS samples of the NAEP state assessments (Cochran, 1977; Qian, 2017; Wolter, 2007), there has been no literature on assessing the consequences of failing to incorporate the FPC effect in the jackknifing procedure by comparing the jackknifed variances with the delta variances.

In this study, a paradigm is proposed to compare the jackknifed variance estimates with those yielded by the delta method that applies adequate Hájek joint probability approximations and then evaluates the consequences of omitting the incorporation of the FPC effect in the jackknifing procedure. The goal is to investigate the impact of neglecting the FPC effect in the jackknifed variance estimation with complex data, in particular for two-stage PPS samples such as the NAEP state samples. The proposed paradigm is to examine changes in the delta variance estimates as the Hájek approximation of joint inclusion probabilities varies and compare the delta estimates with the jackknifed estimates. Instead of presenting the comparison at one point, the paradigm is designed to provide an overall picture of comparisons at all possible values of the Hájek approximation. Then the jackknifed variances without the FPC effect were examined against those of the delta method and against related results in literature. In the study, real data from the eighth-grade NAEP state science assessment samples were used.

Although the empirical results are based on data from a two-stage PPS sampling without replacement, the application of the paradigm can be extended to other complex data such as one-stage unequal sampling with unequal probabilities without replacement. Furthermore, the paradigm can also be used in comparisons involving other resampling methods of variance estimation with or without FPC: for example, BRR and bootstrap (Wolter, 2007).

The next section provides a review of the method applied in the study, including PPS sampling without replacement for a two-stage sampling design, the grouped jackknife procedure, the Horvitz–Thompson (H–T) estimator, the delta methods, and the Hájek approximation of joint probability. The third section consists of the comparison results: the jackknifed variance estimates versus the delta variance estimate with different Hájek approximations. The final section offers a summary and conclusions.

## Method

### Data

The data used in the analysis were the samples of the NAEP 2009 state science assessments (National Center for Education Statistics, 2011), which were selected based on a two-stage *systematic* PPS sample design for schools and SRS for students (Allen et al., 2001). There were three subscales (Physical Science, Life Science, and Earth and Space Sciences) for the NAEP science assessment. The NAEP proficiency scores were reported by using plausible values (Allen et al., 2001; National Center for Education Statistics, 2011). In selecting the state samples, the first step was to partition the states into three categories based on the sizes of their school populations: small, middle, and large, with cut-points of 1,100 and 2,200. Then, I randomly selected 1, 2, and 2 states from the categories, respectively. Table 1 presents the sampling rates of schools, the mean subscale scores, and composite scores for the five state samples. For the QC purpose, Table 1 also included the male

**Table 1** The School Population Sizes, Sampling Rate of Schools, and the Weighted Means of Subscale Scores and Composite Scores for the Five State Samples

| State | School population size | Sampling rate of schools (%) | Mean of physical science | Mean of Earth science | Mean of life science | Mean of composite |
|---|---|---|---|---|---|---|
| All | | | | | | |
| 1 | Middle | 6.62 | 138.35 | 138.86 | 139.71 | 139.05 |
| 2 | Small | 10.11 | 143.03 | 145.23 | 143.54 | 143.89 |
| 3 | Middle | 6.10 | 155.76 | 155.09 | 154.84 | 155.19 |
| 4 | Large | 3.53 | 147.31 | 145.62 | 145.77 | 146.19 |
| 5 | Large | 4.92 | 148.17 | 145.04 | 146.16 | 146.43 |
| Male | | | | | | |
| 1 | Middle | 6.62 | 143.91 | 146.29 | 143.57 | 144.49 |
| 2 | Small | 10.11 | 147.40 | 151.42 | 148.12 | 148.89 |
| 3 | Middle | 6.10 | 163.86 | 163.50 | 161.03 | 162.62 |
| 4 | Large | 3.53 | 151.74 | 149.40 | 147.97 | 149.53 |
| 5 | Large | 4.92 | 153.83 | 150.32 | 149.73 | 151.14 |
| Female | | | | | | |
| 1 | Middle | 6.62 | 136.40 | 134.99 | 139.55 | 137.23 |
| 2 | Small | 10.11 | 142.44 | 143.37 | 143.12 | 142.99 |
| 3 | Middle | 6.10 | 156.93 | 155.89 | 157.92 | 157.01 |
| 4 | Large | 3.53 | 145.74 | 144.69 | 146.42 | 145.70 |
| 5 | Large | 4.92 | 146.15 | 143.32 | 146.23 | 145.33 |

*Note*. The composite results in the far-right column were created by combining Physical Science, Life Science, and Earth Science subscale scores with weights of .3, .3, and .4, respectively. The allocation was consistent with the distribution of items by content area in NAEP science assessments before 2009. However, the results in *The Nation's Report Card* for 2009 were based on a univariate scale including all items (National Center for Education Statistics, 2011). Therefore, the tabled results may differ from those reported.

and female scores. For each state in Table 1, male students scored higher on average than female students, which was consistent with those scores reported by NAEP.

To avoid outlier effects in the computation, the schools with extremely small inclusion probabilities were excluded from the analysis because, for these kinds of schools, $\delta_{\text{Arith},\,ij}$ and $\delta_{\text{RR},\,ij}$ can have excessive values, and $\widehat{d}_{\text{Arith},ij}$ and $\widehat{d}_{\text{RR},ij}$ might have excessive estimates. The exclusion criterion was based on the interquartile range (IQR). A school $j$ with small inclusion probability $p_j$ was treated as an outlier when it was 3 IQR away from the median, i.e., $|p_j - p_{\text{median}}| \geq 3\text{IQR}$ (Hoaglin et al., 1983). For a symmetric distribution, this criterion is similar to that of 3SD (Mosteller & Tukey, 1977). The actual exclusion rate is less than 2%.

## Simple Random Sampling Without-Replacement and Finite Population Correction

A special case of complex sample design is SRS/wor. For a finite population consisting of $N$ sample units, SRS/wor selects a sample of $n$ sample units without replacement from the population; in each selection, all units retained in the population will have the same probability for selection. This selection procedure continues until $n$ units are drawn. For SRS/wor, it is straightforward to verify that every unit in the sample has the same inclusion probability of $n/N$ to be selected.

Let $\overline{Y} = N^{-1}\sum_{i=1}^{N} y_i$ be the population mean of interest, and let $\overline{y} = n^{-1}\sum_{i=1}^{n} y_i$ be the sample mean. For SRS/wor, the expectation of $\overline{y}$ equals $\overline{Y}$ (i.e., $\overline{y}$ is an unbiased estimator of $\overline{Y}$), and the variance of $\overline{y}$ is equal to

$$V\left(\overline{y}\right) = \frac{1-f}{n}S^2,$$

(Cochran, 1977, p. 24), where $f = n/N$, and

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2.$$

The term $1-f=(N-n)/N$ is the FPC factor for the variance, and $\sqrt{(N-n)/N}$ is the FPC factor for standard errors (Cochran, 1977, p. 24). Clearly, the FPC factor is less than 1; with the increase of the sampling rate $n/N$ as $n \to N$, the impact of the FPC factor can be enormous.

## The Grouped Jackknifing Procedure and Variance Estimation

In variance estimation for complex data, the grouped JRR (GJRR) method is widely used because GJRR is easy to be implemented and yields robust results. The complete GJRR method (CGJRR; Haberman et al., 2009) can be applied to estimating the standard errors of all steps for a whole statistical procedure (e.g., the linking procedure used in analyzing assessment data includes several steps such as item response theory [IRT] calibration, item parameter scaling, and IRT linking, all of which are amenable to CGJRR use).

For GJRR, a replicate sample is formed by randomly dropping one group of cases from the sample and making adjustment to weights based on statistical techniques such as minimum discriminant information adjustment (Haberman, 2015), raking, and poststratification (Qian et al., 2013). Let $J$ be the number of total replicate samples. For NAEP, $J$ is usually set to 62. For the whole sample and each jackknife replicate sample, we first compute the parameter estimate of interest. Then we derive the jackknifed SEs of the parameters of interest. Let $\mu$ be the parameter estimated from the whole sample and $\mu_{(j)}$ be the estimate from the $j$th jackknife replicate sample. The variance of $\mu$ was estimated by

$$v(\mu)=\frac{J-1}{J}\sum_{j=1}^{J}\left[\mu_{(j)}-\mu_{(.)}\right]^2, \tag{1}$$

where $\mu_{(.)}$ is the mean of all $\mu_{(j)}$ (Wolter, 2007). A special GJRR is paired grouped jackknifing (PGJRR; Allen et al., 2001), in which two or several balanced groups are aggregated into one stratum. For example, NAEP creates a jackknifing stratum by aggregating a pair of groups (e.g., primary sampling units or schools) in one stratum, and then a replicate sample is formed by randomly dropping one school and doubling the weights for cases in the remaining school. The variance of $\mu$ is estimated by the same formula in Equation 1.

The efficacy of the JRR procedure is based on its large-sample property that the distribution of the estimates yielded by the replicate samples approximates the distribution of the population parameter of interest. Nonetheless, the jackknifing procedure disregards an important feature of finite population sampling: namely, the aforementioned FPC effect in estimation.

## Horvitz–Thompson Estimators and the Delta Method

The variance formula of the delta method (Cochran, 1977, p. 308) discussed in this section is derived based on a two-stage PPS sampling without replacement because, in assessment surveys, the first step is to sample schools and the next is to draw students from the sampled schools.

### *Horvitz–Thompson Estimators*

For a population of $N$ schools, a population total of interest is $\widetilde{Y}=\sum_{i=1}^{N}\widetilde{Y}_i$, a sum of all school totals in the population. The school total is $\widetilde{Y}_i=\sum_{k=1}^{M_i}y_{ik}$, a sum of all students $y_{ik}$ in school $i$, and let $\pi_i$ be the inclusion probability of school $i$. For a school sample of size $n$, the weight for school $i$ ($i=1, 2, \ldots, n$) is equal to the inverse of $\pi_i$, i.e., $w_i=1/\pi_i$ (Allen et al., 2001). Let $M_i$ and $m_i$ be the school size and its sample size for school $i$, respectively. Let $y_{ik}$ be the value of a variable of interest for student $k$ in school $i$. Let $w_{ik}$ be the case weight for student $k$ in school $i$ (Allen et al., 2001). The total for school $i$ can be estimated by $\widetilde{y}_i=\frac{M_i}{m_i}\sum_{k=1}^{m_i}w_{ik}y_{ik}$, and the H–T estimator (Cochran, 1977; Horvitz & Thompson, 1952) for the total is defined as

$$\widetilde{y}_{HT}=\sum_{i=1}^{n}w_i\widetilde{y}_i, \tag{2}$$

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

The H–T estimator of the mean is a ratio estimator $\bar{y}_{HT} = \widetilde{y}_{HT}/\widetilde{w}$, where $\widetilde{w} = \sum_{i=1}^{n}\sum_{k=1}^{m_i} w_{ik}$ is the total of weights. Let the school weights be $w_{i\cdot} = \frac{1}{\pi_i}$ $(i = 1, 2, \ldots, n)$, the reciprocal of $\pi_i$. Although $\bar{y}_{HT}$ is biased, this bias vanishes with increasing sample size $n$ to the order $O(1/n)$ (Cochran, 1977).

### The Variance of a Horvitz–Thompson Estimator of Total

The variance formula of $\widetilde{y}_{HT}$:

$$V\left(\widetilde{y}_{HT}\right) = \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\pi_i\pi_j - \pi_{ij}\right)\left(\frac{\widetilde{Y}_i}{\pi_i} - \frac{\widetilde{Y}_j}{\pi_j}\right)^2 + \sum_{i=1}^{N}\frac{V\left(\widetilde{w}_i\right)}{\pi_i}, \tag{3}$$

where

$$V\left(\widetilde{w}_i\right) = \frac{M_i^2\left(1 - f_{2i}\right)}{m_i}S_{2i}^2,$$

and the variance $S_{2i}^2 = \left(M_i - 1\right)^{-1}\sum_{k=1}^{M_i}\left(y_{ik} - \overline{Y}_{i\cdot}\right)^2$ with $\overline{Y}_{i\cdot} = M_i^{-1}\sum_{k=1}^{M_i} y_{ik}$. The expression is in the Sen-Yates-Grundy (SYG) form of the variance of $\widetilde{y}_{HT}$ (Cochran, 1977, p. 308; Sen, 1953; Yates & Grundy, 1953). The estimate of $V\left(\widetilde{y}_{HT}\right)$ is

$$v\left(\widetilde{y}_{HT}\right) = \sum_{i=1}^{n}\sum_{j>i}^{n}\frac{\left(\pi_i\pi_j - \pi_{ij}\right)}{\pi_{ij}}\left(\frac{\widetilde{y}_i}{\pi_i} - \frac{\widetilde{y}_j}{\pi_j}\right)^2 + \sum_{i=1}^{n}\frac{v\left(\widetilde{w}_i\right)}{\pi_i}, \tag{4}$$

where

$$v\left(\widetilde{w}_i\right) = \frac{M_i^2\left(1 - f_{2i}\right)}{m_i}s_{2i}^2,$$

and the weighted variance is $s_{2i}^2 = \left(m_i^{-1}\left(m_i - 1\right)\sum_{k=1}^{m_i} w_{ik}\right)^{-1}\sum_{k=1}^{m_i} w_{ik}\left(y_{ik} - \bar{y}_{i\cdot}\right)^2$ with $\bar{y}_{i\cdot} = \sum_{k=1}^{m_i} w_{ik}y_{ik}/\sum_{k=1}^{m_i} w_{ik}$ (Cochran, 1977, p. 301). The computation of the formula $v\left(\widetilde{y}_{HT}\right)$ in Equation 4 requires knowledge of the joint inclusion probabilities $\pi_{ij}$ in $n \times n$ matrix of the $\pi_{ij}$'s. Because $\pi_{ij}$ $(i \,\&\, j = 1, 2, \ldots, n)$ are usually unavailable, Hájek approximations are used to approximate these probabilities in this analysis. Several Hájek approximations are introduced in the next section. Note that the term $\pi_i\pi_j - \pi_{ij}$ in the variance estimates of the H–T estimator can be negative because of the variability in its values.

### The Variance of a Horvitz–Thompson Estimator of the Mean

The variance of the mean estimator can be estimated with a Taylor approximation. Using vector symbols, let $\xi = \left(\widetilde{y}_{HT}, \widetilde{w}\right)'$ and $\Xi = \left(\widetilde{Y}, \widetilde{W}\right)'$, $g(\xi) = \widetilde{y}_{HT}/\widetilde{w}$, and $g(\Xi) = \widetilde{Y}/\widetilde{w}$. Let

$$\frac{\partial g(\Xi)}{\partial \xi} = \left(\frac{\partial g(\Xi)}{\partial \widetilde{y}_{HT}}, \frac{\partial g(\Xi)}{\partial \widetilde{w}}\right) = \left(\frac{1}{\widetilde{w}}, -\frac{\widetilde{y}_{HT}}{\widetilde{w}^2}\right). \tag{5}$$

Let symbols $\dot{\xi} = \left(\dot{y}_{HT}, \dot{w}\right)$ be a point between $\boldsymbol{\xi}$ and $\boldsymbol{\Xi}$. The first-order Taylor expansion of $g(\boldsymbol{\xi})$ with the mean value form of the remainder is

$$g(\xi) - g(\Xi) = (\xi - \Xi)'\frac{\partial g(\Xi)}{\partial \xi} + \frac{1}{2!}(\xi - \Xi)'\frac{\partial^2 g\left(\dot{\Xi}\right)}{\partial \xi^2}(\xi - \Xi). \tag{6}$$

Then the mean square error of $\boldsymbol{\xi}$ is

$$E\left(g(\xi) - g(\Xi)\right)^2 = \left(\frac{\partial g(\Xi)}{\partial \xi}\right)' V(\xi)\frac{\partial g(\Xi)}{\partial \xi} = \overline{Y}_{HT}^2\left(\frac{V\left(\bar{y}_{HT}\right)}{\widetilde{Y}_{HT}^2} - 2\frac{\mathrm{Cov}\left(\bar{y}_{HT}, \widetilde{w}\right)}{\widetilde{Y}_{HT}\widetilde{W}} + \frac{V\left(\widetilde{w}\right)}{\widetilde{W}^2}\right) + O\left(n^{-2}\right). \tag{7}$$

Thus, the derived formula of $V\left(\overline{y}_{HT}\right)$ will be consistent with those provided by Cochran (1977, p. 155)

$$V\left(\overline{y}_{HT}\right) \approx \overline{Y}_{HT}^{2}\left(\frac{V\left(\overline{y}_{HT}\right)}{\widetilde{Y}_{HT}^{2}} - 2\frac{\mathrm{Cov}\left(\overline{y}_{HT}, \widetilde{w}\right)}{\widetilde{Y}_{HT}\widetilde{W}} + \frac{V\left(\widetilde{w}\right)}{\widetilde{W}^{2}}\right), \tag{8}$$

which has order $O(n^{-1})$. The variance of $V\left(\overline{y}_{HT}\right)$ can be estimated by

$$v\left(\overline{y}_{HT}\right) = \overline{y}_{HT}^{2}\left(\frac{v\left(\widetilde{y}_{HT}\right)}{\widetilde{y}_{HT}^{2}} - 2\frac{\mathrm{cov}\left(\widetilde{y}_{HT}\right)}{\widetilde{y}_{HT}\widetilde{w}} + \frac{v\left(\widetilde{w}\right)}{\widetilde{w}^{2}}\right), \tag{9}$$

in which the variance terms $v\left(\widetilde{y}_{HT}\right)$ and $v\left(\widetilde{w}\right)$ are the same as in Equation 4. The derivation of the delta method is based on the Taylor series expansion with FPC factored in. Therefore, the variance estimates yielded by the delta method are used as the criterion in comparison with those yielded by the jackknifed approach for complex data drawn from a finite population.

## The Joint Inclusion Probabilities and Hájek Approximation

In application of the delta method, the computation of the formula $v\left(\widetilde{y}_{HT}\right)$ in Equation 4 requires knowledge of the joint inclusion probabilities $\pi_{ij}$ in $n \times n$ matrix of the $\pi_{ij}$'s. As mentioned, these probabilities are often unavailable. Hájek (1964) proposed an approximation of the joint inclusion probabilities of $\pi_{ij}$. Let $c_i = \pi_i(1-\pi_i)$ and $c = \sum_{i=1}^{N} c_i$. Hájek's approximation is defined as

$$\widehat{\pi}_{H,ij} = \pi_i\pi_j\left(1 - \frac{\left(1-\pi_i\right)\left(1-\pi_j\right)}{c}\right). \tag{10}$$

The $c$ in Equation 10 can be estimated by using the H – T estimator

$$\widehat{c}_1 = \sum_{i=1}^{n}\frac{c_i}{\pi_i} = \sum_{i=1}^{n}\left(1-\pi_i\right) = n - \sum_{i=1}^{n}\pi_i. \tag{11}$$

Another form of approximation is Hájek's $d$:

$$\widehat{d}_{\text{Hájek},ij} = \left(\pi_i\pi_j - \widehat{\pi}_{H,ij}\right)/\widehat{\pi}_{H,ij} \ \left(i\&j = 1, 2, \ \ldots \ , n\right).$$

Based on the large-sample properties of $(\pi_i\pi_j - \pi_{ij})/\pi_i\pi_j$, the Hájek *lower bound* (HB) of the estimator of $\widehat{d}_{\text{Hájek},ij}$ is

$$\widehat{d}_{\text{HB},ij} = \frac{1-\delta}{n-1} \tag{12}$$

(see Appendix for details), and $\delta$ is the function of $\pi_i$, $\pi_j$, and $\pi_{ij}$ with value range (0, 1). Based on Equation 12, we can examine how variance estimation changes correspondingly when the variable $\delta$ changes. Clearly, if the school inclusion probabilities are close to each other, for example equal to $c_0$, by Equation 12, then $\widehat{d}_{\text{HB},ij} \approx \left(1-c_0\right)/(n-1)$.

There are three proposed estimators of $\delta$: the arithmetic mean $\delta_{\text{Arith},\,ij} = (\pi_i + \pi_j)/2$, the geometric mean $\delta_{\text{Geo},ij} = \sqrt{\pi_i\pi_j}$, and $\delta_{\text{RR},\,ij} = min(\pi_{i,}\,\pi_j)$.See Qian (2017) and Rizzo and Rust (2011). Therefore, there are three estimators:

$$\widehat{d}_{\text{Arith},ij} = \frac{1-\left(\pi_i+\pi_j\right)/2}{n-1}, \tag{13}$$

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

$$\widehat{d}_{\text{RR},ij} = \frac{1-\min\left(\pi_{i,}\pi_j\right)}{n-1}. \tag{15}$$

Compared with $\widehat{d}_{\text{Arith},ij}$, $\widehat{d}_{\text{Geo},ij}$ is smaller; $\widehat{d}_{\text{RR},ij}$ is the least among three.

## Results

### The Empirical Curve of the Variance Estimates Yielded by the Delta Method

For the five state samples of the NAEP 2009 state science assessments, both the delta method and the jackknife approach were used to estimate the variance of the subscale scores for the same sets of complex data. The focus was on whether the grouped jackknife approach yielded larger variance estimates than those of the delta method due to absence of the FPC effect. In applying the delta method, three different joint probability approximations, such as the δ estimators represented in Equations 13 – 15, were employed. The variance of the composite scores, a weighted average of three subscale means, was derived based on the variance formula for a stratified sample (Cochran, 1977, p. 92). Table 2 presents the means and standard deviations of the three types of δ estimators for the five state samples. The overall means of $\widehat{d}_{\text{Arith},ij}$, $\widehat{d}_{\text{Geo},ij}$, and $\widehat{d}_{\text{RR},ij}$ were 0.431, 0.388, and 0.284, respectively. Obviously, $\widehat{d}_{\text{Geo},ij}$ was smaller than $\widehat{d}_{\text{Arith},ij}$, and $\widehat{d}_{\text{RR},ij}$ was the least. Although the range of δ is (0, 1), based on Table 2, the empirical range of δ should be between 0.15 and 0.55.

Based on Equation 12, the examination was focused on how the variance estimated by the delta method changes correspondingly when the variable δ changes. The delta standard error estimates were then compared with those yielded by the jackknife approach and examined as to whether the jackknifed variances were larger. Figure 1 presents the standard error curves estimated by the delta method with δ increased from 0 to 1 in Equation 12 and the horizontal lines of the jackknifed standard errors for the five state samples. When $\delta = 1$, the approximation $\widehat{d}_{\text{HB},ij} = (1 - \delta) / (n - 1)$ in Equation 12 is at its minimum, which implies that the standard error estimate contains only the term of within-school errors. On the contrary, when $\delta = 0$, $\widehat{d}_{\text{HB},ij}$ reaches its maximum. There is a total of 21 estimates of standard errors on each curve as δ takes the values from 0 to 1 with an interval of .05.

The sampling rates of schools for States 1 – 3 were all above 6%, higher than those for States 4 and 5. Their standard error lines of the jackknife approach were above the standard error curves of the delta method, meaning all jackknifed standard errors were larger than the standard error estimates yielded by the delta method in the entire range of δ. These standard errors without the FPC effect for States 1 – 3 were consistent with those reported in the literature (Bellhouse, 1985; Kali et al., 2011; Wolter, 2007). The comparisons in this study were conducted using empirical analyses, and the true variance is unknown. The trend of these results can be assessed together with those of the delta method to illustrate how the variance estimation changes corresponding to the levels of the Hájek approximation. In this study, the cutoff points were determined based on a 6% sampling rate; however, this is not a golden criterion that can be generalized to populations with different characteristics. It is necessary to adjust the cutoff points according to the specific conditions of a population. For example, the cutoff points for a population of primary schools must differ from those for a population of colleges.

However, for State 4, the standard error line of the jackknife approach was below the delta standard error curve; that is, the jackknifed standard error was smaller than the delta standard errors. The school sampling rate for State 4 was small, 3.53%, and its school population size was large, so the FPC effects were small. This can be one of the main reasons for the small jackknifed standard error for State 4, and homogeneity among school clusters can be another one. For State 5, the standard error curve of the delta method was slightly above the line of the jackknifed standard error on the left segment and was below the line on the right. Similar to State 4, State 5 also had a large school population and a small sampling rate of 4.92%. By the empirical results of this educational assessment survey, 6% or less can be treated as a threshold of a small sampling rate.

**Table 2** The Mean and Standard Deviation (*SD*) of the Three Types of δ Estimators for the Five State Samples

| | $\delta_{Arith}$ | | $\delta_{Geo}$ | | $\delta_{RR}$ | |
|---|---|---|---|---|---|---|
| State | Mean | *SD* | Mean | *SD* | Mean | *SD* |
| 1 | 0.3400 | 0.1407 | 0.3134 | 0.1306 | 0.2319 | 0.1184 |
| 2 | 0.5533 | 0.1937 | 0.5177 | 0.1937 | 0.3971 | 0.2016 |
| 3 | 0.4112 | 0.1587 | 0.3792 | 0.1577 | 0.2842 | 0.1567 |
| 4 | 0.4516 | 0.2673 | 0.3724 | 0.2537 | 0.2525 | 0.2593 |
| 5 | 0.4003 | 0.2007 | 0.3584 | 0.1763 | 0.2561 | 0.1535 |
| Average | 0.4313 | | 0.3882 | | 0.2843 | |

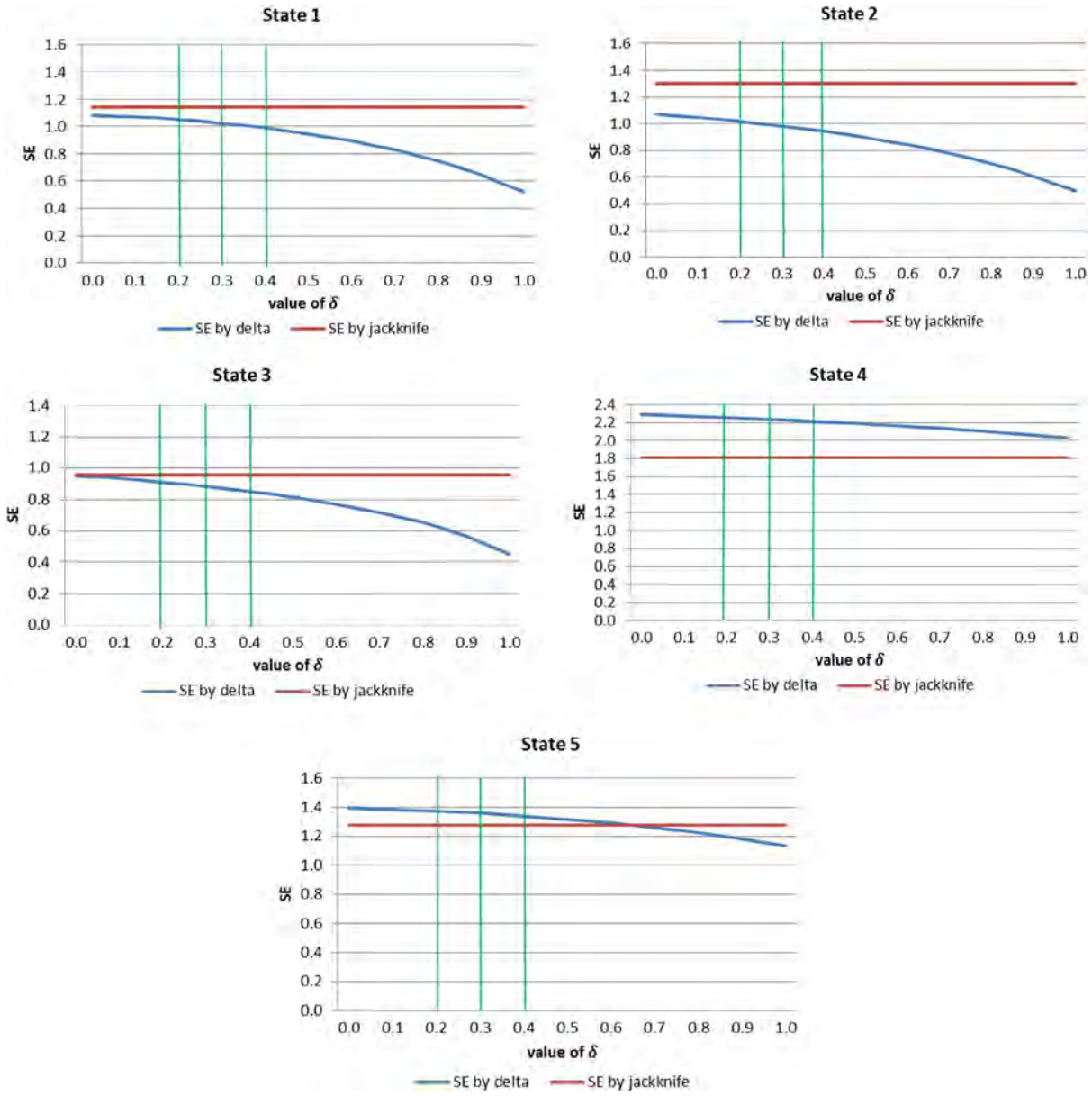


**Figure 1** The standard errors of mean estimated by the delta method and jackknife approach, States 1 – 5.

In Figure 1, the average ratios of the jackknifed standard error to the delta standard error were 1.03, 1.05, and 1.09 at three δ value points of 0.2, 0.3, and 0.4, respectively. These comparisons showed that the jackknifing procedure without a built-in FPC effect tended to yield standard errors that were about 6% larger than those produced by the delta method. The findings, in average across the states, suggest that for a sample from a finite population with a high sampling rate, it is necessary to incorporate the FPC effect into the jackknife approach. There are several approaches to incorporating the FPC effects in the jackknifing procedure, such as applying the scaling factor (Wolter, 2007) or using the adapted replicate weights proposed by Kali et al. (2011) and Rizzo and Rust (2011).

## The Paired Jackknife Approach for Variance Estimation With Adjusted Group Weights

For some grouped jackknifing procedures, the group weights used in forming replicate samples can be adapted to incorporate the FPC effect. Beginning in 2011, the NAEP paired grouped jackknife approach has incorporated the FPC effect (Kali et al., 2011; Rizzo & Rust, 2011). For a two-stage PPS sampling without replacement, the details of forming replicate

**Table 3** The Standard Error (*SE*) of the Mean of Subscales and Mean Score Yielded by the Delta Method and the Jackknife Approach Based on Different Hájek's $d_{*_{ij}}$ Estimates for the Five State Samples

| State | Sampling rate of schools (%) | *SE* of physical science | *SE* of earth science | *SE* of life science | *SE* of mean composite |
|---|---|---|---|---|---|
| | | $\widehat{d}_{\text{Arith},ij}$ | | | |
| 1 | 6.62 | 1.501 | 1.493 | 1.572 | 0.894 |
| 2 | 10.11 | 1.419 | 1.396 | 1.400 | 0.819 |
| 3 | 6.10 | 1.422 | 1.338 | 1.353 | 0.798 |
| 4 | 3.53 | 3.832 | 3.691 | 3.755 | 2.192 |
| 5 | 4.92 | 2.241 | 2.251 | 2.220 | 1.305 |
| | | $\widehat{d}_{\text{Geo},ij}$ | | | |
| 1 | 6.62 | 1.681 | 1.666 | 1.766 | 1.002 |
| 2 | 10.11 | 1.744 | 1.714 | 1.702 | 1.001 |
| 3 | 6.10 | 1.575 | 1.481 | 1.494 | 0.882 |
| 4 | 3.53 | 3.860 | 3.721 | 3.789 | 2.210 |
| 5 | 4.92 | 2.260 | 2.279 | 2.249 | 1.319 |
| | | $\widehat{d}_{\text{RR},ij}$ | | | |
| 1 | 6.62 | 1.820 | 1.802 | 1.915 | 1.085 |
| 2 | 10.11 | 1.913 | 1.879 | 1.858 | 1.095 |
| 3 | 6.10 | 1.685 | 1.580 | 1.591 | 0.941 |
| 4 | 3.53 | 3.895 | 3.758 | 3.829 | 2.232 |
| 5 | 4.92 | 2.388 | 2.414 | 2.384 | 1.396 |

sample $r$ from school pairs can be found in the Grouped Jackknifing Procedure and Variance Estimation section. Let $A_r$ be the set of retained students and $D_r$ be the set of deleted students for replicate sample $r$ ($r = 1, 2, \ldots, R$). The replicate weight for school $i$ is

$$w_i(r) = \begin{cases} w_i\left(1 + \sqrt{\widehat{d}_{\text{Hájek},ij}}\right) & \text{i} \in A_r \\ w_i\left(1 - \sqrt{\widehat{d}_{\text{Hájek},ij}}\right) & i \in D_r \\ w_i \text{ otherwise} \end{cases}, \tag{16}$$

where $\widehat{d}_{\text{Hájek},ij}$ is following Equation 11 and $j$ refers to the dropped school in forming a replicate sample (Rizzo & Rust, 2011). Note that, without considering FPC effects, the school replicate weight $w_i(r)$ would be $2 \cdot w_i$.

In the proposed paradigm, the delta method approximates the standard error based on a Taylor series expansion. Although the approximation can offset the accuracy of the results to a certain degree, its variance formula has the order $O(1/n)$.See Cochran (1977, p. 155); the error is usually much smaller than the sampling rate $n/N$.

In addition, there can be other factors confounding the results. For example, the schools in a sample frame are usually sorted, either explicitly or implicitly. Although the sorting effects on variance estimation have been discussed (Cochran, 1977; Kish, 1965), such sorting effects on FPC are unknown and remain to be explored. For the paired grouped jackknife, because the schools in a pair are similar and the pairs of similar schools are treated as strata, the sorting effects on paired JRR are expected to be very limited.

## Impact of Different Approximations of the Joint Probability on the Delta Method

Table 3 presents the standard errors of mean scores yielded by the delta method and the standard errors of the composite scores based on the above $\widehat{d}_{ij}$ estimates in Equations 13 – 15. As expected, the standard errors estimated with H – T estimators and the delta method $\widehat{d}_{\text{Arith},ij}$ were the smallest, and those with $\widehat{d}_{\text{RR},ij}$ were the largest. Compared with those estimated with $\widehat{d}_{\text{RR},ij}$, measured by relative absolute errors, the standard errors estimated using $\widehat{d}_{\text{Geo},ij}$ were about 5.5, 3.5, and 3.8% smaller, whereas those using $\widehat{d}_{\text{Arith},ij}$ were about 13.3, 9.0, and 9.5% smaller.

The results confirmed that the approximation $\widehat{d}_{\text{Arith},ij}$ is the smallest among the three. Thus, using $\widehat{d}_{\text{Arith},ij}$ can underestimate the standard errors because the Hájek approximation $\delta_{Arith,\,ij}$ would overestimate the joint inclusion probabilities,

in particular when there are large differences between two school sizes (Qian, 2017). In contrast, the approximation $\widehat{d}_{\mathrm{RR},ij}$ can cause overestimation in the standard error under the same circumstances. These can also be observed from the plots in Figure 1. Because the value of $\widehat{d}_{\mathrm{Geo},ij}$ is always between $\widehat{d}_{\mathrm{Arith},ij}$ and $\widehat{d}_{\mathrm{RR},ij}$, the delta standard errors yielded by the approximation $\widehat{d}_{\mathrm{Geo},ij}$ were larger than those by $\widehat{d}_{\mathrm{Arith},ij}$ but smaller than those by $\widehat{d}_{\mathrm{RR},ij}$, as shown in Table 3. However, to validate the true adequacy and robustness of the standard errors, we would need to conduct a simulation study, which is beyond the scope of the current study.

## Summary

The proposed paradigm is to study the effects of ignoring the FPC effect on the jackknife approach. The jackknifed variance estimates were compared with those estimated by the delta method. Based on the large-sample properties of the Hájek approximation, the examination was focused on how variance estimation changes correspondingly when the variable δ (in Equation 12) changes. Although this study employed the jackknife approach, the proposed paradigm can be readily extended to comparisons using other resampling methods such as BRR and bootstrap. Moreover, the paradigm can be applied to other types of complex data, such as one-stage unequal sampling with unequal probabilities without replacement. Correspondingly, instead of using Equation 4 in the Variance of a Horwitz–Thompson Estimator of Total section, the formula used in the delta method needs to be replaced with the equations in 9A.43 and 9A.44 in Cochran (1977, p. 261).

The empirical results showed that the jackknifing procedure, with a missing FPC effect, tends to yield larger standard errors than those yielded by the delta method. For three samples with the school sampling rate greater than 6%, as shown in the plots of States 1–3 in Figure 1, the jackknifed variances were evidently larger than those of the delta method. For two states with the school sampling rates less than 6%, mixed results of the jackknifed variances showed no clear tendency. In general, compared with the delta standard errors, the average of the jackknifed standard erros for means was also greater. The empirical results revealed the relationship between the school sampling rate and the FPC effect. Other factors can also have an impact on variance estimation—for example, clustering effects (Cochran, 1977). The impact of clustering effects is beyond the scope of this study because clustering effects have no direct relationship with FPC, as the FPC effect is due to the sampling without replacement from a finite population.

The empirical results revealed that the delta method can be used to yield appropriate standard error estimates if the joint inclusion probabilities are estimated by proper Hájek approximation. In this study, the H–T estimator was proposed to estimate the $c$ in Equation 11 and A2; the variance of the estimator $\widehat{c}_1$ was also derived (see Appendix). For three approximations, $\widehat{d}_{\mathrm{Arith},ij}$, $\widehat{d}_{\mathrm{Geo},ij}$, and $\widehat{d}_{\mathrm{RR},ij}$ in Equations 13–15, $\widehat{d}_{\mathrm{Arith},ij}$ tended to yield smaller SEs because $\delta_{Arith,\ ij}$ would overestimate the joint inclusion probabilities when there is a huge discrepancy between school sizes (i.e., difference in school inclusion probabilities). On the contrary, when there are large differences between two school sizes, $\widehat{d}_{\mathrm{RR},ij}$ tends to yield larger standard errors because $\delta_{RR,\ ij}$ would underestimate the joint inclusion probabilities. The results show that the delta method provides robust results when applying approximation $\widehat{d}_{\mathrm{Geo},ij}$, even if the school size differences in school pairs are large in the sample.

The results suggest that the delta method is certainly a proper method for variance estimation in analyzing complex data without replacement. The jackknife approach can also be used, in particular for samples with a small sampling rate. For samples with a large sampling rate, to avoid overestimation for the jackknife approach, it is necessary to consider methods including certain scaling factors related to FPC, as discussed in the the Empirical Curve of the Variance Estimates Yielded by the Delta Method section.

To validate the findings in the current study, a more exhaustive simulation is needed for future studies: in particular, a simulation to aid in choosing the appropriate Hájek approximation in applying the delta method and finding the empirical formula of scaling factors for the jackknife approach. In a future simulation study, it would be interesting to apply the paradigm to compare the jackknifed standard errors using the adapted school weights (in Equation 16) with those using the original weights and to further evaluate whether the jackknifed standard errors with the adapted weights or the delta standard errors are better aligned with true standard errors.

In addition, this study is based on the NAEP state data drawn with systematic PPS sampling. Nevertheless, the variance formula (in Equation 3) for the delta method was derived based on a strict two-stage PPS sampling. Although the PPS sample without replacement and the systematic PPS sample are equivalent (Kish, 1965; Rust et al., 2001), to obtain

theoretically accurate results for the delta method requires a reconfirmation of the findings based on strict PPS samples without replacement, which can be drawn by the rejective sampling approach (Hájek, 1964). This is a demanding task that needs vast resources to accomplish.

## Acknowledgments

The author thanks Matthew Johnson, Rebecca Zwick, Yue Jia, Bruce Kaplan, Shuhong Li, Shelby Haberman, and Keith Rust for their suggestions and comments. The author thanks David Freund and Scott Davis for their assistance in assembling data. The author also thanks Kim Fryer and Ayleen Gontz for their editorial help. Any opinions expressed in this paper are those of the author and not necessarily those of Educational Testing Service.

## References

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

Bellhouse, D. R. (1985). Computing methods for variance estimation in complex surveys. *Journal of Official Statistics*, *1*(3), 323–329.

Berger, Y. G. (2004). A simple variance estimator for unequal probability sampling without replacement. *Journal of Applied Statistics*, *31*(3), 305–315. https://doi.org/10.1080/0266476042000184046

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). Wiley.

Fay, R. E. (1984). Some properties of estimates of variance based on replication methods. *JSM proceedings: Survey research methods section* (pp. 495–500). American Statistical Association.

Fay, R. E. (1989). Theory and application of replicate weighting for variance calculations. *JSM proceedings: Survey research methods section* (pp. 212–218). American Statistical Association.

Fuller, W. A. (1996). *Introduction to statistical time series*. Wiley. https://doi.org/10.1002/9780470316917

Fuller, W. A. (1998). Replicate variance estimation for two phase samples. *Statistica Sinica*, *8*, 1153–1164.

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics*, *40*(3), 254–273. https://doi.org/10.3102/1076998615574772

Haberman, S. J., Lee, Y., & Qian, J. (2009). *Jackknifing techniques for evaluation of equating accuracy* (Research Report No. RR-09-39). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02196.x

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, *35*(4), 1491–1523. https://doi.org/10.1214/aoms/1177700375

Hájek, J. (1981). *Sampling from a finite population*. Dekker.

Hartley, H. O., & Rao, J. N. K. (1962). Sampling with unequal probabilities and without replacement. *The Annals of Mathematical Statistics*, *33*(2), 350–374. https://doi.org/10.1214/aoms/1177704564

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*. Wiley.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 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). *JSM proceedings: Survey research methods section*. American Statistical Association.

Kalton, G. (1979). Ultimate cluster sampling. *Journal of the Royal Statistical Society. Series A (General)*, *142*(2), 210–222. https://doi.org/10.2307/2345081

Kalton, G. (2002). Models in the practice of survey sampling (revisited). *Journal of Official Statistics*, *18*(2), 129–154.

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

Lee, H., Rancourt, E., & Särndal, C.-E. (1995). Variance estimation in the presence of imputed data for the generalized estimation system. *JSM proceedings: Survey research methods section* (pp. 384–389). American Statistical Association.

Mosteller, F., & Tukey, J. (1977). *Data analysis and regression*. Addison-Wesley.

National Center for Education Statistics. (2011). *The nation's report card: Science 2009* (NCES 2011-451). U.S. Department of Education, Institute of Education Sciences.

Neidorf, T. S., Binkley, M., Gattis, K., & Nohara, D. (2006). *Comparing mathematics content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Programme for International Student Assessment (PISA) 2003 assessments* (NCES 2006-029). U.S. Department of Education, National Center for Education Statistics.

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 2001-07). U.S. Department of Education, National Center for Education Statistics.

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

Qian, J., Jiang, Y., & von Davier, A. (2013). *Weighting test samples in IRT linking and equating: Toward an improved sampling design for complex equating* (Research Report No. RR-13-39). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02346.x

Rao, J. N. K., & Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, *79*(4), 811–822. https://doi.org/10.1093/biomet/79.4.811

Rao, J. N. K., & Sitter, R. R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, *82*(2), 453–460. https://doi.org/10.1093/biomet/82.2.453

Rao, J. N. K., & Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, *83*(401), 231–241. https://doi.org/10.1080/01621459.1988.10478591

Rizzo, L., & Judkins, D. (2004). Replicate variance estimation for the National Survey of Parents and Youths. *JSM proceedings: Survey research methods section* (pp. 4257–4263). American Statistical Association.

Rizzo, L., & Rust, K. (2011). Finite population correction for NAEP variance estimation. *JSM proceedings: Survey research methods section* (pp. 2501–2515). American Statistical Association.

Rust, K. F. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, *1*(4), 381–397.

Rust, K. F., Wallace, L., & Qian, J. (2001). Sample design for the state assessment. In N. L. Allen, J. R. Donoghue, & T. L. Schoeps (Eds.), *The NAEP 1998 technical report* (NCES 2001-09; pp. 61–77). National Center for Education Statistics.

Särndal, C. E., Swenson, B., & Wretman, J. (1992). *Model assisted survey sampling*. Springer-Verlag.

Sen, A. R. (1953). On the estimate of variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, *5*(1), 119–127.

Steel, P., & Fay, R. E. (1995). Variance estimation for finite populations with imputed data. *JSM proceedings: Survey research methods section* (pp. 374–379). American Statistical Association.

Wolter, K. (2007). *Introduction to variance estimation*. Springer.

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 (Methodological)*, *15*(2), 253–261. https://doi.org/10.1111/j.2517-6161.1953.tb00140.x

## Appendix

## Hájek Approximation of the Joint Inclusion Probabilities

The formula $v\left(\widetilde{y}_{HT}\right)$, in Equation 4, requires knowledge of the joint inclusion probabilities $\pi_{ij}$, (i.e., an $n \times n$ matrix of the $\pi_{ij}$'s). Let $c_i = \pi_i(1-\pi_i)$ and $c = \sum_{i=1}^{N} c_i$. Based on rejective sampling, Hájek (1981, p. 75) provided an asymptotically valid approximation of $\pi_{ij}$ :

$$\widehat{\pi}_{H,ij} \approx \pi_i \pi_j \left( 1 - \frac{\left(1-\pi_i\right)\left(1-\pi_j\right)}{c} \right). \tag{A1}$$

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 $\pi_i\pi_j - \widehat{\pi}_{H,ij}$ 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 $\pi_N$; however, most of the assessment data sets contain only $\pi_1, \pi_2, \ldots,$ and $\pi_n$. Thus, the parameter $c$ cannot be computed directly and must be estimated.

One method of estimation uses the H – T estimator:

$$\widehat{c}_1 = \sum_{i=1}^{n} \frac{c_i}{\pi_i} = \sum_{i=1}^{n} \left(1 - \pi_i\right) = n - \sum_{i=1}^{n} \pi_i. \tag{A2}$$

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

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

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

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

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

$$\widehat{\pi}_{\widehat{c}_1,ij} = \pi_i\pi_j \left[1 - \frac{\left(1 - \pi_i\right)\left(1 - \pi_j\right)}{\widehat{c}_1}\right]. \tag{A3}$$

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

$$\widehat{c}_2 = \frac{N}{n} \sum_{i=1}^{n} c_i; \tag{A4}$$

thus

$$\widehat{\pi}_{\widehat{c}_2,ij} = \pi_i\pi_j \left[1 - \frac{\left(1 - \pi_i\right)\left(1 - \pi_j\right)}{\widehat{c}_2}\right]. \tag{A5}$$

### 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 A3 and Equation A5, 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

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

Moreover, instead of estimating $\pi_{ij}$, 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

$$\frac{n\left(\pi_i\pi_j - \pi_{ij}\right)}{\pi_i\pi_j} \to 1, \tag{A6}$$

when $N \to \infty$ while $n$ is fixed, and there is one set of the Hartley–Rao conditions (Hájek, 1964, p. 1495; Hartley & Rao, 1962). The property in Equation A6 is false if $\sum_{i=1}^{N} \pi_i \left(1 - \pi_i\right) \to \infty$; the formulae based on Equation A6 are applicable if $N$ is much larger than $n$ (Hájek, 1964, p. 1496). The large-sample property (in Equation A6) implies

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

Therefore, the form

$$\widehat{d}_{\mathrm{HB},ij} = \frac{1-\delta}{n-1}, \tag{A7}$$

(Rizzo & Rust, 2011) can be treated as the *lower bound* of the estimator of $\widehat{d}_{\mathrm{Hájek},ij}$, where $\delta$ is a small positive number and can be a function of $\pi_i$, $\pi_j$, and $\pi_{ij}$. Although $\widehat{d}_{\mathrm{Hájek},ij}$ can also be expressed in the form $1/(\gamma - 1)$ with $\gamma$ having a value range $(n, \infty)$, it is more straightforward to discuss modified estimators in the form of Equation A7. Note that, in application, Equation A7 is applicable only if $N$ is much larger than $n$, which may not be true for a sample drawn from a small state.

The empirical results in the Results section show that $\delta$ can be estimated adequately by the geometric mean and the arithmetic mean of $\pi_i$ and $\pi_j$:

$$\widehat{d}_{\mathrm{Arith},ij} = \frac{1 - \left(\pi_i + \pi_j\right)/2}{n-1}, \tag{A8}$$

and

$$\widehat{d}_{\mathrm{Geo},ij} = \frac{1 - \sqrt{\pi_i \pi_j}}{n-1}. \tag{A9}$$

Compared with $\widehat{d}_{\mathrm{Arith},ij}$, the estimator $\widehat{d}_{\mathrm{Geo},ij}$ is smaller. The approximation of $\widehat{d}_{\mathrm{HR},ij}$ with $\eta = \min(\pi_i, \pi_j)$ is

$$\widehat{d}_{\mathrm{RR},ij} = \frac{1 - \min\left(\pi_i, \pi_j\right)}{n-1}, \tag{A10}$$

(Rizzo & Rust, 2011). Compared with school inclusion probabilities, if the joint inclusion probabilities are very small, the estimation of $\widehat{d}_{\mathrm{RR},ij}$ can be conservative and have variance overestimated. Clearly, if the probabilities of school inclusion are close to each other, perhaps equal to $c_0$,then $\widehat{d}_{\mathrm{HB},ij} \approx \left(1 - c_0\right)/(n-1)$. Because the NAEP school selection was based on systematic sampling, the joint inclusion probabilities can be very small, and thus $\widehat{d}_{\mathrm{RR},ij}$ can be overestimated.

**Suggested citation:**

Qian, J. (2020). *Variance estimation with complex data and finite population correction – A paradigm for comparing jackknife and formula-based methods for variance estimation* (Research Report No. RR-20-11). Educational Testing Service. https://doi.org/10.1002/ets2.12294

**Action Editor:** Rebecca Zwick

**Reviewers:** Yue Jia