Variance Estimation for Random-Groups Linking in Large-Scale Survey Assessments

Bingchen Liu, Yue Jia, and John Mazzeo

Abstract The random-groups design is frequently used in equating and linking scores from two tests, in which the linking functions are derived from the test scores of two samples of the test-taker population. In this paper, we consider estimating variances of test score population statistics for large-scale survey assessments (LSAs), where the random-groups design is used in linking latent variable test scores. Examples of LSAs include National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Programme for International Student Assessment (PISA). In estimating variances of population statistics in LSAs, the common practice takes into account the uncertainties due to sampling and latency. In this paper, we propose a variance estimation method as an extension of the existing procedure that takes into account the random-groups linking. We illustrate the method using a NAEP dataset for which a linear linking function is used in linking test scores from a computer-based test to those from a paper-and-pencil test. The proposed method can be easily extended when random-groups equating and linking are applied to other assessment contexts, with linking functions being parametric or non-parametric.

B. Liu (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: bliu@ets.org

Y. Jia
Educational Testing Service, Princeton, NJ, USA
e-mail: yjia@ets.org

J. Mazzeo
(emeritus), Educational Testing Service, Princeton, NJ, USA
e-mail: mazzeo123@comcast.net

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_11

1 Introduction

In educational assessments, score linking is a general term that refers to relating scores from different tests or test forms (American Educational Research Association et al., 2014). This paper focuses on the random-groups linking in which one sample drawn from the population is administered one test form, while another sample drawn from the same population is administered a different test form. Based on the two samples selected from the common population, a linking function can be derived to transform the scores of one test form to the scores on the other test form (Kolen & Brennan, 2004).

Large-scale survey assessments (LSAs) are those used to monitor academic performance for populations (e.g., US fourth graders). One of the most important uses of LSAs is to track population statistics at a given time and changes in population statistics over time, such as how countries differ in students' mean scores on reading or how the mean reading scores in a country or a region change over time. Examples of LSAs include US National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Programme for International Student Assessment (PISA).

LSAs apply item response theory (IRT) latent variable regression models to directly estimate score distributional statistics for the population and subpopulations, such as population means and the percentage of students above specified proficiency levels (Mislevy, 1984, 1985). To provide a means to estimate population statistics, the programs also make available plausible values for individuals sampled from the population. Plausible values are random draws, or multiple imputations, from the performance distribution for individuals, conditional on the IRT latent regression model parameter estimates, response data, and contextual information (Mislevy, 1991; Braun & von Davier, 2017). In addition, LSAs make use of sampling weights to draw inferences from the probability-based samples to the population of interest. See, for example, von Davier et al. (2006) and Mazzeo (2018) on the design, sampling, and analysis of LSAs.

For LSAs, standard errors are estimated along with the population statistics. Typically, two general sources of variance are accounted for: sampling of test takers and latency of the test scores. The sampling variance accounts for the variability among the units in the population. The size of the sampling variance is in part a function of the sample design (see, e.g., Johnson & Rust, 1992). The latency variance reflects the uncertainty due to the statistics being estimated from the test-taker performance on a set of test questions and other auxiliary information used in the latent variable regression models. The latency variance is also referred to as between-imputation variance. Details on how these variances are estimated for LSAs are reviewed in Sect. 2.

One approach to estimate the sampling variance of a statistic is to use resampling methods such as the jackknife, balanced repeated replication (BRR), or bootstrap methods. These resampling methods create a number of subsamples and use the variability among the estimates from the subsamples to estimate the variance of

the statistic. An alternative approach is to linearize the statistic (e.g., using the delta method or Taylor series expansion) and then estimate the variance of the linearized statistic analytically. Wolter (2007) described both approaches. Kish and Frankel (1974) showed in simulation studies that using a multistage design with two primary sampling units per stratum, both the jackknife and BRR gave acceptably low bias in estimated variance for various statistics. They also showed that these two methods gave results that were similar to those achieved via the Taylor series linearization. Many theoretical and empirical studies have also supported that the resampling methods perform well and result in comparable standard error estimates as the linearization approach (e.g., Krewski & Rao, 1981; Rao & Wu, 1985; Valliant, 1990; Shao, 1996).

In this paper, we consider the random-groups design where a sample of test-takers (referred to as the target sample) is administered assessment T, while another sample (referred to as the source sample) is administered assessment S. The scores from assessments S and T are estimated on the two separate latent variable scales. In addition, the scores from assessment S are linked to assessment T via random-groups linking. One example is to link scores from a paper-and-pencil test to a test given on a computer (Eignor, 2007; Jewsbury et al., 2020). Other examples are the studies in linking scores between two different LSAs (Johnson, 1998; Johnson et al., 2005; Jia et al., 2011).

For the random-groups design, the linking function coefficients are statistics calculated based on the source and target samples and using the test scores that are subject to latency variance. Kolen and Brennan (2004) discussed the use of the bootstrap to estimate the sampling variance of statistics for assessments with the random-groups linking. However, we are not aware of any real-data applications.

When the random-groups design is applied in linking the LSA test scores, uncertainty in the linking function is typically ignored. Mazzeo et al. (in press) offered an approach to approximate the variance associated with the linking function, as an additional source of variance, adding to the sampling and latency variances typically estimated for the population statistics. Jewsbury (2019) derived analytic equations for variance estimation of population statistics such as averages, percentiles, and standard deviations. He suggested that the resampling methods might be more tractable in practice to cover a wide range of statistics. In this paper, we propose a variance estimation method that incorporates the uncertainty of the linking function into the sampling and latency variance estimates. The proposed method can be used to estimate variances for both linear and nonlinear statistics. Further, the method can be used when the two samples used in linking are either dependent or independent from each other.

In Sect. 2, we review the variance estimation approach currently used in LSAs. In Sect. 3, we introduce the new variance estimation method, which is an extension and modification of the existing method. We illustrate the method with a dataset from NAEP in Sect. 4. The conclusion follows in Sect. 5.

2 Variance Estimation in Large-Scale Survey Assessments

For complex survey data, analytical variance estimators for nonlinear statistics are difficult to develop, and some do not have a closed form. For LSAs, one common practice is to use the jackknife repeated replication (JRR) with replicate weights to estimate sampling variance. Several studies (Hansen et al., 1985; Kovar et al., 1988) have shown that JRR provides reasonable variance estimates for both linear and nonlinear statistics. Briefly, a total of H strata are formulated, and each replicate is created by excluding a random set of data in a stratum while keeping the remaining subset from that stratum and all the data in the other $H - 1$ strata. The replicate weights are then calculated for each of the H replicates which reflect the complex sample design. Those replicate weights also help protect the survey participants' information because the more detailed sampling information, such as stratification, primary sampling units (PSUs), clusters, etc., are not needed with the availability of the replicate weights. Details are provided in the next section. Applications include NAEP and TIMSS.

Using NAEP as an example, we now review how the sampling and latency variances are estimated. Let $\boldsymbol{W}_{\mathbf{orig}}$ represent the original sampling weights for the full sample, and $\boldsymbol{W}_j$ represent the jth set of jackknife replicate weights, $j = 1, 2, \ldots, N_r$, respectively, for a total of N_r sets of replicate weights. Further, let $\boldsymbol{v_i}$ denote the ith set of plausible values which is on an arbitrary IRT scale T, $i = 1, 2, \ldots, M$. Then the population statistic on scale T, denoted as $\hat{t}$ (e.g., population average score), can be calculated as

$$\hat{t} = \frac{\sum_{i=1}^{M} \hat{t}_i}{M} \tag{1}$$

where $\hat{t}_i$ is calculated using $\boldsymbol{v_i}$ with weight $\boldsymbol{W}_{\mathbf{orig}}$. The sampling variance of $\hat{t}$ is calculated as $\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N_r}\left(\hat{t}_{ij} - \hat{t}_i\right)^2$ where $\hat{t}_{ij}$ denotes the statistic calculated using $\boldsymbol{v_i}$ with replicate weight $\boldsymbol{W_j}$.

In practice, the sampling variance is often approximated based only on one set of plausible values to reduce computational burden. For example, using the first set of plausible values, the sampling variance can be estimated as

$$\widehat{\mathrm{Var}}_{\mathrm{samp}}\left(\hat{t}\right) = \sum_{j=1}^{N_r}\left(\hat{t}_{1j} - \hat{t}_1\right)^2 \tag{2}$$

Based on the work of Rubin (2004), the latency variance of $\hat{t}$ is estimated as follows:

$$\widehat{\mathrm{Var}}_{\mathrm{lat}}\left(\hat{t}\right) = \left(1 + \frac{1}{M}\right)\frac{\sum_{i=1}^{M}\left(\hat{t}_i - \hat{t}\right)^2}{M-1}. \tag{3}$$

The total variance for the statistic $\hat{t}$ is the sum of sampling and latency variances:

$$\widehat{\text{Var}}_{\text{total}}\left(\hat{t}\right) = \widehat{\text{Var}}_{\text{samp}}\left(\hat{t}\right) + \widehat{\text{Var}}_{\text{lat}}\left(\hat{t}\right) \tag{4}$$

3 Variance Estimation to Incorporate Uncertainty in Random-Groups Linking

As mentioned in Sect. 1, we consider that a target sample of test-takers is administered assessment T, while a source sample is administered assessment S. Assessment T results are on latent scale T, while assessment S results are on latent scale S. The objective is to apply a linear function to link assessment S results from scale S to scale T by aligning the mean and standard deviation (SD) of the sample taking assessment S to those of assessment T. For example, during the NAEP transition from paper-based assessment (PBA) to digitally based assessment (DBA), the sample who took the PBA is the target sample, and the sample who took the newly implemented DBA is the source sample. The linking function is then derived to link the DBA results to the latent scale for PBA, so that the DBA and PBA results can be compared.

For the source sample statistics that are linked to scale T, we propose a new resampling approach for variance estimation. Under the method, the variance consists of the sampling and measurement variance components, each taking into consideration the random-groups linking. We first discuss the JRR method for the estimation of the sampling variance that involves resampling both the target and source samples simultaneously and then the estimation of the latency variance. The proposed method is an extension of the method discussed in Sect. 2. The method works when the two samples are dependent or independent.

To be more specific, let:

- $\boldsymbol{x_i}$ represent the ith set of plausible values for the target sample on scale T,
- $\boldsymbol{\theta_i}$ represent the ith set of plausible values for the source sample on scale S,
- $\boldsymbol{y_i}$ represent the ith set of plausible values for the source sample that has been transformed to scale T, $i = 1, 2, \ldots, M$.

Further, let $\overline{\theta}_S$ and $\widehat{\sigma}_S$ denote the mean and SD of the source sample plausible values on scale S, weighted by $\boldsymbol{W}_{\mathbf{orig}}$, the original student sampling weights of the source sample. Similarly, let $\overline{X}_T$ and $\widehat{\sigma}_T$ denote the mean and SD of the target sample plausible values on scale T, weighted by $\boldsymbol{W}'_{\mathbf{orig}}$, the original student sampling weights of the target sample.

The coefficients $\hat{a}$ and $\hat{b}$ of the linear linking function are calculated as

$$\hat{a} = \frac{\widehat{\sigma}_T}{\widehat{\sigma}_S} \tag{5}$$

and

$$\hat{b} = \overline{X}_T - \hat{a}\overline{\theta}_S \tag{6}$$

Apply $\left(\hat{a}, \hat{b}\right)$ to transform $\boldsymbol{\theta_i}$ from scale S onto scale T:

$$\boldsymbol{y_i} = \hat{a}\boldsymbol{\theta_i} + \hat{b}, i = 1, 2, \ldots, M. \tag{7}$$

Last, we calculate the statistic $\hat{t}$ for the source sample on scale T using Eq. 1, with $\hat{t}_i$ being estimated using $\boldsymbol{y_i}$ with $\boldsymbol{W_{orig}}$, for $i = 1, 2, \ldots, M$.

In the text below, we describe the procedure in estimating the variance of the statistic $\hat{t}$.

3.1 Estimation of Sampling Variance

In this section, we describe the procedure used in estimating the sampling variance of the source sample statistic $\hat{t}$ as defined in Eq. 1, which is linked to scale T through random-groups linking. We further introduce the following notations $\boldsymbol{W_j}$, $\boldsymbol{W'_j}$, which represent the jth set of jackknife replicate weights of the source and target samples, respectively, $j = 1, 2, \ldots, N_r$. In the random-groups linking design, it is common that the two samples to be linked have the same number of replicate weights. Therefore, in our method, we assume the source and target samples have the same number of replicate weights (denoted as N_r here). To reduce the computational intensity, we use only the first set of plausible values from both samples for the calculation.

Using the jth pair of replicate weights $\left(\boldsymbol{W_j}, \boldsymbol{W'_j}\right), j = 1, 2, \ldots, N_r$, we conduct the following steps of calculation:

1. Compute $\overline{\theta}_{S_j}$ and $\widehat{\sigma}_{S_j}$, the mean and SD of the first set of plausible values for the source sample on scale S, weighted by $\boldsymbol{W_j}$, as well as $\overline{X}_{T_j}$ and $\widehat{\sigma}_{T_j}$, the mean and SD of the first set of plausible values for the target sample on scale T, weighted by $\boldsymbol{W'_j}$;
2. Calculate the coefficients of the linear linking function $\left(\hat{a}_j, \hat{b}_j\right)$ based on Eqs. 5 and 6, with $\overline{\theta}_{S_j}$, $\widehat{\sigma}_{S_j}$, $\overline{X}_{T_j}$, and $\widehat{\sigma}_{T_j}$;
3. Apply $\left(\hat{a}_j, \hat{b}_j\right)$ to transform $\boldsymbol{\theta}_1$ of the source sample from scale S onto scale T of the target sample, i.e. $\boldsymbol{y_1^j} = \hat{a}_j\boldsymbol{\theta}_1 + \hat{b}_j$, where $\boldsymbol{y_1^j}$ is the transformed plausible values for the source sample, $j = 1, 2, \ldots, N_r$;
4. Calculate $\hat{t}'_{1j}$, using $\boldsymbol{y_1^j}$ with replicate weight $\boldsymbol{W_j}$, $j = 1, 2, \ldots, N_r$.

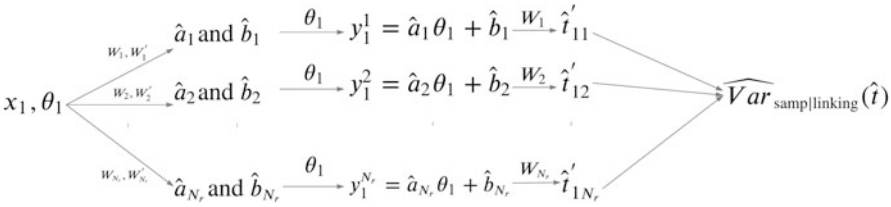

Fig. 1 The calculation process of sampling variance estimation for the source sample

The sampling variance of statistic $\hat{t}$ can then be approximated as

$$\widehat{\text{Var}}_{\text{samplinking}}\left(\hat{t}\right) = \sum_{j=1}^{N_r}\left(\hat{t}'_{1j} - \bar{\hat{t}}_1\right)^2, \tag{8}$$

where

$$\bar{\hat{t}}_1 = \frac{1}{N_r}\sum_{j=1}^{N_r}\hat{t}'_{1j} \tag{9}$$

Figure 1 illustrates the calculation process of $\widehat{\text{Var}}_{\text{samplinking}}\left(\hat{t}\right)$ in Eq. 8. Alternatively, one can approximate the sampling variance of statistic $\hat{t}$ as

$$\widehat{\text{Var}}'_{\text{samplinking}}\left(\hat{t}\right) = \sum_{j=1}^{N_r}\left(\hat{t}'_{1j} - \hat{t}'_1\right)^2, \tag{10}$$

where $\hat{t}'_1$ is calculated by using the original weights $\boldsymbol{W}_{\textbf{orig}}$ and the first set of plausible values that are linked to scale T. The scale transformation follows steps 1-3 described above while using the original student weights $\left(\boldsymbol{W}_{\textbf{orig}}, \boldsymbol{W}'_{\textbf{orig}}\right)$.

We point out that when calculating $(\hat{a}_j, \hat{b}_j)$, $j = 1, 2, \ldots, N_r$, we pair the replicate weights once and in their corresponding sequential order (i.e., pairing the jth replicate weights from both the source and target samples). For the source and target samples that are dependent, pairing the replicate weights of the two samples in this matter properly accounts for the dependency between the samples. On the other hand, if the source and target samples are independent, then the pairings between the source and target samples can be random. In fact, there are $N_r!$ possible ways to pair the replicate weights between the two samples. In theory, one can calculate the variance estimate for all $N_r!$ sets of pairings and then take an average. In practice, $N_r!$ is usually a very large number. To reduce computational burden, a practical approach is to randomly select a subset from the $N_r!$ sets of pairings. Suppose the N_s ($N_s < N_r!$) sets of random pairings are generated and for ith set of pairing the sampling variance estimate is $\widehat{\text{Var}}^{(i)}_{\text{samplinking}}\left(\hat{t}\right), i = 1, 2, \ldots, N_s$, which is

calculated using Eq. 8. Then the sampling variance is estimated as the average of the N_s estimates:

$$\widehat{\text{Var}}^{*}_{\text{samp|linking}}(\hat{t}) = \frac{1}{N_s}\sum_{i=1}^{N_s}\widehat{\text{Var}}^{(i)}_{\text{samp|linking}}(\hat{t}) \tag{11}$$

The choice of the value N_s is a balance between the computation intensity and the stability of the variance estimate.

The above procedure described how to calculate the sampling variance for the source sample statistics only. There are also situations where the statistics are computed based on combining the source and target samples. Next, we show that the procedure can be generalized to estimate the sampling variance for the combined sample as well.

To do that, after getting the transformed plausible values $\boldsymbol{y}_{\mathbf{1}}^{j}$, we concatenate $\boldsymbol{y}_{\mathbf{1}}^{j}$ with $\boldsymbol{x}_{\mathbf{1}}$ of the target sample as the combined set of plausible values, $\boldsymbol{z}_{\mathbf{1}}^{j} = \begin{pmatrix} \boldsymbol{y}_{\mathbf{1}}^{j} \\ \boldsymbol{x}_{\mathbf{1}} \end{pmatrix}$, $j = 1, 2, \ldots, N_r$. Then the statistic of interest based on the combined sample can be calculated using $\boldsymbol{z}_{\mathbf{1}}^{j}$ with weight $\boldsymbol{W}_{\boldsymbol{j}}^{\mathbf{comb}}$, which is the replicate weights for the combined sample. Note that in practice, $\boldsymbol{W}_{\boldsymbol{j}}^{\mathbf{comb}}$ are created specially to the analysis of the combined sample. The rest of the calculation is the same as shown in Eq. 8.

Figure 2 shows the calculation process for statistics of the combined sample. Note that the replicate weights are paired following their corresponding sequential order as (1 to 1), (2 to 2), etc. As discussed earlier, when the source and target samples are independent of each other, the pairing of plausible values from the two samples can be random.

3.2 *Estimation of Latency Variance*

We now discuss the procedure of calculating the latency variance of the source sample statistics $\hat{t}$ as defined in Eq. 1, which is linked to scale T through the random-groups linking.

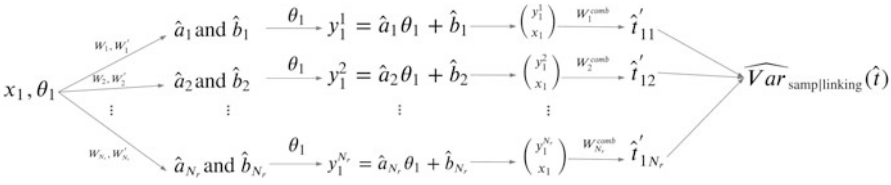

Fig. 2 The process of sampling variance estimation for the combined sample

Using the M sets of plausible values from the source sample and the target sample, we conduct the following steps:

1. Calculate $\overline{\theta}_{S_i}$ and $\widehat{\sigma}_{S_i}$, the mean and SD of the scale scores using the ith set of plausible value in the source sample on scale S with $\boldsymbol{W}_{\mathbf{orig}}$;
2. Calculate $\overline{X}_{T_i}$ and $\widehat{\sigma}_{T_i}$, the mean and SD of the scale scores using the ith set of plausible value in the target sample on scale T with $\boldsymbol{W}'_{\mathbf{orig}}$;
3. Calculate the transformation coefficients $\left(\hat{a}_i, \hat{b}_i\right)$ based on Eqs. 5 and 6 with $\left(\overline{\theta}_{S_i}, \widehat{\sigma}_{S_i}\right)$ and $\left(\overline{X}_{T_i}, \widehat{\sigma}_{T_i}\right)$, $i = 1, 2, \ldots, M$;
4. Apply $\left(\hat{a}_i, \hat{b}_i\right)$ to transform $\boldsymbol{\theta_i}$ from scale S onto scale T, i.e., $\boldsymbol{y}_{\boldsymbol{i}}^{*}=\hat{a}_i\boldsymbol{\theta_i}+\hat{b}_i$;
5. Calculate the statistic of interest $\hat{t}_i^{*}$, using $\boldsymbol{y}_{\boldsymbol{i}}^{*}$ with $\boldsymbol{W}_{\mathbf{orig}}$, $i = 1, 2, \ldots, M$;
6. Calculate the latency variance of the source sample statistics.

$$\widehat{\mathrm{Var}}_{\mathrm{lat|linking}}\left(\hat{t}\right) = \left(1 + \frac{1}{M}\right)\frac{\sum_{i=1}^{M}\left(\hat{t}_i^{*} - \hat{t}^{*}\right)^2}{M-1} \tag{12}$$

where

$$\hat{t}^{*} = \frac{\sum_{i=1}^{M}\hat{t}_i^{*}}{M} \tag{13}$$

The process of calculating the latency variance is illustrated in Fig. 3.

In the above procedure, the plausible values from the two samples are paired when calculating the linking function coefficients $\left(\hat{a}_i, \hat{b}_i\right)$, $i = 1, 2, \ldots, M$. The plausible values for the source and target samples are multiple imputations that were drawn independently using two latent regression models and therefore are independent regardless whether the two samples are dependent or independent of each other.

There are a total of $M!$ possible sets of pairings of the plausible values from the two samples, with M sets of plausible values for each sample. In practice, we can choose a subset of random pairings to reduce computation intensity. Let's assume

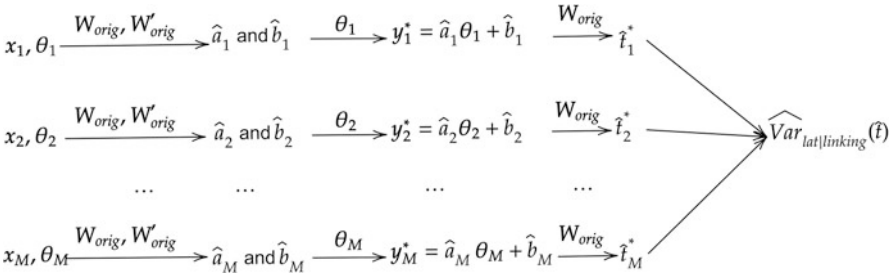

Fig. 3 The calculation process of latency variance estimation for the source sample

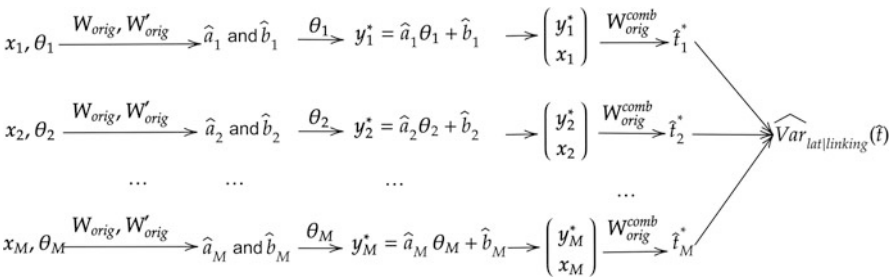

Fig. 4 The process of latency variance estimation for the combined sample

N_s ($N_s < M!$) sets of random pairings are generated, and for ith set of pairing, the latency variance estimate is $\widehat{\text{Var}}^{(i)}_{\text{latllinking}}\left(\hat{t}\right), i = 1, 2, \ldots, N_s$ which is calculated using Eq. 12. Then the latency variance can be estimated as the average of the N_s estimates:

$$\widehat{\text{Var}}^{*}_{\text{latllinking}}\left(\hat{t}\right) = \frac{1}{N_s}\sum_{i=1}^{N_s}\widehat{\text{Var}}^{(i)}_{\text{latllinking}}\left(\hat{t}\right) \tag{14}$$

The choice of N_s is a balance between the computation capacity and reducing variability of the variance estimation.

The above procedure to calculate the latency variance is for the source sample statistics only. Similar to the estimation of sampling variance, we can extend the method to calculate latency variance for the statistics based on the combined source and target sample. To do that, after transforming the source sample plausible values from $\boldsymbol{\theta_i}$ to $\boldsymbol{y_i^*}$ using $\left(\hat{a}_i, \hat{b}_i\right), i = 1, 2, \ldots, M$, we concatenate $\boldsymbol{y_i^*}$ with $\boldsymbol{x_i}$ as $\boldsymbol{z_i^*} = \begin{pmatrix} \boldsymbol{y_i^*} \\ \boldsymbol{x_i} \end{pmatrix}$. Then the statistic of interest based on the combined sample can be calculated using $\boldsymbol{z_i^*}$ with weight $\boldsymbol{W^{comb}_{orig}}$, which is the original weights for the combined sample. The rest of the calculation is the same as shown in Eq. 12.

Figure 4 displays this calculation procedure for the combined sample, with the pairing of the plausible values following a (1 to 1), (2 to 2), etc. fashion.

Finally, the total variance of the statistic $\hat{t}$ is the sum of the sampling and latency variances. When the source and target samples are dependent, the total variance is estimated as

$$\widehat{\text{Var}}_{\text{totalllinking}}\left(\hat{t}\right) = \widehat{\text{Var}}_{\text{samplinking}}\left(\hat{t}\right) + \widehat{\text{Var}}^{*}_{\text{latllinking}}\left(\hat{t}\right) \tag{15}$$

3.3 Properties of the Proposed Variance Estimation Method

In this study, we consider linear linking in a random-groups design. That is, a linear function is applied to align the mean and SD of the source sample score distribution to the mean and SD of the target sample score distribution. Next, we show that regardless of the sample size and other features of the source sample, its mean and SD are fixed to be the same as those of the target sample as the expected result of the linking. Recall the linear function has the following form:

$$\boldsymbol{y_i} = \hat{a}\boldsymbol{\theta_i} + \hat{b},\ i = 1, 2, \ldots, M. \tag{16}$$

where $\hat{a} = \frac{\widehat{\sigma}_T}{\widehat{\sigma}_S}$ and $\hat{b} = \overline{X}_T - \hat{a}\overline{\theta}_S$, as defined in Eqs. 5 and 6.

Let $\overline{Y}_S$ and $\widehat{\sigma}_S^Y$ denote the mean and SD of the transformed scores of the source sample, then given how the $\hat{a}$ and $\hat{b}$ are constructed, we have

$$\overline{Y}_S = \hat{a}\overline{\theta}_S + \hat{b} = \overline{X}_T \tag{17}$$

and

$$\widehat{\sigma}_S^Y = \hat{a} * \widehat{\sigma}_S = \widehat{\sigma}_T \tag{18}$$

The above property is true when the weights used in the calculation are the original weights or the replicate weights. Therefore, for the estimation of sampling variance discussed in Sect. 3.1, $\hat{t}'_{1j}$, $j = 1, 2, \ldots, N_r$, for the source sample are the same as the corresponding statistics of the target sample. According to Eq. 8, the sampling variances of the overall mean and SD for the source sample are the same as those for the target sample, provided the point estimates used in the formula are also the same between the two samples.

Similarly, for the latency variance estimation, $\hat{t}_i^*$, $i = 1, 2, \ldots, M$, of the source sample are the same as the corresponding statistics of the target sample. Following the same logic as for the sampling variance, the latency variances of the overall mean and SD for the source sample are the same as those for the target sample.

Now for the combined source and target sample, we have plausible values $\boldsymbol{z_i} = \begin{pmatrix} \boldsymbol{y_i} \\ \boldsymbol{x_i} \end{pmatrix}$, $i = 1, 2, \ldots, M$. Then the mean of the combined sample

$$\overline{Z} = \frac{\overline{Y}_S n_S + \overline{X}_T n_T}{n_S + n_T} = \frac{\overline{X}_T (n_S + n_T)}{n_S + n_T} = \overline{X}_T \tag{19}$$

where n_S and n_T are the weighted sample size of the source and target samples. Similarly, for the SD of the combined sample,

$$\begin{aligned}\widehat{\sigma}_Z &= \sqrt{\frac{\left(\widehat{\sigma}_S^Y\right)^2 n_S + (\widehat{\sigma}_T)^2 n_T}{n_S + n_T}} \\ &= \sqrt{\frac{(\widehat{\sigma}_T)^2 n_S + (\widehat{\sigma}_T)^2 n_T}{n_S + n_T}} \\ &= \widehat{\sigma}_T \sqrt{\frac{n_S + n_T}{n_S + n_T}} \\ &= \widehat{\sigma}_T\end{aligned} \tag{20}$$

That is, the combined sample, after the scale linking, has the same mean and SD as those for the target sample. Moreover, the variances of the overall mean and SD for the combined sample are also the same as those for the target sample. The argument is the same as for the source sample.

In addition, we point out that the variance estimation considering random-groups linking does not necessarily result in a larger estimated value than those procedures in which the uncertainty due to linking is ignored. For example, as described above, the variances of the mean estimates are the same between the source and target samples after linking. The property holds even when the source sample has much smaller sample size than the target sample. For subgroups, as will be shown in the empirical data below, it is possible to obtain a variance estimate that is smaller when considering the uncertainty due to linking.

4 Applications

4.1 Empirical Results

In this section, we use the data from NAEP to illustrate our proposed method. A study with the random-groups design and linear linking was implemented to link the scores from DBA to PBA. The study involved administering the DBA and PBA to two samples of students, respectively, namely, the DBA sample and the bridge PBA sample. A total of 13,400 students were selected in the study to take either the DBA or PBA. The DBA and bridge PBA samples are dependent with comparable sizes.

Table 1 displays the comparison between the DBA and bridge PBA samples. We can see the demographic distributions between the two samples are comparable.

The bridge PBA and DBA samples were analyzed separately using the IRT latent regression models, and the results were expressed on two separate IRT scales. Following the NAEP operational convention, a total of 20 plausible values were imputed for each student in the 2 samples. In addition, for each sample, the original weight and 62 replicate weights were provided for each student. The results for the bridge PBA sample were estimated on the existing NAEP trend scale, where the

Table 1 Weighted percentage of students by subgroup between the bridge PBA and DBA samples: a NAEP dataset

		Bridge PBA	DBA
Gender	Male	51%	51%
	Female	49%	49%
Race/ethnicity	White	49%	49%
	Black	15%	14%
	Hispanic	27%	27%
	Others	10%	10%
School type	Public	91%	93%
	Non-public	9%	7%

Table 2 Sample sizes, standard errors of estimates of means with and without linking error: the combined DBA/PBA sample

Group	*N*	SE	SE*	SE Ratio
All students	13,400	0.87	0.69	1.26
Male	6900	0.95	0.78	1.22
Female	6500	0.92	0.78	1.18
White	5900	1.01	0.93	1.09
Black	2100	1.30	1.18	1.10
Hispanic	3900	1.20	1.01	1.19
Asian	700	1.83	1.77	1.03
American-Indian/Alaska	200	13.89	14.07	0.99
Northeast	2000	1.85	1.78	1.04
Midwest	2300	1.99	2.01	0.99
South	5400	1.10	0.95	1.16
West	3700	1.19	1.09	1.09

mean and SD of the scale were set operationally to be 150 and 35. For the DBA sample, the results were generated on an arbitrary IRT scale with mean 0 and SD 1. The plausible values of these separate analyses were then used to develop a linear linking function (Eqs. 5 and 6) which allowed for the expression of the DBA results on the bridge PBA scale. Since the DBA and bridge PBA samples are dependent, when calculating the sampling variance, we applied the (1 to 1), (2 to 2), ..., (62 to 62) fashion of pairing the replicate weights between the DBA and bridge PBA samples.

Table 2 presents the standard errors of the mean estimates for the combined DBA/PBA sample, using the proposed new method (Eqs. 8, 14, and 15). For comparison purpose, we also include the usual NAEP variance estimates which do not contain linking variance. Column SE contains the standard errors calculated using our proposed methods, and column SE* contains the standard errors without accounting for random-groups linking. For the race/ethnicity variable, the students in the Native Hawaiian/Other Pacific Islander and the Two or More Races categories are not listed in the table.

We can see from Table 2 that for the overall mean estimate, $\mathrm{SE}\left(\overline{X}_{\text{all\_student}}\right) = 0.87$, $\mathrm{SE}^{*}\left(\overline{X}_{\text{all\_student}}\right) = 0.69$, with a ratio of 1.26. The change in standard errors for subpopulation means, with and without accounting for random-groups linking, is less than the value for the overall population. For the displayed subgroups, the ratios range from about 0.99–1.22.

Furthermore, we observe that the ratios in standard error vary for different subgroups, but have little relationship with the sample size of the group in question. For example, the male and female students are about 50% of the overall population; the ratios in standard error with and without accounting for linking errors are 1.22 and 1.18, respectively. On the other hand, White subgroup is about half the overall population, but the ratio in standard error is 1.09. The linear linking functions were derived based on the overall population, not the subgroups whose results were being transformed by the function. As a result, the ratios in standard error are expected to vary across subgroups. Analytical results of the effect on subgroup standard errors are found in Jewsbury (2019).

4.2 Further Considerations on Latency Variance Estimation

As mentioned in Sect. 4.1, in NAEP, there are 20 plausible values for each student in the source and target samples. When calculating the latency variance, the pairing of the 20 sets of plausible values between source and target sample can be random given the source and target sample plausible values are independent. For example, one way of pairing the plausible values is to follow their corresponding sequential order (i.e., pairing the ith set of plausible values from both the source and target samples). As another example, one could pair the plausible values from the source and target samples following the sequence as (1 to 2), (2 to 3), ..., (20 to 1). In theory, there are 20! possible ways to pair the plausible values between the source and target samples.

We point out that while the latency variance can be estimated based on a single set of pairings of the source and target sample plausible values, averaging the latency variance estimates over multiple sets of pairings, N_s ($N_s < 20!$), is expected to improve the stability of the latent variance estimates. Using the NAEP data, we conducted a simulation study to examine how the latency variance estimates vary with different values for N_s.

In the simulation study, we considered five conditions, with N_s being 1, 5, 10, 25, and 50. For each of the five conditions, we calculated the latency variance 100 times, using the method discussed in Sect. 3. Figure 5 shows the box-plot of the standard errors due to latency for the male students average score for the 100 replications. We can see that as N_s increases, the variation of the standard error estimates decreases. The most noticeable variability reduction is from 1 to 5 random pairings. When N_s equals to 5, the difference between the maximal and minimal standard error estimates among the 100 replications is less than 0.04. In

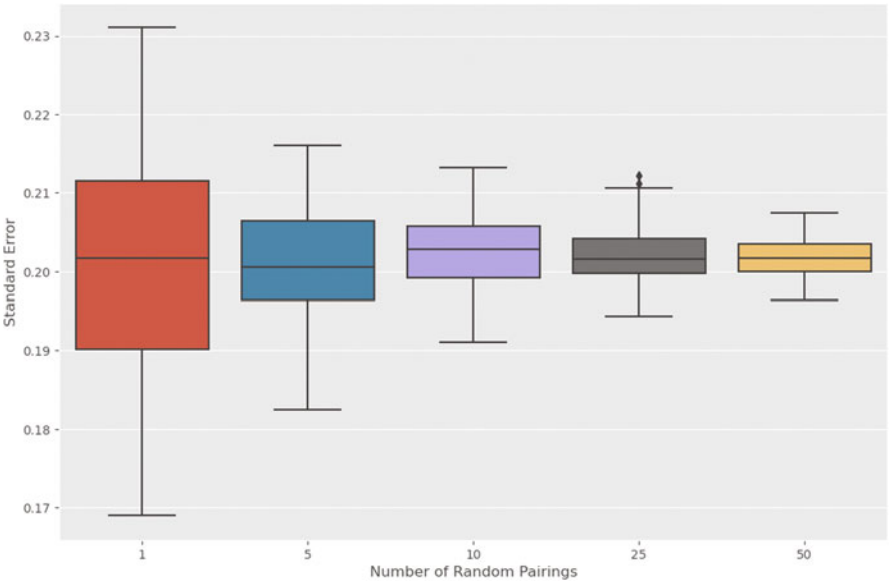

Fig. 5 Box-plot of standard errors due to latency for the male students' average score

this application, we estimated latency error based on five sets of random pairings, considering the latency error estimation is acceptably stable given the magnitudes of subgroup standard errors (as listed in Table 2) and that the latency standard error estimates are typically around 0.2 to 0.4. In practice, similar simulation studies can be helpful to specify the number of random pairings.

5 Conclusion

With complex survey data, it is desirable to have resampling methods that utilize the existing estimation system for variance estimation. For the large-scale survey assessments, the variance of the population statistics is estimated as the sum of two components, the sampling and latency variances. In this paper, we proposed a resampling method for variance estimation when random-groups linking design is applied, incorporating linking error into both the sampling and latency variance estimates. The method is applicable to both linear or nonlinear statistics.

We proposed the estimation procedure in the context of linear linking function. However, the approach applies to both parametric and non-parametric linking functions. Further, it can be applied when the linking sample are dependent or independent.

Acknowledgments The research reported here was supported under the National Assessment of Educational Progress (ED-IES-13-C-0017) to Educational Testing Service (ETS) as administered by the Institute of Education Sciences, US Department of Education. The opinions expressed are those of the authors and do not represent views of the Institute or the US Department of Education. The authors wish to thank Paul Jewsbury, Daniel F McCaffrey, Mei-Jang Lin, John R Donoghue, and Xueli Xu at Educational Testing Service for contributing to this work. We would also like to thank Keith Rust at Westat and the Design and Analysis Committee of the National Assessment of Educational Progress for useful discussions and advice. We are grateful to the anonymous reviewers for the useful suggestions.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Braun, H., & von Davier, M. (2017). The use of test scores from large-scale assessment surveys: Psychometric and statistical considerations. *Large-scale Assessments in Education, 5*(1), 1–16.

Eignor, D. R. (2007). Linking scores derived under different modes of test administration. In *Linking and aligning scores and scales* (pp. 135–159). Springer.

Hansen, M. H., Dalenius, T., & Tepping, B. J. (1985). The development of sample surveys of finite populations. In A. C. Atkinson & S.E. Fienberg (Eds.), *A celebration of statistics* (pp. 327–354). Springer.

Jewsbury, P. (2019). Error variance in common population linking bridge studies. (Research Report No. RR-19-42). Princeton, NJ: Educational Testing Service.

Jewsbury, P., Finnegan, R., Xi, N., Jia, Y., Rust, K., & Burg, S. (2020). 2017 NAEP transition to digitally based assessments in mathematics and reading at grades 4 and 8: Mode evaluation study. https://nces.ed.gov/nationsreportcard/subject/publications/main2020/pdf/transitional_whitepaper.pdf

Jia, Y., Phillips, G., Wise, L., Rahman, T., Xu, X., Wiley, C., & Diaz, T. (2011). NAEP-TIMSS linking study: Technical report on the linking methodologies and their evaluations (NCES 2014-461).

Johnson, E., Cohen, J., Chen, W., Jiang, T., & Zhang, Y. (2005). 2000 NAEP–1999 TIMSS linking report.

Johnson, E. G. (1998). *Linking the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS): A technical report*. US Department of Education, Office of Educational Research and Improvement.

Johnson, E. G., & Rust, K. F. (1992). Chapter 5: Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17*(2), 175–190.

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

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. Springer.

Kovar, J., Rao, J., & Wu, C. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics, 16*(S1), 25–45.

Krewski, D., & Rao, J. N. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics* 1010–1019.

Mazzeo, J. (2018). Large-scale group-score assessments. In *Handbook of item response theory* (pp. 297–311). Chapman and Hall/CRC.

Mazzeo, J., Liu, B., Donoghue, J., & Xu, X. (in press). Approximate standard errors for NAEP results that incorporate linking error under the random group design.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*(3), 359–381.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*(392), 993–997.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196.

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

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. Wiley.

Shao, J. (1996). Invited discussion paper resampling methods in sample surveys. *Statistics, 27*(3–4), 203–237.

Valliant, R. (1990). Comparisons of variance estimators in stratified random and systematic sampling. *Journal of Official Statistics, 6*(2), 115–131.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions. *Handbook of Statistics, 26*, 1039–1055.

Wolter, K. M. (2007). *Introduction to variance estimation* (2nd ed., Vol. 53). Springer.