Measuring the Power of Learning.™

Research Report
ETS RR–18-16

Grouping Effects on Jackknifed Variance Estimation for Item Response Theory Scaling and Equating With Cluster-Based Assessment Data

Lin Wang

Jiahe Qian

Yi-Hsuan Lee

December 2018

ETS Research Report Series

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Grouping Effects on Jackknifed Variance Estimation for Item Response Theory Scaling and Equating With Cluster-Based Assessment Data

Lin Wang, Jiahe Qian, & Yi-Hsuan Lee

Educational Testing Service, Princeton, NJ

Educational assessment data are often collected from a set of test centers across various geographic regions, and therefore the data samples contain clusters. Such cluster-based data may result in clustering effects in variance estimation. However, in many grouped jackknife variance estimation applications, jackknife groups are often formed by a random grouping method that ignores the cluster structures of the data. In this study, we constructed both random and cluster-based jackknife groups for data known to have cluster structures and compared the jackknifed standard errors, yielded by two different grouping methods, of item response theory (IRT) scaling coefficient estimates and equated scores. Three independent data samples from an international test of English were used for the study. The cluster-based jackknife group results showed relatively larger jackknifed standard errors of scaling coefficient estimates and scale scores than the results of the random jackknife groups for all three data samples. For cluster-based assessment data, the cluster-based jackknife approach provides a more appropriate way to estimate the standard errors of the parameters of IRT calibration, scaling, and equating analyses.

Keywords Grouped jackknife procedure; random grouping method; cluster-based grouping method; IRT-based item calibration; Stocking–Lord test characteristic curve linking

doi:10.1002/ets2.12204

Educational testing programs that administer multiple test forms over time routinely conduct test score equating to adjust for minor differences in the difficulty levels among different test forms so that test scores from different test administrations are comparable; this is critical to score validity and fairness. Because equating is a statistical procedure that analyzes the responses from a sample of test takers to a collection of test questions, it is inevitable that “random equating errors are introduced through estimation of equating parameters” (Haberman, Lee, & Qian, 2009, p. 1). Equating error is a form of sampling error and affects precision in estimation of equating parameters that relate to the quality of equated test scores. Evaluation of equating error is an important means to understanding and monitoring the quality of test scores.

In practice, equating is a complex procedure that involves a series of operations. For instance, in true score equating based on item response theory (IRT), the whole equating process may include estimating item parameters from response data (calibration), transforming item parameter estimates to the test’s IRT scale (scaling), and computing scale scores by using the scaled item parameter estimates (equating). The complexity of this whole process makes it difficult to compute equating errors analytically. Resampling techniques, such as bootstrap, jackknife, or balanced repeated replications (Efron, 1979; Shao, 1996; Wolter, 2007), have been used in estimating sampling variance when parametric estimation is not feasible (Zaman & Alakus, 2015). For example, the jackknife method has been used for variance estimation in large-scale educational assessments such as the National Assessment of Educational Progress (Allen, Donoghue, & Schoeps, 2001; Oranje, 2006; Phillips, 2014; Qian, 2005), the Program for International Student Assessment (Neidorf, Binkley, Gattis, & Nohara, 2006; Nohara, 2001), and the Trends in International Mathematics and Science Study (Neidorf et al., 2006; Nohara, 2001).

To assess equating errors, the jackknife method is an effective alternative to analytical delta approaches (Haberman et al., 2009). Implementation of the jackknife method has been improved since its introduction by Tukey (1958). For example, when the sample size of an analysis is small, one outlier observation may significantly impact the outcome of the analysis, and the traditional jackknife method that deletes one observation at a time (Tukey, 1958) is typically applied. On the other hand, when the sample size is large, which is typical of educational assessment data, any one observation may

Corresponding author: L. Wang, E-mail: lwang@ets.org

not have noticeable impact on the outcome. The jackknife application may need to delete a group of observations that share some common features at a time. This is called the *grouped jackknife method* (Miller, 1964). Haberman et al. (2009) provided a formal justification for applying the grouped jackknife method to evaluating equating accuracy and error estimation, which was the first research to establish the theoretical grounds for employing this method in educational measurement applications.

The application of the grouped jackknife method can take the underlying structure of assessment data into account. For instance, the test used for illustration in this study is an international English proficiency assessment with data collected from test centers worldwide. Not all test centers are available for every test administration — usually only 70 – 80% of the test centers available would be involved in each test administration. Either a single test center or an aggregate of test centers from the same geographic region can be treated as a cluster in these assessment samples. Previous studies (Lee & Haberman, 2013, 2018) have found that test takers' performance on English proficiency assessments is related to geographic regions defined by the country or region in which test takers took the test (called *test center country* in this report). Lee and Haberman (2013) found that predictors for regional effects explained more variation in the mean scores of the administrations than did other predictors, such as seasonality. With different data sets from the same test, Lee and Haberman (2018) showed that test takers' geographic region, defined either by test center country or by both test center country and native country (test takers' self-reported country of birth), was the most important source of variation in test takers' scores compared to other background variables, such as gender and age. It is clear that test takers' performance on such an English test tended to vary from region to region. In sampling, this phenomenon is called *clustering effects*. Incorporating such cluster information into jackknife analyses can take into account potential sampling errors due to clustering effects that are present in a data set.

In the reported studies of grouped jackknifing applications in educational assessments, instead of forming cluster-based jackknife groups, a random grouping approach is often employed to form jackknife groups even though the data may have varying degrees of clustering features (Lu, Haberman, Guo, & Liu, 2015; Wang, Qian, & Lee, 2013). In sampling, clustering effects are caused by intercluster correlations, and cluster-based sampling will yield larger variance estimates than those yielded by random sampling (Cochran, 1977, p. 240; Kish, 1965, p. 162). A fundamental statistical principle is that the methods of estimation should be compatible with the sample structure of data because the frame structure and sampling approach determine the functional form and statistical properties of the frequency distribution. Cochran (1977, p. 150) claimed that when one knows the functional form of the frequency distribution followed by the data in the sample, the method of estimation should be carefully geared to this type of distribution and data structure. In educational research, it is relatively easy to conduct random group-based jackknife analyses because this does not require test takers' background information, which is often not readily available. The issue is that the randomly formed groups (clusters) tend to be uniform and have little variation across these uniform clusters. This method will underestimate the variances of interest if the assessment data actually contain nonuniform clusters (e.g., test takers from different geographic regions). Therefore the jackknifing approach based on the random grouping method is not optimal for educational assessment data with a cluster data structure because it disregards the cluster effects.

The purpose of the current study was to investigate the cluster-based jackknife grouping effects on jackknifed variance estimates for IRT scaling and equating for assessment data with cluster structures. This research was focused on whether the cluster-based grouping method would yield appropriate variance estimates in comparison with the random grouping method when analyzing data with cluster structures. The expectation of the study was that the random grouping method would underestimate the variance with the cluster-based data. Both the random and the cluster-based grouping methods were employed in this investigation.

The following section describes the data source, the design of the study, and the analyses. Results are presented in the third section, followed by a discussion and conclusions in the final section.

Method

Data Source

This study employed operational data from an English reading test that was administered to nonnative English speakers worldwide. The test contained selected-response items only; most items were scored dichotomously (0 or 1 point), and a few items were scored polytomously (0 – 2 points). The data were from three test administrations that were approximately

2 weeks apart; all three test administrations occurred in similar geographic regions of the world. One unique test form was used for each test administration; no overlapping items were among the three test forms (called Form 1, Form 2, and Form 3 hereafter). The sample sizes were 4,564, 2,976, and 3,378 for Forms 1, 2, and 3, respectively. All the test forms were constructed to meet the same content and statistical specifications. For the purpose of this study, we considered that all the potential test takers in the regions where a test was administered constituted the study population (population hereafter) for that test administration, and the test takers who took the test formed the sample from its population.

Designs for Constructing Random and Cluster-Based Groups

The random grouping method constructs K groups from the entire sample of test takers (N) for a test administration; these K groups are simple random samples of equal size (N/K) without overlap. On the other hand, the cluster-based groups are constructed in a nonrandom manner by utilizing the data structures, mainly test centers and aggregates of test centers for the analysis of interest. After the jackknife groups are constructed, grouped jackknife replicate samples can then be constructed by deleting one jackknife group at a time in a sequential manner; a grouped jackknife replicate sample consists of $K-1$ groups. For this study, both random and cluster-based groups were constructed for the jackknifing analysis, with 100 groups per grouping type for each form (Forms 1, 2, and 3). The following steps were taken to construct the random and cluster-based groups for each test form sample.

Constructing 100 Random Groups

Step 1: The test takers in the sample of a test form were randomly sorted and aggregated into $K = 100$ disjoint groups of the same size, if possible (if not, consider Step 2).

Step 2: When the sample size N of a test form was not the exact multiple of 100 (this was true for all three test forms), N' cases were left ($N' < 100$) and needed to be distributed to N' of the 100 groups. First, one case was randomly selected from the N' cases and assigned to the first group. Next, another case from the $N' - 1$ cases was randomly selected and assigned to the second group, and so on until the last (N'th) case was assigned to the K'th group ($K' < K$). Therefore 100 random jackknife groups were constructed in total, with K' groups having one more case than the rest of the 100 groups. For example, Form 1 had 4,564 cases; 4,500 of the 4,564 cases were first randomly assigned to 100 disjoint groups of 45 per group. The remaining 64 cases were then randomly assigned to the first 64 jackknife groups, one case per group. As a result, 64 groups had 46 cases per group, and the other 36 groups had 45 cases per group.

Constructing 100 Cluster-Based Groups

The principal idea of constructing cluster-based groups is to make all groups have almost the same composition with regard to the data features that define the groups, including the geographic proximity and the test performance similarity (e.g., mean scores) of the test centers. For example, a cluster-based group may be formed to consist of test takers from the same test center; aggregates of test centers can be formed within a country, by country, or by multiple geographically adjacent countries, because the countries that participated in each test administration might have varying numbers of test takers, whereas the jackknife groups should be of similar sizes. In this study, the cluster-based groups were constructed from aggregates of test centers for two considerations. One was that the test center information was available for each test taker and was accurate. The other consideration was that prior studies (Lee & Haberman, 2013, 2018) had found regional effects on test taker performance.

Because 100 cluster-based groups were needed per test form, the average jackknife group sizes, N_{avg}, should be approximately 46, 30, and 34 for Forms 1, 2, and 3, respectively (see section “Data Source” for the three sample sizes). For each test form, if the total number of the test takers from a cluster of test centers in a country was about the size of N_{avg}, all the test takers from this country were aggregated into one “cluster,” with some adjustments, if needed. This was applied to all three test forms. The operational process to build the clusters consisted of two steps:

Step 1: For each test form, the countries were first classified into three categories by size (large, medium, and small). The cluster of test centers in a medium-sized country had an appropriate sample size of test takers that was generally in the range of (0.5 N_{avg}, 1.5 N_{avg}). The only exception was Form 1, where the upper range of country size was as large as 1.57 N_{avg}, as Form 1's sample size was considerably larger than Form 2's or Form 3's. The cluster of test centers in a large-sized country had a sample size larger than 1.5 N_{avg}, and the cluster of test centers in a small-sized country had a sample size less than 0.5 N_{avg}.

Step 2: The aggregate of test centers in a medium-sized country was considered a single cluster-based group. On the other hand, the test takers from a large-sized country were aggregated into multiple cluster-based groups by first sorting the test takers by their test center cities and codes and then assigning the test takers to groups of similar sizes according to the center locations, time zones, test center city, native language, test center code, and so on, which indicated the geographic proximity of the test centers. Finally, the test takers from multiple small-sized countries of adjacent geographic regions (e.g., in Africa or the Middle East) were pooled into one cluster. The rationale for the adjustments with the large- and small-sized countries was that, in the absence of other information, we might assume that test takers from adjacent areas tend to have similar levels of proficiency being measured by the English test. There were, however, two possible issues in the choice of constructing clusters from the test centers in this study. One was that the main focus of this study was the methodological features of the grouped jackknife analysis rather than whether test center would be the most appropriate variable for defining cluster-based groups. The other was that, while test center-based clusters may make sense in most regions of the world, this may not be true for the US sample, because there is not always a correlation between the geographic locations of the US test centers and the test performance of the test takers in those test centers.

Table 1 summarizes the number of cluster-based jackknife groups for each country size category for the three test forms. For example, Form 1 had 4,564 test takers, of which 3,349 were from large countries, 627 from medium countries, and 588 from small countries; the target cluster-based group size was around 46 for Form 1. Therefore 3,349 test takers in the large country category were assigned to 73 cluster-based jackknife groups, the test takers from the medium-sized countries formed 14 groups, the test takers from small-sized countries formed 13 groups, and the total number of cluster-based jackknife groups for Form 1 was 100.

Table 2 provides the distributions of the cluster-based jackknife group sizes by test form. Unlike the random groups that had similar group sizes (the difference was only 1), the cluster-based group sizes varied as a result of forming the groups from the three test forms that had different overall sample sizes. For example, the median sizes of the cluster-based groups were 46, 30, and 34 for Forms 1, 2, and 3, respectively. The group sizes were also different within each test form.

After the jackknife groups were formed by either the random or the cluster-based method, the next step was to create jackknife replicate samples. A jackknife replicate sample was created by dropping one jackknife group from the overall sample (of 100 groups) and therefore contained 99 jackknife groups; repeating this process 100 times yielded 100 jackknife replicate samples. In the jackknifing analysis of equating, each jackknife replicate sample constituted an equating sample; the jackknifed standard error of the statistics of interest was estimated over all the replicate samples. The formulas in the following section show how the jackknifed standard errors were computed.

Analysis

The following analyses were conducted on all of the jackknife replicate samples for the random and cluster-based methods for each of the three test forms:

Table 1 Distributions of Test Takers and Cluster-Based Jackknife Groups by Country Size Category

	Test center country category by size			
	Large	Medium	Small	Total
Form 1				
N	3,349	627	588	4,564
Jackknife groups	73	14	13	100
Form 2				
N	2,350	242	384	2,976
Jackknife groups	79	8	13	100
Form 3				
N	2,773	283	322	3,378
Jackknife groups	82	8	10	100

Table 2 Distributions of the Cluster-Based Jackknife Group Sizes by Test Form

	Form 1	Form 2	Form 3
Total number of test takers	4,564	2,976	3,378
Total number of jackknife groups	100	100	100
Jackknife group size			
Median	46	30	34
Maximum	71	49	47
Minimum	31	22	25
5th percentile of jackknife group sizes (rounded)	38	25	29
95th percentile of jackknife group sizes (rounded)	61	35	37

1. Each test form (Forms 1, 2, and 3) was a new form and needed to be equated to the base form of the test through a "common-item equating to a calibrated item pool" scheme (Kolen & Brennan, 1995, p. 200). Each new test form contained some previously administered items (common items) whose IRT item parameter estimates were already on the IRT scale of the test. In equating operations, these common items are also called anchor items, as their item parameters are used as reference item parameters to scale the new items on a new test form. An IRT true score equating of scores on a new test form consists of three phases: IRT calibration of item parameters (*a* for item discrimination and *b* for item difficulty), item parameter transformation (scaling), and IRT true score equating.
2. For this study, the IRT calibration of item parameters was conducted by using the Parscale software package (ETS version by Muraki & Bock, 1999). The calibrated item parameters were not on the IRT scale of the test and needed to be transformed. The test characteristic curve method by Stocking and Lord (1983) was used to estimate the scaling coefficients *A* and *B* from the anchor items; the *A* and *B* coefficients were then used to transform (scale) the item parameter estimates from the calibration phase so that the new items' parameter estimates were put on the reference IRT scale of the test. The scaled parameter estimates of the operational items of the new form were used in the IRT true score equating phase to compute equated test scores (for a detailed illustration, see Kolen & Brennan, 1995, pp. 174–180). Both the item parameter transformation and the IRT true score equating phases were accomplished using ICEDOG software (Robin, Holland, & Hemat, 2006).
3. The statistics of interest in this study were the standard errors of the estimated scaling coefficients *A* and *B* in the item parameter transformation (linking) phase and the standard errors of the scale scores. The grouped jackknifing procedure was implemented to estimate the standard errors in both the linking and the equating phases (Qian, 2005; Wang et al., 2013) on the entire sample data of a test form and on each jackknife replicate sample. The jackknifed standard errors of the statistics of interest were estimated as follows.

   Let $\hat{\theta}$ be the parameter estimated from the whole sample and $\hat{\theta}_{(j)}$ be the estimate from the *j*th jackknife replicate sample. The jackknifed variance of $\hat{\theta}$ was estimated by

   $$v_J\left(\hat{\theta}\right) = \frac{J-1}{J}\sum_{j=1}^{J}\left(\hat{\theta}_{(j)} - \hat{\hat{\theta}}.\right)^2, \tag{1}$$

   where $J = 100$ is the number of jackknife replicate samples and $\hat{\hat{\theta}}.$ is the mean of $\hat{\theta}_{(j)}$ ($j = 1, 2,. .., 100$). The jackknifed standard error was then computed by

   $$se_J\left(\hat{\theta}\right) = \sqrt{v_J\left(\hat{\theta}\right)}. \tag{2}$$

   (Wolter, 2007).

Results

Jackknife Standard Errors of the Scaling Coefficient Estimates

Table 3 presents the means and jackknifed standard errors of the estimated *A* and *B* coefficients of the 100 jackknife replicate samples for the random and cluster-based methods. The jackknifed standard errors were computed by using

Table 3 Jackknifed Error Estimates of Scaling Coefficients *A* and *B*

	Form 1	Form 2	Form 3
N	4,564	2,976	3,378
Jackknife replicates (per group type)	100	100	100
Mean (*A*)	0.992	1.022	0.985
$se_R(A)$ [a]	0.015	0.014	0.014
$se_{CB}(A)$ [b]	0.019	0.024	0.018
Ratio ($se_{CB}(A)/se_R(A)$)	1.267	1.714	1.286
Mean (*B*)	−0.081	−0.236	−0.082
$se_R(B)$	0.017	0.014	0.020
$se_{CB}(B)$	0.041	0.048	0.038
Ratio ($se_{CB}(B)/ se_R(B)$)	2.412	3.249	1.900

[a] Jackknifed standard error (*SE*) for the random groups. [b] Jackknifed *SE* for the cluster-based groups.

Equations (1) and (2). All the mean values of coefficient estimate *A* were close to 1; the mean *B* values of Forms 1 and 3 were almost identical and were both close to 0; Form 2 had a lower mean *B* value than Forms 1 and 3.

The jackknifed standard errors (*SE*) of estimated coefficients *A* and *B* were very small for both the random and cluster-based groups, but the cluster-based groups had relatively larger *SE* values ($se_{CB}(A)$, $se_{CB}(B)$) than the random groups ($se_R(A)$, $se_R(B)$). For example, for Form 1, $se_{CB}(A)$ and $se_R(A)$ were 0.019 and 0.015, respectively; the ratio of $se_{CB}(A)$ to $se_R(A)$ was 1.267. A ratio of 1 indicates that the two *SE* values were the same; a ratio greater than 1 suggests that the cluster-based group method yielded a larger *SE* than the random group method, and vice versa. Similarly, for Form 1, the ratio of $se_{CB}(B)$ to $se_R(B)$ was 2.412. Form 2 had the largest difference of the three test forms in the *SE* estimates between the random method and the cluster-based method.

Jackknifed Standard Errors of the Scale Scores

In the last phase of the grouped jackknifing application, for each of the 100 jackknife replicate samples, a scale score was computed for each raw score on a test. This operation resulted in 100 scale scores for each raw score, and the jackknifed standard error was computed from the 100 scale scores.

Table 4 presents a summary of the jackknifed standard errors from the random and the cluster-based grouping methods; the information includes the mean, median, minimum, and maximum values of the random group jackknifed standard errors (se_R), the cluster-based jackknifed standard error (se_{CB}), and their ratios (se_{CB}/se_R). The jackknifed standard errors at each raw score point on the three test forms are provided in the appendix. Overall, the jackknifed standard errors were small on all three test forms. For example, the median jackknifed standard errors for the random groups se_R were 0.065, 0.075, and 0.066 for Forms 1, 2, and 3, respectively. The median jackknifed standard errors for the cluster-based groups (se_{CB}) were 0.081, 0.087, and 0.074 for Forms 1, 2, and 3, respectively. The median ratios of the two errors (se_{CB}/se_R) were 1.246, 1.160, and 1.121 for Forms 1, 2, and 3, respectively. The cluster-based method had larger mean and median jackknifed standard errors than the random group method on all three test forms.

Distributions of the Jackknifed Standard Errors

In addition to the summary information in Table 4, Figures 1–3 plot the jackknifed standard errors by raw score points for the random group and cluster-based group methods (see Tables A1–A3 in the appendix for the corresponding results). The figures may assist readers in inspecting how the two types of jackknifed standard errors spread across the raw score scale and how much the two types of estimated jackknifed standard errors differ from each other across all the score points. On each test form, the random group standard errors (the solid line) are shown to be smaller than the cluster-based standard errors (the dotted line) for almost all raw score points. Of the three test forms, Form 1 showed the largest differences between the two sets of standard errors; Forms 2 and 3 had similar differences between the two sets of standard errors. The patterns of the differences varied from form to form.

Figure 4 shows the ratios of the cluster-based jackknifed standard errors to the random group jackknifed standard errors by raw score points for the three test forms; the ratios indicate the differences in the jackknifed standard errors between the two methods. As was mentioned earlier, of the three test forms, Form 1 (solid line) tended to have the largest differences, especially between raw scores of 33 and 39.

Table 4 Jackknifed Standard Errors of Scale Scores by Form

	se_R	se_{CB}	Ratio (se_{CB}/se_R)
Form 1			
Mean	0.080	0.098	1.225
Median	0.065	0.081	1.246
Minimum	0.047	0.046	0.979
Maximum	0.148	0.176	1.189
Form 2			
Mean	0.085	0.095	1.118
Median	0.075	0.087	1.160
Minimum	0.024	0.033	1.375
Maximum	0.151	0.150	0.993
Form 3			
Mean	0.080	0.088	1.100
Median	0.066	0.074	1.121
Minimum	0.035	0.040	1.143
Maximum	0.137	0.151	1.102

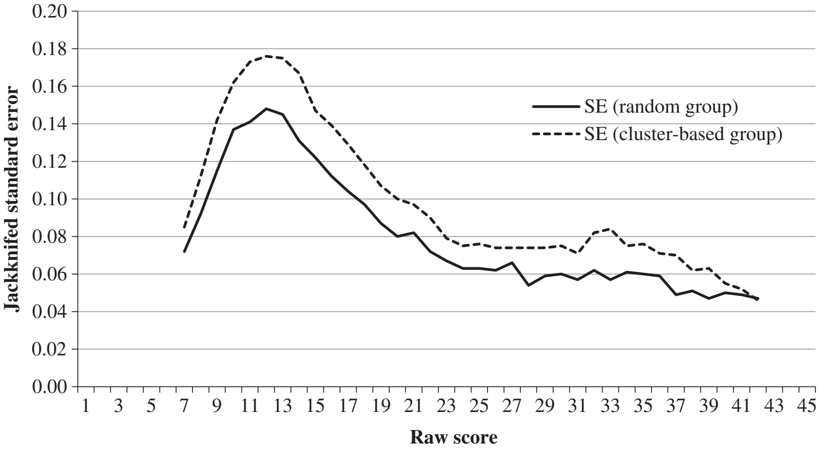

Figure 1 Random and cluster-based jackknifed standard errors of scale scores of Form 1.

Discussion

Jackknifed Standard Errors of the Scaling Coefficient Estimates

The estimated scaling coefficients (A and B) are used to transform the IRT parameter estimates (a, b) of the new items to the IRT scale defined for the test. After transformation, all the item parameter estimates are on the same scale and comparable. Because scale scores are derived from the transformed parameter estimates (a, b) of the items on the test, the estimated scaling coefficients (A and B) determine the transformed parameter estimates of the items on a new test form and consequently influence the scale scores. Small standard errors associated with the A and B coefficient estimates suggest stable and desirable equating outcomes.

The absolute values of the jackknifed standard errors of the estimated A and B coefficients were small for both the random and the cluster-based groups ($se_R(A)$, $se_{CB}(A)$, $se_R(B)$, and $se_{CB}(B)$); this was found to be true for all three test forms in this study. However, although each method yielded small error estimates, the cluster-based method had relatively larger errors than the random groups, as was shown by the ratios $se_{CB}(\mathrm{A})/se_R(\mathrm{A})$ and $se_{CB}(B)/se_R(\mathrm{B})$, respectively. The relatively larger standard errors of A and B from the cluster-based groups on all three test forms could be due to both the cluster structure of the underlying data and the way the groups were formed. Because the data in this study were known to have cluster structures, and the random grouping method did not take those underlying structures into consideration, the random grouping method results should be viewed as approximations at best. The cluster-based method by which the groups were formed took into consideration the cluster structures of the data and seemed to be consistent with our

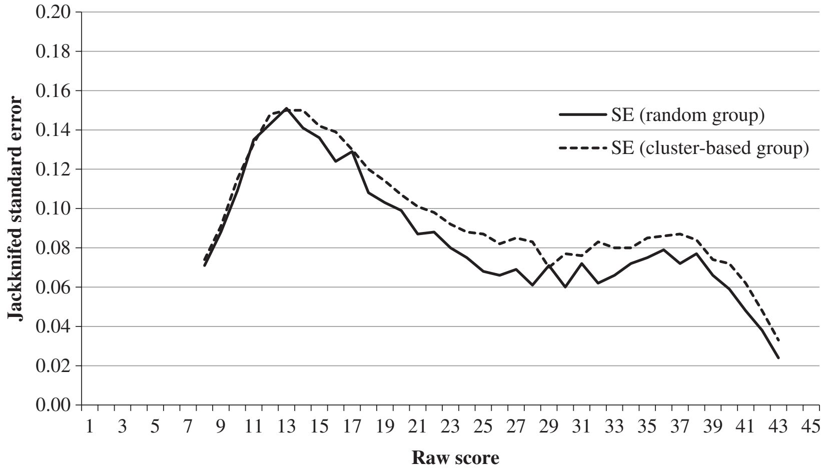

Figure 2 Random and cluster-based jackknifed standard errors of scale scores of Form 2.

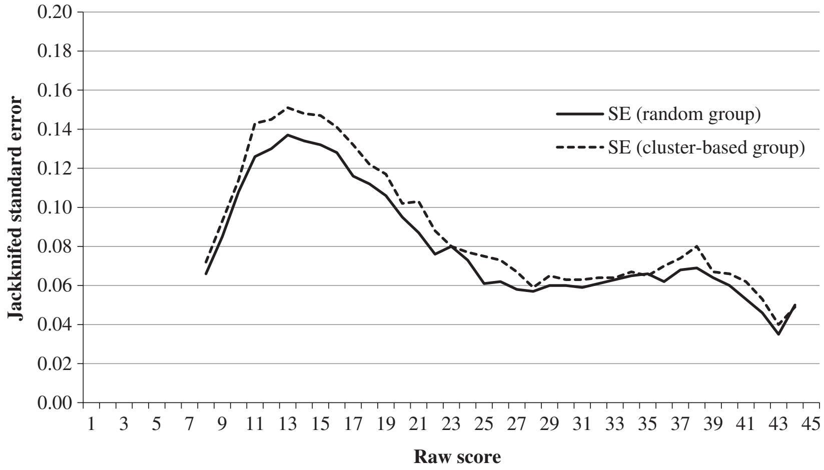

Figure 3 Random and cluster-based jackknifed standard errors of scale scores of Form 3.

theoretical expectations. They also agreed with the findings by Lee and Haberman (2013) that test scores were correlated with the geographic regions of the test takers.

Jackknifed Standard Errors of Scale Scores

The two previously mentioned observations on the jackknifed standard errors of the estimated scaling coefficients were also true for the jackknifed standard errors of the scale scores presented in Table 4. Because the scale scores are reported to score users, the quality of the scale scores is the ultimate concern of a testing program. The fact that the jackknifed errors were small for the three test forms was a good indication that the scale scores were estimated accurately and stably by both methods. As expected, the cluster-based method yielded larger jackknifed errors than the random method. For example, as shown in Table 4, the ratio of mean jackknifed standard errors was close to 1.23, with the cluster-based estimate being 23% larger than the random group estimate.

The differences in jackknifed standard errors between the random group and the cluster-based group methods were consistent across the three test forms, as shown in Figures 1–4. Because the three test forms were independent of one another and the same findings were replicated, it is unlikely that the consistent differences between the two methods could have been due to mere coincidence.

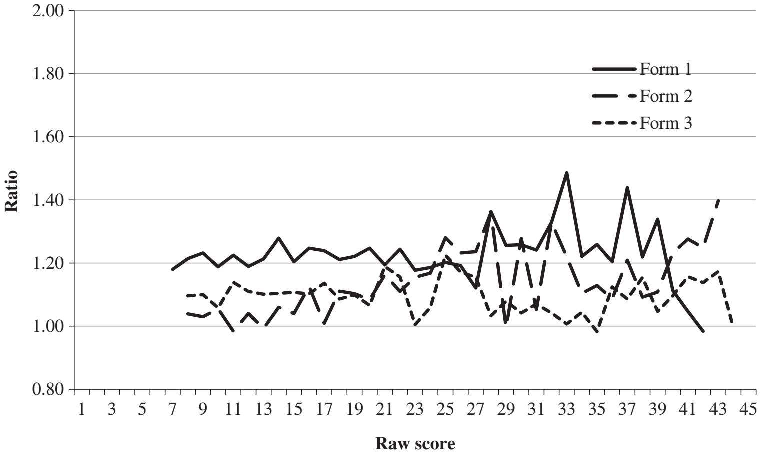

Figure 4 Ratios of cluster-based jackknifed standard errors to random group jackknifed standard errors of scale scores.

Conclusions

Two grouping methods for grouped jackknifing were explored and compared in this study. When both the random jackknife groups and the cluster-based jackknife groups were created for the same data in this study, the jackknifed standard errors of the estimated scaling coefficients and scale scores were found to be relatively larger for the cluster-based group method than for the random group method; this was true for all three data samples. The consistent differences between the two methods across the three data samples are compatible with the statistical properties of the grouping methods. The larger standard errors yielded by the cluster-based group method can be explained by the clustering effects associated with the data structures and to some extent by the way that the jackknife groups and replicates were constructed for this study. For the assessment considered in this study, the cluster-based group jackknifed standard errors seemed to be estimated more appropriately than the random group jackknifed standard errors, as the cluster-based method was compatible with the data structures.

In grouped jackknife applications, constructing cluster-based jackknife groups using background variables remains a challenge, as full and accurate background data are often not available. In this study, for example, the data samples were from a large-scale computer-delivered English assessment that was administered in many countries. It would seem straightforward simply to consider country for the grouping variable, but that was constrained by the varying number of test takers (test volume) across the countries participating in a test administration. The varying test volume across the countries made it impossible to construct equal-sized cluster-based groups for all countries. As an alternative, aggregates of test centers were used in creating jackknife groups by partitioning a large-volume country into multiple groups of test centers in geographic proximity and with similar levels of proficiency measured by the test and also by combining adjacent small-volume countries into one cluster-based group. The principle of aggregation of test centers in this study was based on the geographic proximity and the test performance similarity of the test centers; the clustering effects on the jackknifed standard errors were mainly due to test centers or aggregates of test centers instead of countries. The magnitude of the clustering effects on the grouped jackknife variance estimation, in theory, depends on the intracluster correlations of the test takers within test centers (Cochran, 1977), so the effects can vary across different assessments.

A limitation of this study, as mentioned in the section "Constructing 100 Cluster-Based Groups," is that there is not always a logical association between the US test centers (clusters) and test performance. This is simply because test takers in the United States are from all over the world and do not typically choose to go to test centers based on their native countries. In this sense, some cluster-based groups from the US samples might resemble random groups; this situation may complicate the interpretation of the results from the two methods.

In closing, incorporating cluster information into jackknife analyses takes into account sampling errors due to clustering effects that are present in a data set. Successful applications of grouped jackknifing with a cluster-based grouping method depend on reliable background or ancillary information that can be used to construct jackknife groups to assess jackknifed standard errors. In this study, aggregates of test centers formed the clusters, as reliable information on test centers was available. In practice, a combination of background variables can be used to form cluster grouping in jackknifing analysis on assessment data that include additional background information.

Acknowledgments

The content of this manuscript was presented at the 2015 annual meeting of the American Educational Research Association in Chicago, Illinois, April 16–20, 2015.

References

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

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

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*, 1–26.

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

Kish, L. (1965). *Survey sampling*. New York, NY: John Wiley. https://doi.org/10.1002/j.2333-8504.2009.tb02196.x

Kolen, M. J., & Brennan, R. L. (1995). *Test equating methods and practices*. New York, NY: Springer.

Lee, Y.-H., & Haberman, S. J. (2013). Harmonic regression and scale stability. *Psychometrika, 78*, 815–829. doi:https://doi.org/10.1007/s11336-013-9337-1

Lee, Y.-H., & Haberman, S. J. (2018). *Studying scale stability with harmonic regression: Three approaches to adjustment of examinee-specific demographic data*. Unpublished manuscript.

Lu, R., Haberman, S. J., Guo, H., & Liu, J. (2015). *Use of jackknifing to evaluate effects of anchor item selection on equating with the nonequivalent groups with anchor test (NEAT) design* (Research Report No. RR-15-10). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12056

Miller, R. G. (1964). A trustworthy jackknife. *Annals of Mathematical Statistics, 53*, 1594–1605.

Muraki, E., & Bock, R. (1999). *PARSCALE 3.5: IRT item analysis and test scoring for rating-scale data*. Lincolnwood, IL: Scientific Software.

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 asssessments* (Report No. NCES 2006-029). Washington, DC: 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)* (Report No. NCES 2001-07). Washington, DC: National Center for Education Statistics.

Oranje, A. (2006). *Jackknife estimation of sampling variance of ratio estimators in complex samples: Bias and the coefficient of variation* (Research Report No. RR-06-19). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2006.tb02025.x

Phillips, G. W. (2014). *Linking the 2011 National Assessment of Educational Progress (NAEP) in Reading to the 2011 Progress in International Reading Literacy Study (PIRLS)*. Washington, DC: American Institutes for Research. Retrieved from ERIC database. (ED545246)

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

Robin, F., Holland, P., & Hemat, L. (2006). *ICEDOG* [computer software]. Princeton, NJ: Educational Testing Service.

Shao, J. (1996). Resampling methods in sample surveys. *Statistics, 27*, 203–254.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics, 29*, 614–623.

Wang, L., Qian, J., & Lee, Y.-H. (2013). *Exploring alternative designs using reduced equating samples and shortened anchor test length* (Research Report No. RR-13-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02309.x

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

Zaman, T., & Alakus, K. (2015). Analysis of the invariance and generalizability of multiple linear regression model results obtained from Maslach Burnout scale through jackknife method. *Open Journal of Statistics, 5*, 645–651. doi:https://doi.org/10.4236/ojs.2015.57065

Appendix

Jackknifed Standard Errors of Scale Scores

Tables A1 – A3 provide the standard errors of the scale scores for Forms 1, 2, and 3, respectively. In each table, raw scores are given in the first column, followed by a column for the number of test takers having that score on the test form. The jackknifed standard error estimates were computed on the unrounded scale scores (not listed in the tables) corresponding to the raw scores from the 100 jackknife replicate samples. The next two columns show the jackknifed standard errors of the scale scores from the random (se_R) and the cluster-based (se_{ST}) groups, respectively. The last column is the ratio of the cluster-based jackknifed standard error to the random group jackknifed standard error.

Table A1. Jackknifed Standard Errors of Scale Scores of Form 1

Raw score	*N*	se_R	se_{CB}	Ratio (se_{CB}/se_R)
3	1			
4	3			
5	2			
6	3			
7	9	0.072	0.085	1.180
8	12	0.092	0.112	1.214
9	21	0.115	0.142	1.232
10	42	0.137	0.162	1.188
11	35	0.141	0.173	1.225
12	43	0.148	0.176	1.189
13	80	0.145	0.175	1.213
14	102	0.131	0.167	1.279
15	116	0.122	0.147	1.204
16	116	0.112	0.139	1.247
17	154	0.104	0.129	1.239
18	156	0.097	0.118	1.211
19	164	0.087	0.107	1.221
20	177	0.080	0.100	1.247
21	199	0.082	0.097	1.194
22	169	0.072	0.090	1.244
23	188	0.067	0.079	1.177
24	201	0.063	0.075	1.186
25	194	0.063	0.076	1.202
26	199	0.062	0.074	1.192
27	176	0.066	0.074	1.121
28	177	0.054	0.074	1.363
29	177	0.059	0.074	1.256
30	169	0.060	0.075	1.258
31	177	0.057	0.071	1.241
32	154	0.062	0.082	1.329
33	169	0.057	0.084	1.486
34	163	0.061	0.075	1.221
35	132	0.060	0.076	1.259
36	115	0.059	0.071	1.204
37	127	0.049	0.070	1.439
38	96	0.051	0.062	1.219
39	108	0.047	0.063	1.339
40	75	0.050	0.055	1.113
41	61	0.049	0.052	1.047
42	53	0.047	0.046	0.984
43	27			
44	20			
45	2			

Note. $N = 4{,}564$. The se_R and se_{CB} statistics were estimated based on the scale scores corresponding to their raw scores from the 100 jackknife replicates. The missing estimates at the top and bottom of the table were due to insufficient and inadequate sample sizes for jackknifing computation.

Table A2. Jackknifed Standard Errors of Scale Scores of Form 2

Raw score	N	se_R	se_{CB}	Ratio (se_{CB}/se_R)
5	3			
6	6			
7	6			
8	15	0.071	0.074	1.039
9	18	0.088	0.091	1.030
10	34	0.109	0.115	1.055
11	31	0.135	0.133	0.984
12	45	0.143	0.148	1.040
13	71	0.151	0.150	0.993
14	58	0.141	0.150	1.060
15	59	0.136	0.142	1.040
16	59	0.124	0.139	1.126
17	80	0.129	0.130	1.009
18	82	0.108	0.120	1.111
19	92	0.103	0.114	1.103
20	100	0.099	0.107	1.084
21	86	0.087	0.101	1.163
22	99	0.088	0.098	1.110
23	94	0.080	0.092	1.155
24	120	0.075	0.088	1.168
25	114	0.068	0.087	1.280
26	110	0.066	0.082	1.232
27	116	0.069	0.085	1.236
28	121	0.061	0.083	1.360
29	113	0.071	0.070	0.996
30	119	0.060	0.077	1.278
31	106	0.072	0.076	1.053
32	95	0.062	0.083	1.329
33	111	0.066	0.080	1.220
34	119	0.072	0.080	1.104
35	97	0.075	0.085	1.129
36	102	0.079	0.086	1.087
37	80	0.072	0.087	1.209
38	96	0.077	0.084	1.092
39	91	0.066	0.074	1.108
40	62	0.059	0.072	1.232
41	44	0.048	0.062	1.276
42	43	0.038	0.048	1.249
43	46	0.024	0.033	1.397
44	26			
45	7			

Note. $N = 2{,}976$. The se_R and se_{CB} statistics were estimated based on the scale scores corresponding to their raw scores from the 100 jackknife replicates. The missing estimates at the top and bottom of the table were due to insufficient and inadequate sample sizes for jackknifing computation.

Table A3. Jackknifed Standard Errors of Scale Scores of Form 3

Raw score	N	se_R	se_{CB}	Ratio (se_{CB}/se_R)
5	2			
6	1			
7	3			
8	10	0.066	0.072	1.096
9	11	0.085	0.093	1.100
10	16	0.108	0.114	1.058
11	25	0.126	0.143	1.139
12	24	0.130	0.145	1.110
13	44	0.137	0.151	1.101
14	34	0.134	0.148	1.104
15	45	0.132	0.147	1.107
16	54	0.128	0.141	1.102
17	70	0.116	0.132	1.136
18	94	0.112	0.122	1.086
19	75	0.106	0.117	1.099
20	89	0.095	0.102	1.066
21	93	0.087	0.103	1.189
22	97	0.076	0.088	1.156
23	125	0.080	0.080	1.005
24	151	0.073	0.077	1.059
25	123	0.061	0.075	1.225
26	130	0.062	0.073	1.171
27	139	0.058	0.067	1.155
28	147	0.057	0.059	1.033
29	161	0.060	0.065	1.079
30	151	0.060	0.063	1.042
31	143	0.059	0.063	1.070
32	163	0.061	0.064	1.042
	151	0.063	0.064	1.007
34	115	0.065	0.067	1.044
35	123	0.066	0.065	0.983
36	122	0.062	0.070	1.124
37	120	0.068	0.074	1.086
38	104	0.069	0.080	1.154
39	100	0.064	0.067	1.047
40	105	0.060	0.066	1.096
41	79	0.053	0.062	1.157
42	53	0.046	0.053	1.138
43	44	0.035	0.040	1.173
44	29	0.050	0.049	0.995
45	13			

Note. $N = 3{,}378$. The se_R and se_{CB} statistics were estimated based on the scale scores corresponding to their raw scores from the 100 jackknife replicates. The missing estimates at the top and bottom of the table were due to insufficient and inadequate sample sizes for jackknifing computation.

Suggested citation

Wang, L., Qian, J., & Lee, Y.-H. (2018). *Grouping effects on jackknifed variance estimation for item response theory scaling and equating with cluster-based assessment data* (Research Report No. RR-18-16). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12204

Action Editor: Marna Golub-Smith

Reviewers: Sooyeon Kim, Anna Kubiak, and Frederic Robin

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/