Research Report

Linking Errors in Trend Estimation in Large-Scale Surveys: A Case Study

Xueli Xu

Matthias von Davier

April 2010

ETS RR-10-10

Linking Errors in Trend Estimation in Large-Scale Surveys: A Case Study

Xueli Xu and Matthias von Davier
ETS, Princeton, New Jersey

April 2010

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Shelby Haberman

Technical Reviewers: Yue Jia, Yi-Hsuan Lee

Abstract

One of the major objectives of large-scale educational surveys is reporting trends in academic achievement. For this purpose, a substantial number of items are carried from one assessment cycle to the next. The linking process that places academic abilities measured in different assessments on a common scale is usually based on a concurrent calibration of adjacent assessments using item response theory (IRT) models. It can be conjectured that the selection of common items has a direct effect on the estimation error of academic abilities due to item misfit, small changes in the common items, position effect, and other sources of construct-irrelevant changes between measurement occasions. Hence, the error due to the common-item sampling could be a major source of error for the ability estimates. In operational analyses, generally two sources of error are accounted for in variance estimation: student sampling error and measurement error. A double jackknifing procedure is proposed to include a third source of the estimation error, the error due to common-item sampling. Three different versions of the double jackknifing were implemented and compared. The data used in this study were item responses from Grade 4 students who took the NAEP 2004 and 2008 math long-term trend (LTT) assessments. These student samples used in this study are representative samples of Grade 4 student population in 2004 and in 2008 across the US. The results showed that these three double jackknifing approaches resulted in similar standard error estimates that were slightly higher than the estimates from the traditional approach, regardless of whether an item sampling scheme was used or items were dropped at random.

Key words: linking error, trend estimates, double jackknifing, long-term trend assessments

Trend measurement and reporting is a major focus in large-scale surveys (Mazzeo & von Davier, 2008). In practice, the trend is maintained through a set of common items across adjacent assessments. If the trend estimates are interpreted within the limit of the trend items, there is no need to investigate linking errors caused by the selection of trend items. However, as pointed out by Monseur, Sibberns, and Hastedt (2008), an improvement in student performance based on the trend items is currently interpreted by report users and policy-makers as an improvement in student performance for the whole domain assessed by the study. Hence, the inclusion of a linking error component based on item sampling and student sampling in reporting trends would be consistent with how trends are presently interpreted. The selection of common items might have a direct effect on the estimation error of academic abilities, which are latent in *item response theory* (IRT) models, due to item misfit, small changes in the common items, position effect, and other factors. Consequently, the error due to the common-item sampling could be a substantial source of error for the ability estimates.

Although maintenance of a meaningful trend line is an important focus in large-scale educational surveys, the number of studies devoted to linking errors in large-scale surveys is surprisingly small. The reason might be partly due to the complexity of large-scale surveys, since most of these assessments employ *partially balanced incomplete block* (pBIB) design, stratified student sampling, IRT, and latent regression modeling to make inferences on the abilities defined in the framework for subgroups of interest. The complex sampling of items and students makes linking errors difficult to estimate and understand. In current operational analysis procedures, the student sampling uncertainty and measurement uncertainty were taken into account when calculating the estimation error of ability estimates. Cohen, Johnson, and Angeles (2001) attempted to account for the estimation error of ability estimates by considering both item and student sampling variation. A double jackknife procedure was employed to examine the effect of item sampling in addition to student sampling error. However, there is some concern about their derived formula for the standard errors (Haberman, 2005). Recently, Haberman, Lee, and Qian (2009) derived a formula for group jackknifing on both the item and student sampling. Their approach is to randomly drop one group of items and one group of students simultaneously. In fact, item jackknifing is not new. Sheehan and Mislevy (1988) looked into item jackknifing by dropping a group of equivalent items one at a time, and calculated the errors of the linear constants in the true-score equating.

Their findings were that item sampling was an important source of estimation error. In the study conducted by Michaelides and Haertel (2004), the authors pointed out that error due to common-item sampling depends not on the size of the examinee sample but on the number of common items used.

In this study, we used double jackknifing to investigate the linking error in one of the National Assessment of Educational Progress (NAEP) assessments. The data we used was the long-term trend (LTT) math data from the 2004 and 2008 administrations. A compensatory general diagnostic model (GDM; von Davier, 2005) was used to calibrate the items as well as the subgroup ability distributions. The software mdltm (von Davier, 1995) was used for item calibration and for estimating standard errors, using the jackknife procedure. The rest of this paper is organized as follows: The first section briefly introduces the GDM, the second section describes the detailed procedure of double jackknifing used in this study, and the final section shows the results and includes a brief discussion.

The Logistic Formulation of a Compensatory GDM

A logistic formulation of the compensatory GDM under multiple-group assumption is introduced in this section. The probability of obtaining a response x for item i in the multiple-group GDM is expressed as

$$P\left(X_i = x \middle| \vec{\beta}_i, \vec{q}_i, \vec{\gamma}_i, \vec{a}, g\right) = \frac{exp\left[\beta_{xig} + \sum_{k=1}^{K} x\gamma_{ikg} q_{ik} a_k\right]}{1 + \sum_{y=1}^{m_i} exp\left[\beta_{yig} + \sum_{k=1}^{K} y\gamma_{ikg} q_{ik} a_k\right]}, \quad (1)$$

where x is the response category for item i ($x \in \{0,1,...,m_i\}$); $\vec{a} = (a_1,...,a_K)$ represents a K-dimensional skill profile containing discrete, user-defined skill levels $a_k \in \{s_{k1},...,s_{kl},...,s_{kL_k}\}$ for $k = 1,...,K$; $\vec{q} = (q_{i1},...,q_{iK})$ are the corresponding Q-matrix entries relating item i to skill k ($q_{ik} \in (0,1,2...)$ for $k = 1,...,K$); the item parameters $\overrightarrow{\beta_i} = (\beta_{ixg})$ \ and $\overrightarrow{\gamma_i} = (\gamma_{ikg})$ are real-valued thresholds and K-dimensional slope parameters, respectively; and g is the group membership indicator. For model identification purposes, the researcher can impose necessary constraints on $\sum_k \gamma_{ikg}$ and $\sum \beta_{xig}$; also, with a

nonzero *Q*-matrix entry, the slopes γ_{ikg} help determine how much a particular skill component in $\vec{a} = (a_1, ..., a_K)$ contributes to the conditional response probabilities for item *i,* given membership in group g. For multiple-group models with a common scale across populations, the item parameters are constrained to be equal across groups, so that $\beta_{ixg} = \beta_{ix}$ for all items *i* and thresholds *x*, as well as $\gamma_{ikg} = \gamma_{ik}$ for all items *i* and skill dimensions *k*. It should be noted that even if the total number of ability dimension *K* and the number of levels for each dimension are moderate, the number of parameters in the discrete latent ability distribution is large for multiple-group analysis. For example, for a test measuring four dimensions and four levels specified for each dimension, the number of parameters in the latent ability distribution only to be estimated using Model 1 for four-group analysis is $4 \times (4^4 - 1) = 1020$! Xu and von Davier (2008) took further steps to reduce the number of parameters in the discrete latent ability distribution by utilizing a loglinear model to capture basic features of the discrete latent ability distribution. Specifically, the joint probability of the discrete latent ability distribution can be modeled as

$$\log(P_g(a_1, a_2, ..., a_K)) = \mu + \sum_{k=1}^{K} \lambda_{kg} a_k + \sum_{k=1}^{K} \eta_k {a_k}^2 + \sum_{i \neq j}^{K} \delta_{ij} a_i a_j \ , \qquad (2)$$

where μ, λ_{kg}, η_k, and δ_{ij} are parameters in this loglinear smoothing model, and g is a group index (Haberman, von Davier, & Lee, 2008; Xu & von Davier, 2008).

Data and Model

In this study, the LTT math assessment data were used for illustration to examine the difference between the double jackknifing and student-jackknifing. A couple of features in LTT mathematics assessment appear to make this database an ideal starting point for explorations with a double jackknife approach. One feature is that the assessment framework of the NAEP LTT defines the target of interest as a unidimensional ability variable. The other feature is the number of items taken by each student. Although a pBIB design is employed in the LTT math assessment, each student took about 50 items on average, which makes the LTT assessment a reasonably long test for an educational survey. This implies that student ability

can be estimated rather accurately, even without using latent regression models commonly applied to borrow information in shorter assessments. Hence, in this study, we did not use the latent regression model. Instead, we used a multiple-group GDM to calibrate the items and estimate the latent ability distributions for subgroups of interest. Due to the design of the LTT math assessments, a simplified version of the multiple-group GDM in Models 1 and 2 can be applied. Specifically, for each item i there are only two categories, $x = \{0,1\}$, and the number of skill dimensions is $K = 1$. In addition, in this study, 31 quadrature points, distributed evenly from -4 to 4, were specified for the ability dimension. Hence, the model used in this particular study is written as

$$P\left(X_i = 1 \middle| \vec{\beta}_i, \vec{q}_i, \vec{\gamma}_i, \vec{a}, g\right) = \frac{exp[\beta_{1i} + \gamma_i q_i a]}{1 + exp[\beta_{1i} + \gamma_i q_i a]} \tag{3}$$

and

$$log(P_g(a)) = \mu + \lambda_g a + \eta a^2, for \alpha = \{-4, -3.7333, ..., 3.7333, 4\}.$$

It is noted that the group indicator g is dropped in Model 3 to calibrate all subgroups of interest on the same scale.

It is well known that identifiability is a concern in IRT models, and this also applies to the GDM, which is used in our study as a general modeling framework that includes IRT as a special case. One prerequisite of identifiability in IRT models is that the indeterminacy of the IRT scale is removed. In order to achieve this, we fixed the mean and standard deviation of the ability distribution of the groups defined by ethnicity in the 2004 assessment to 0.0 and 1.0, respectively. We chose to use the White ethnicity group assessed in 2004 as the reference group to the indeterminacy of the IRT scale. (This can be changed for future research, depending on the purpose of the study.) For our purposes, we needed an arbitrarily-chosen reference group so that the means of the other 2004 and the 2008 groups could be interpreted in terms of differences to this reference group.

The data used in this study were item responses from Grade 4 students who took the NAEP 2004 and 2008 math LTT assessments. These student samples used in this study are representative samples of Grade 4 student population in 2004 and in 2008 across the US. The sample sizes for these two assessments are approximately 8,000 and 7,200, respectively.

There were six blocks within the 2004 and 2008 administrations, and five of them were trend blocks (i.e., five of the six blocks administered in 2004 were also administered in 2008.) This resulted in 112 trend items across these two administrations. Most students in the 2008 assessment had taken two trend blocks. Each block contained about 20 to 26 items.

Double Jackknifing in LTT Math Assessment

The operational set of replicate weights was used for the student jackknifing. These weights were developed by first forming 62 pairs of *primary sampling units* (PSU). The two PSUs within each pair were assumed to be similar to each other in terms of their background features. Then, the jackknife samples were created by randomly dropping one PSU in one pair by assigning zero weight, and assigning double weight to the other PSU within this pair. Consequently, we obtained 62 weights for each student.

Three approaches were employed to conduct the item jackknifing. The first approach was to create the jackknife samples by randomly selecting one item for each trend block and dropping these items. This yielded 23 jackknifing samples. This approach is referred to as *random-item jackknifing*. The second approach was to create the jackknife samples by first grouping the items into five groups within each trend block, based on their discrimination parameter estimates obtained from using original full data, and then dropping one such group at a time. This also yielded 23 jackknifing samples. This approach is referred to as *A-item jackknifing*. The third approach was similar to the second approach, only this time the grouping was based on the difficulty parameter estimates. This approach is referred to *B-item jackknifing*. The purpose of the second and third approaches was to examine the relationship between the item characteristics and the estimation error of group ability estimates.

Double jackknifing is a combination of *student jackknifing* and *item jackknifing*. Specifically, for each jackknife sample, one of the 62 sets of weights was used, and five trend items on average were dropped from the assessment. Then, the jackknifed sample of 2004 and 2008 assessments were calibrated concurrently to putting these assessments onto the same scale. Thus, for each approach (random-item jackknifing, A-item jackknifing and B-item jackknifing), there were 62*23 concurrent calibrations, for which the group mean and variance estimates were produced.

Analysis and Results

Table 1 presents the subgroup mean estimates of these two assessment years across the three different jackknifing schemes. One can observe that different jackknifing schemes lead to mean estimates close to those from using the full data set. Table 2 shows the linking error under different jackknifing schemes. The linking error was calculated using the formula derived by Haberman (2005), "Let θ be the true values for statistics of interest, such as group mean and standard deviation, and let $\tilde{\theta}_{ij}$ be the estimate by dropping one group of items (indexed by i) and one group of students (indexed by j)" (p. 2). Then, the jackknife estimate can be written as

$$\tilde{\theta} = \sum_i \sum_j \tilde{\theta}_{ij} / IJ , \tag{4}$$

where I, J are the total number of jackknife groups for items and students, respectively. Let $d_{ij} = \tilde{\theta}_{ij} - \tilde{\theta}$, then we have

$$\begin{aligned} \overline{d_{i.}} &= \frac{\sum_j d_{ij}}{J} \\ \overline{d_{.j}} &= \frac{\sum_i d_{ij}}{I} \\ e_{ij} &= d_{ij} - \overline{d_{i.}} - \overline{d_{.j}} . \end{aligned} \tag{5}$$

Finally, the jackknife error from the double jackknifing is calculated by

$$\sigma^2_{d-jack} = \frac{I-1}{I} \sum_i \overline{d_{i.}}^2 + \frac{J-1}{J} \sum_j \overline{d_{.j}}^2 - \frac{(I-1)(J-1)}{IJ} \sum_i \sum_j e_{ij}^2 . \tag{6}$$

The jackknife error estimate from *student jackknifing only* is estimated from a different procedure. That is, no item is dropped to form a jackknife sample. Instead, 62 jackknife samples with different sets of student replicate weights are formed and used to estimate the jackknife error. (A total of 62 samples were selected in NAEP operational

analysis by design.) Specifically, the jackknife error from *student jackknifing only* is calculated by aggregating these 62 squared differences,

$$\sigma_{student-jack} = \sum_{t=1}^{62} \left(t_i - \bar{t} \right)^2, \tag{7}$$

where t_i denotes the estimator of the parameter obtained from the i^{th} jackknife sample and $\bar{t}$ is the average of t_i s (Qian, Kaplan, Johnson, Krenzke, & Rust, 2001). For further discussion of the variance estimation procedure used by NAEP, interested readers may refer to the paper by Johnson (1989).

Table 1 presents the estimates of ability means by subgroups defined by ethnicity across the 2004 and 2008 assessment cycles obtained in a joint calibration. Recall that the estimates obtained with the student-only and the three double-jackknifing schemes are based on constraints that set the mean of the 2004 White group to 0.0 and the standard deviation of that group to 1.0.

Table 1

The Group Mean Estimates From Different Sampling Schemes

Group	Original skill mean	Student jackknifing with all items Skill mean	Student jackknifing with random-item jackknifing Skill mean	Student jackknifing with A-item jackknifing Skill mean	Student jackknifing with B-item jackknifing Skill mean
2004 White	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]
2004 Black	-0.683	-0.683	-0.683	-0.683	-0.683
2004 Hispanic	-0.476	-0.476	-0.476	-0.476	-0.476
2004 Asian	0.604	0.604	0.605	0.605	0.604
2008 White	0.159	0.159	0.159	0.159	0.159
2008 Black	-0.592	-0.592	-0.592	-0.592	-0.592
2008 Hispanic	-0.346	-0.346	-0.346	-0.346	-0.346
2008 Asian	0.723	0.723	0.723	0.724	0.723

[a]These numbers were fixed to 0 to make the model identifiable.

As shown in Table 2, the error associated with a particular subgroup mean is similar across different double jackknifing schemes. Moreover, the estimation error produced by

double jackknifing is slightly larger than that produced by student-sample-only jackknifing. Note that the reference group is 2004 White, so there are no estimates available for this group.

Table 2

The Standard Error of Group Mean From Different Sampling Schemes

Group	Student jackknifing with all items	Student jackknifing with random-item jackknifing	Student jackknifing with a-item jackknifing	Student jackknifing with b-item jackknifing
2004 White	–	–	–	–
2004 Black	0.066	0.067	0.068	0.070
2004 Hispanic	0.046	0.051	0.054	0.055
2004 Asian	0.139	0.140	0.145	0.142
2008 White	0.061	0.060	0.062	0.063
2008 Black	0.056	0.058	0.060	0.061
2008 Hispanic	0.050	0.053	0.054	0.056
2008 Asian	0.116	0.118	0.120	0.121

Table 3 presents the estimates of group variances across years under different jackknifing schemes. For a particular subgroup, the jackknife estimates are similar to each other and are close to the estimates using the full set of items.

Table 3

The Group Standard Deviation Estimates From Different Sampling Schemes

Group	Original skill SD	Student jackknifing with all items Skill SD	Student jackknifing with random-item jackknifing Skill SD	Student jackknifing with a-item jackknifing Skill SD	Student jackknifing with b-item jackknifing Skill SD
2004 White	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]
2004 Black	0.919	0.919	0.919	0.919	0.919
2004 Hispanic	0.977	0.977	0.977	0.978	0.977
2004 Asian	1.180	1.180	1.182	1.182	1.181
2008 White	1.027	1.027	1.028	1.027	1.028
2008 Black	0.985	0.985	0.985	0.985	0.985
2008 Hispanic	0.956	0.955	0.956	0.956	0.956
2008 Asian	1.327	1.327	1.328	1.329	1.327

[a]These numbers were fixed to 1 to make the model identifiable.

Table 4 presents the standard error of the estimated standard deviation under different jackknifing schemes. Note that the reference group is 2004 White, so there are no estimates available for this group.

Table 4

The Standard Error of Group Standard Deviation Estimates From Different Sampling Schemes

Group	Student jackknifing with all items	Student jackknifing with random-item jackknifing	Student jackknifing with a-item jackknifing	Student jackknifing with b-item jackknifing
2004 White	–	–	–	–
2004 Black	0.032	0.039	0.048	0.039
2004 Hispanic	0.034	0.041	0.044	0.038
2004 Asian	0.062	0.063	0.064	0.062
2008 White	0.027	0.029	0.027	0.029
2008 Black	0.038	0.039	0.047	0.044
2008 Hispanic	0.028	0.030	0.028	0.028
2008 Asian	0.064	0.067	0.078	0.079

As shown in Table 4, the estimation errors obtained from using double jackknifing are similar to those obtained from using other approaches and are, in most cases, slightly larger than the estimation error obtained from the one-sided jackknifing with the student sample.

Discussion

The results for the LTT data showed that the double jackknife is feasible and results in slightly increased estimates of standard errors of ability distribution parameters. Note, however, that NAEP LTT data were chosen for a number of reasons, first and foremost to obtain information about the feasibility of the double jackknife approach using a relatively long assessment instrument. The LTT data are characterized by observations that contain 50 responses on average per student, which is on the high side when compared to other large-scale survey assessments. In shorter assessments, the differences across approaches may look more dramatic, in the sense that a double jackknife with dropping 5/50 of the item set did not produce substantially increased errors.

The good news is that the increase did, under the conditions outlined, not depend on the specific selection of items to be dropped. More specifically, the jackknife schemes that dropped items according to their discrimination (or difficulty) parameters did not result in

inflated jackknife estimates of standard errors compared to a random selection of dropped items. This implies the LTT mathematic assessment linkage is robust, so researchers can have confidence in interpreting the improvement of student performance in these assessments as an improvement for the whole domain assessed by the NAEP study.

Note that this research has used the comprehensive reestimation of all parameters of the multiple group IRT model as described in Hsieh, Xu, and von Davier (2009). A less comprehensive approach like the one currently used operationally may have resulted in a larger difference between full item set and double jackknife. Further research is needed in this direction, as well as research on the effect of dropping items from shorter scales, or double jackknifing in models with multidimensional ability variables.

References

Cohen, J., Johnson, E., & Angeles, J. (2001, April). *Estimates of the precision of estimates from NAEP using a two-dimensional jackknife procedure.* Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

Haberman, S. (2005). *Jackknifing and two-dimensional sampling.* Unpublished manuscript. Princeton, NJ: ETS.

Haberman, S. J., von Davier, M., & Lee, Y. H. (2008). *Comparison of multidimensional item response models: multivariate normal ability distributions versus multivariate polytomous ability distributions* (ETS Research Rep. No. RR-08-45). Princeton, NJ: ETS.

Haberman, S., Lee, Y., & Qian, J. (2009). *Jackknifing techniques for evaluation of equating accuracy* (ETS Research Rep. No. RR-09-39). Princeton, NJ: ETS.

Hsieh, C., Xu, X., & von Davier, M. (2009). Variance estimation for NAEP data using a resampling-based approach: An application of cognitive diagnostic models. *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments, 2,* 161–174.

Johnson, E. G. (1989). Considerations and techniques for the analysis of NAEP data. *Journal of Educational Statistics*, *14*(4), 303–334.

Mazzeo, J., & von Davier, M. (2008). *Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results.* Retrieved from http://www.ierinstitute.org/IERI_Monograph_Volume_01_Chapter_6.pdf

Michaelides, M. P., & Haertel, E. H. (2004). *Sampling of common items: An unrecognized source of error in test linking.* Los Angeles, CA: Center for the Study of Evaluation, University of California.

Monseur, C., Sibberns, H., & Hestedt, D. (2008). Linking errors in trend estimates for international surveys in education. *IERI Monograph Series: Issues and Methodologies in Large-scale Assessments,1*, 113–122.

Qian, J., Kaplan, B. A., Johnson, E. G., Krenzke, T., & Rust, K. F. (2001). Weighting procedures and estimation of sampling variance for the national assessment. In N. A. Allen, J. R. Donoghue, & T. L. Schoeps (Eds), *The NAEP 1998 technical report* (NCES 2001-452). Washington, DC: U.S. Department of Education, Institute of Education Sciences, Department of Education, Office for Educational Research and Improvement.

Sheehan, K., & Mislevy, R. (1988*). Some consequences of the uncertainty in IRT linking procedures* (ETS Research Rep. No. RR-88-38). Princeton, NJ: ETS.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Rep. No. RR-05-16). Princeton, NJ: ETS.

Xu, X., & von Davier, M. (2008). *Fitting the structural general diagnostic model to NAEP data* (ETS Research Rep. No. RR-08-27). Princeton, NJ: ETS.