The Alberta Journal of Educational Research *Vol. XLIX, No. 3, Fall 2003, 252-263*

Richard G. Wolfe

and

Jennifer L. Dunn

Ontario Institute for Studies in Education of the University of Toronto

The Jackknife and Multilevel Modeling: A New Application of an Old Trick

In this article the authors demonstrate two instances where the jackknife can be used to enhance hierarchical linear model (HLM) analyses. The jackknife was used to improve the HLM estimates of composite measures by jackknifing over items. The first study examined fixed-effects and variance component estimation. The jackknife appeared to reduce the bias in the estimates both of slopes and of variances by implicitly adjusting for item-by-person and item-by-group interactions. The second study examined the utility of the jackknife as a multilevel item analysis tool. The results suggest that pseudovalues offer a unique opportunity for isolating item variability in multilevel data. The jackknife seems to offer enhancements and insights to conventional HLM analyses.

In recent years practice in psychometrics has shifted from classical approaches to more complicated analyses. As the sophistication of our techniques increases, lessons learned from classical principles may be lost. In advanced analyses such as linear structural equation modeling or multilevel modeling (including hierarchical linear modeling [HLM]), it is easy to get sidetracked by the intricacies of the models and the complexities of the estimations and to forget that the original measures are errorful and not true scores. As a result, we may fail to take into account the biasing effects that measurement error can have on estimation and interpretation. For example, multiple indicators are often built into structural equation or HLM models in an attempt to separate the effects of latent variables from those of uncorrelated measurement errors. However, the use of these indicators requires serious consideration of relevant measurement techniques and error structures, especially the relationship among items, constructs, and the item sample size. Any variable can include error and these errors can cumulate and correlate, producing bias. Classical approaches are designed specifically to address error structures at both test and item levels and, therefore, could theoretically be used to improve sophisticated analyses.

The purpose of this article is to meld classical theories and methods with advanced models and estimation procedures. Specifically, the utility of the jackknife in HLM analyses of composite measures is examined. The jackknife is used to try to improve fixed-effects estimation, to find true-score variance

---

Richard Wolfe is an associate professor and Coordinator of the Measurement and Evaluation Program.

Jennifer Dunn is a doctoral candidate in the Measurement and Evaluation Program.

components by implicitly adjusting for item-by-person and item-by-group interactions, and to quantify item variability at both student and classroom levels. Special attention is directed to the sizes and locations of the interactions and the implications of these findings for educational measurement.

Hierarchical Linear Modeling

The hierarchy inherent in educational data makes such data appropriate for HLM analyses, as HLM analyzes regression equations at multiple levels simultaneously (Raudenbush & Byrk, 2002). For example, consider a two-level analysis of student achievement with students nested within classrooms. A multiple regression model relates independent student variables (e.g., gender, initial knowledge) and the outcome variable (final student achievement). Independent classroom-level variables (e.g., teacher experience, classroom resources) influence the expected student achievement and also modify the relative effects of student-level variables on achievement. HLM also estimates the overall outcome variability and the variability of the effects.

Although HLM techniques allow for sophisticated modeling of educational data, the resulting estimates remain, from a psychometric perspective, point estimates that are influenced by systematic measurement error and are probably biased. The measurement error is conceptualized as arising because most educational indicators are composites of individual variables: that is, they are constructed from item samples. Specifically, the dependent variables—and often independent variables—are test scores based on responses to multiple items, and each item response contains measurement error. Computing test scores as averages of the item responses reduces the error relative to the error of single items, but is still dependent on item heterogeneity and on test length. In an HLM analysis, item-by-person interactions or item-by-group (classroom) interactions are subsumed in the variances of the composite test scores, but are confounded with and therefore may bias the estimates of structural coefficients and variances.

Jackknife

The jackknife originally developed by Quenouille (1956) and later modified by Mosteller and Tukey (1977) can be used to quantify the amount of bias in an estimate, reduce the bias in the estimate, and finally place confidence intervals around the new unbiased estimate. Essentially, the jackknife is a re-sampling technique that uses multiple estimates based on subsets of the original observations to correct for the bias. The unknown estimates are first calculated using the entire sample and then recalculated using parts of the sample. Weighted combinations of the subsample estimates and the original estimate are used to create a jackknife estimate in which the magnitude of bias will be reduced (Miller, 1964; Quenouille, 1956). The jackknife works to remove sampling bias and can expand the applications of procedures that would normally be restricted to large samples or samples that meet normality assumptions (Rogers, 1976).

Jackknife methods are believed to provide direct numerical approximations of both bias and standard error and to give reliable confidence limits. In a comprehensive review of the jackknife, Miller (1974) noted successful application of the jackknife to ratio estimators, *u*-statistics, regression estimates,

variance inferences, and multivariate analyses. He went on to hypothesize successful use of the jackknife in linearly transformed order statistics, correlation coefficients, and outlier analyses. Since that review, jackknife principles have also been applied to factor loadings, multiple matrix sampling, reliability, and maximum likelihood estimates (Brennan, Harris, & Hanson, 1987; Huitema, & McKean, 1994; Wilcox, 1997). In terms of confidence interval applications, the jackknife has been shown to be robust against non-normality, useful with small sample sizes, and a valuable tool for the internal replication of a study (Pandey & Hubert, 1975; White, 2000).

Objectives

Melding HLM and jackknifing techniques offers multiple advantages. First, the jackknife has been proven somewhat useful for removing the bias in the variance of student achievement scores (Kifer, Wolfe, & Schmidt, 1993; Kifer & Wolfe, 1986; Miller, 1974). Second, the jackknife has been shown to be a useful procedure for estimating the true score variance of an individual item (Schmidt & Wolfe, 1983). The pseudovalues provide a form of item analysis that can help to investigate the influence of each item on the other items. In HLM, because analyses are being conducted on multiple levels, the situation is further complicated. Biased estimates have the potential to arise at each of the levels. We performed two studies to investigate whether the jackknife can be used (a) to remove the bias, and (b) to support item analyses at multiple data levels.

Study 1

Purpose

Theoretically, by jackknifing over items, the bias in HLM variance component estimates resulting from composite measurements will be reduced. In HLM, individual systematic errors can also influence the accuracy of estimates on multiple levels. Specifically, when items are administered to multiple students across multiple classrooms, the item response systematic errors can lead to biased estimates at both student and classroom levels. The purpose of the first study is to investigate the feasibility of applying the jackknife to HLM variance estimates.

Data Source

The IEA Second International Mathematics Study (SIMS; Burstein, 1993) was the first—and is so far the only—major international educational achievement survey to have collected data at more than one point in time on the same students. In six other countries and two Canadian provinces, the same students, classrooms, and schools were sampled at the beginning and at the end of grade 8. The students were tested in mathematics knowledge and skill. These achievement data, together with a rich array of questionnaire data from students, teachers, and schools, provide a basis for studying cognitive growth and its correlates.

To investigate the effects of measurement errors in composite measures on HLM analyses, a set of items, 8 algebra and 8 geometry, from the "core" SIMS test form was used. These items were answered by all students at the beginning and the end of the school year. In addition, at the end of the year the classroom teachers answered an opportunity to learn (OTL) question about the items: Did the students in the classroom have the opportunity to learn the mathematics

necessary to answer the question? OTL was incorporated into the model as a classroom-level variable whereas student gender was considered a student-level demographic variable. The analysis was conducted for four of the SIMS populations: France (7,226 students in 335 schools), Ontario (Canada, 3,877 students in 161 schools), New Zealand (4,469 students in 175 schools), and the United States (4,846 students in 284 schools).

HLM

A conventional HLM analysis of these data would consider the end-of-year achievement (T_2 achievement) as the dependent variable and the beginning-of-year achievement (T_1 achievement) and student gender (sex; 0 for female, 1 for male) as independent student-level variables. OTL is considered an independent classroom-level variable. The intercept and slope coefficients in the regression of T_2 achievement on T_1 achievement and sex would be considered to vary from class to class in part systematically depending on OTL and in part randomly. This analysis, following the notation of Byrk and Raudenbush (2002), is as follows:

$$T_{2ij} = \beta_{0j} + \beta_{1j} * \text{SEX} + \beta_{2j} * T_{1ij} + r_{ij},$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01} * \text{OTL} + u_{0j},$$
$$\beta_{1j} = \gamma_{10} + u_{1j},$$
$$\beta_{2j} = \gamma_{20} + u_{2j}.$$

The first formula is the structure of the regression within classroom, and says that student i's achievement at time 2 (T_{2ij}) is a function of the average achievement (β_{0j}) in classroom j, the gender effect (β_{1j}) in classroom j, the time 1 achievement effect (β_{2j}) in classroom j, and random student error (r_{ij}). The next formula says that the average within-classroom achievement (β_{0j}) depends on the grand mean achievement (γ_{00}), the effect of OTL (γ_{01}) and random classroom error (u_{0j}). The last two formulas state that the regression coefficients for *SEX* and T_1 achievement are dependent on the average gender effect (γ_{10}) and average T_1 achievement effect (γ_{20}) and random variation (u_{1j} and u_{2j}) respectively. All the independent variables were group centered and the classroom averages are set at the female level of 0. Although all the items were fitted with the same model, the algebra and geometry items were analyzed separately.

Jackknife

The three variables T_2 achievement, T_1 achievement, and OTL are composites over each set (algebra and geometry) of eight items. They have measurement error because those eight items are a sample from a potentially much larger set of items, and potential item-by-student interactions (some items were better understood by some students at the beginning of the year or were more frequently learned), item-by-classroom interactions (some items were on average better understood, learned more and/or taught more in certain classrooms), and item-by-variable effect (T_1 achievement, sex, and OTL) interactions.

To disaggregate some of these measurement effects, we jackknifed *over items*. That is, the HLM model is analyzed first using the composite measures for all eight items, then, in turn, one at a time, each item is dropped and the analysis is rerun. (Note that the metric for the composite measures is kept

constant by rescaling the seven-item scores to an eight-item total.) Pseudovalues are constructed by subtracting the estimates obtained for each of the replicated analyses (8 analyses of 7 items) from the first analysis estimates (all 8 items). The average of these pseudovalues represents the jackknifed (unbiased) estimate. The pseudovalues were also used to compute standard errors and t-statistics for the jackknifed estimates as a measure of the precision and significance.

The entire procedure was repeated for each test (algebra and geometry) and each of the four countries.

Results

The results for each country are presented in Table 1. The estimates from the conventional HLM analysis (including all 8 items) are labeled as *ALL*. The jackknife results (based on the average difference between *ALL* and the 7-item analyses) are labeled *JK*. They represent estimates of what the HLM parameters and variances would be if we were able to use a composite measure that included the *population* of items from which our eight items is a sample. In psychometric terms, the estimates are "disattenuated for unreliability." The original HLM (*ALL*) and jackknifed (*JK*) HLM estimates are plotted in Figure 1.

Discussion

Across geometry and algebra and across the four countries, there is consistent evidence that a conventional analysis underestimates the slope (γ_{20}) of T_1 achievement. In all cases, except geometry for the US, the jackknifed slope is higher than the original slope. This means that the errors in the student scores T_1 and T_2 were attenuating their correlations. (The US geometry result is anomalous, as is clear in Figure 1, and requires further investigation because it is producing negative error variances.) From our psychometric perspective, we are obtaining information that the slopes in the population of items are higher than they seem in the sample of items.

The variances in the slopes of the initial testing (u_{2j}) are small, but they are consistently larger when estimated with the jackknife. This means that the item slopes vary in different ways and that the jackknife has removed some bias.

The influence of OTL (γ_{01}) is about the same with the original and the jackknife estimation, so item-by-classroom interactions in OTL affecting achievement are not pronounced.

There is only a significant overall SEX effect (γ_{10}) for geometry in France. There is significant variability over classrooms in the SEX slopes (u_{1j}), for all countries and both subjects except geometry in Ontario. None of the gender coefficients is very different between the original and the jackknifed estimates, indicating that the effects are homogeneous over items.

The jackknifed error variances for students (r_{ij}) and classrooms (u_{0j}) are always lower than in the original HLM analysis. This means that for both classrooms and students within classrooms, the item-based measurement error was inflating the variance estimates. Generally, the inflation was greater for individuals, suggesting that individual-by-item interactions are relatively greater than classroom-by-item interactions. Our jackknifed variances are estimates of the variability if a long test were used. The implication is that with

Table 1
SIMS Original and Jackknife HLM Analyses by Country and Topic

	Geometry				*Algebra*			
	All	*JK*	*JKSE*	*JKT*	*All*	*JK*	*JKSE*	*JKT*
France								
Grand Mean	1.26	.69	.182	3.78	2.44	1.74	.441	3.95
OTL	.19	.24	.060	3.97	.14	.16	.047	3.42
Sex Slope	.35	.28	.087	3.20	.24	.20	.133	1.52
T_1 Achievement Slope	.44	.63	.063	10.01	.38	.54	.048	11.41
Classroom variance	.62	.51	.082	6.17	.71	.61	.082	7.44
Sex variance	.37	.29	.071	4.14	.33	.28	.083	3.35
T_1 variance	.09	.11	.051	2.15	.09	.18	.057	3.07
Student variance	1.41	.92	.066	13.83	1.43	.92	.085	10.83
Ontario								
Grand Mean	1.88	1.27	.300	4.23	1.56	.91	.375	2.44
OTL	.21	.23	.044	5.24	.19	.22	.068	3.19
Sex Slope	.08	.03	.097	.27	−.05	−.06	.076	−.83
T_1 Achievement Slope	.49	.66	.039	17.00	.45	.63	.044	14.16
Classroom variance	.69	.62	.110	5.68	.70	.73	.177	4.10
Sex variance	.15	.00	.085	.06	.38	.36	.054	6.73
T_1 variance	.11	.14	.037	3.83	.15	.21	.026	8.03
Student variance	1.59	1.12	.042	26.43	1.51	1.02	.084	12.12
New Zealand								
Grand Mean	1.84	1.15	.449	2.57	1.77	1.20	.268	4.46
OTL	.16	.18	.074	2.46	.13	.13	.052	2.43
Sex Slope	−.02	−.09	.073	−1.28	.02	.01	.109	.12
T_1 Achievement Slope	.54	.74	.044	16.92	.43	.63	.068	9.30
Classroom variance	.84	.70	.137	5.10	.56	.31	.174	1.81
Sex variance	.45	.45	.113	4.00	.27	.20	.137	1.50
T_1 variance	.08	.09	.026	3.38	.10	.14	.035	4.07
Student variance	1.52	1.02	.031	32.56	1.47	.95	.075	12.69
United States								
Grand Mean	1.27	−1.86	.287	−6.47	1.25	.79	.351	2.26
OTL	.23	.43	.052	8.23	.28	.28	.038	7.46
Sex Slope	.03	−.04	.036	−1.15	−.04	−.05	.085	−.64
T_1 Achievement Slope	.39	.15	.042	3.53	.36	.51	.044	11.64
Classroom variance	.85	−.27	.160	−1.67	.92	.79	.175	4.54
Sex variance	.24	.31	.110	2.85	.25	.25	.090	2.79
T_1 variance	.15	.16	.022	7.46	.11	.17	.041	4.22
Student variance	1.37	−.03	.128	−.26	1.48	1.00	.081	12.37

such a "true score" the importance of the between-classroom variances increases relative to the between-student variances.

Discussion

The jackknife was found consistently to reduce all estimates of variability and a few of the slope estimates. According to the results of this study, the greatest amount of bias occurs in the level 1 variance estimates and is probably due to

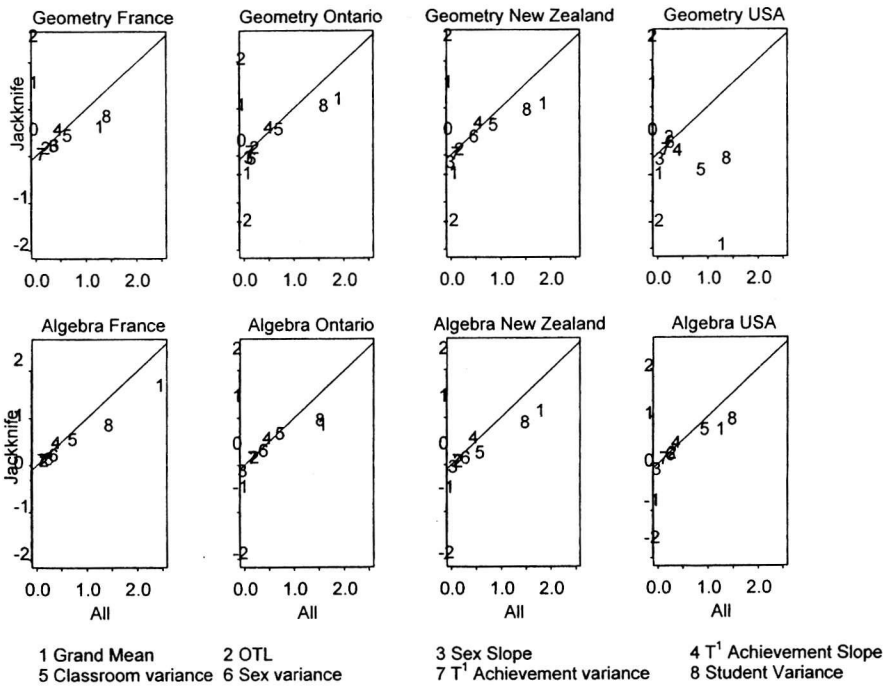

Figure 1. SIMS original (All) versus jackknife estimates for HLM parameters by country and topic.

item-by-person interactions. The classroom level variance estimates appear slightly biased; however, the amount of bias is quite small.

Study 2

The first study expanded the application of the jackknife to include the quantification and removal of bias in HLM estimates for multilevel data. The pseudovalues (produced by the jackknife procedure) have been shown to be a useful item analysis tool for estimating the true score variance of an individual item (Schmidt & Wolfe, 1983). In HLM these item relationships have the potential to vary across multiple levels. Theoretically a "good" item at the student-level may be a "poor" item at the school level. Alternatively, poor items at the student level may function well at the school level. The purpose of the second investigation is to determine the usefulness of combining the jackknife and HLM for multilevel item analyses.

Data Source

The Third International Mathematics and Science Study (TIMSS) was conducted in 1994-1995 to compare student performance across more than 40 countries. In addition to achievement items, the students also responded to questions about classroom practices. The questions about practices in mathematics classrooms were either oriented toward student behaviors or teacher behaviors. For example, the students were asked how often they copied notes from the board or discussed completed homework and how often the teacher explained the rules or questioned students' knowledge. Questions were

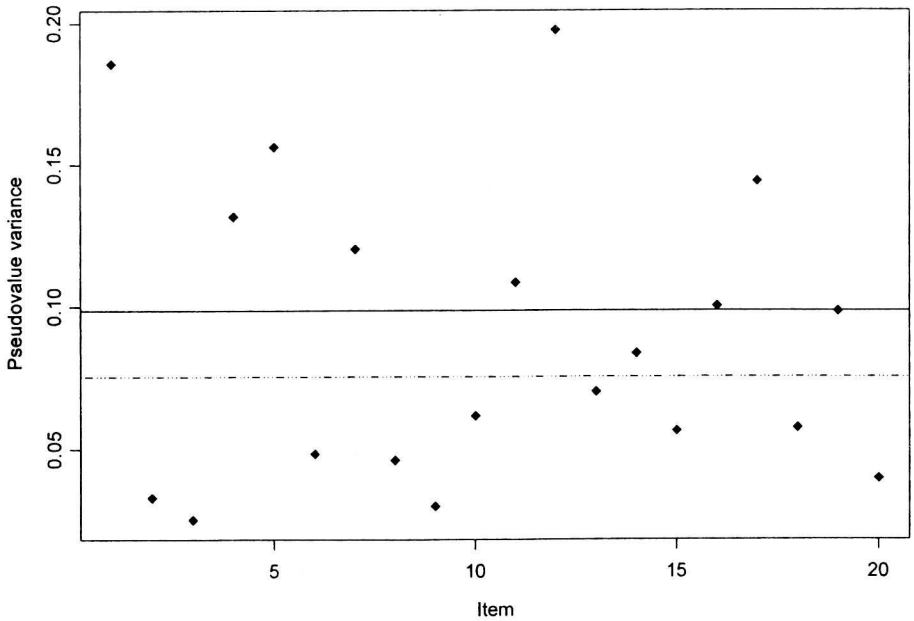

Figure 2. The distribution of Level 1 (student) pseudovalues.

designed on a 4-point Likert-type scale ranging from almost always (1) to never (4).

Although Study 1 was able to analyze data from four countries, the sampling method employed in TIMSS (one classroom per school) resulted in classroom effects being confounded with school effects, thereby limiting the data appropriate for an HLM item analysis. Sweden was one of a few countries to select multiple classrooms within each school and the only country to select multiple classrooms from the same academic stream. The responses of 916 grade 8 students with no missing data from 32 schools in Sweden were used in this study. A total of 20 opinion items were examined.

The item responses were averaged to form an overall score. In addition, 20 subscale scores were formed by dropping one item at a time. For example, Subscale 1 was the average of Items 2 through 20; Subscale 2 was the average of all items except Item 2; Subscale 3 was the average of all items except Item 3.

HLM

A two-level HLM analysis was conducted on the total scale (including all items) with no additional variables. The main purpose of this study was to examine the feasibility of performing multilevel item analyses using the jackknife and HLM. The influence of additional independent variables was not of primary interest. Therefore, a simple variance decomposition analysis was sufficient. The HLM model was applied 20 additional times using the different subscale scores as the dependent variable.

Jackknife

Two variance component estimates were obtained for each of the 21 HLM analyses: one for the student level and one for the school level. Because variance component estimates theoretically have a lower bound of zero, but

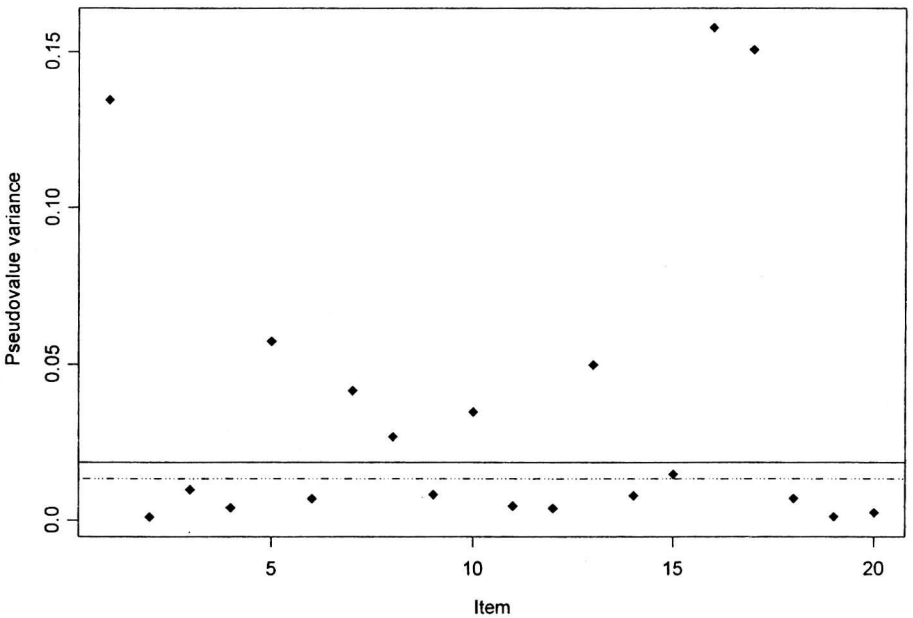

Figure 3. The distribution of Level 2 (school) pseudovalues.

computationally negative variances can result, the log of the estimates was taken before implementing the jackknife formula. Pseudovalues (*Pseudo*) were calculated for each item by taking the weighted difference between the log of the subscale variance estimates ($\sigma^2_{(i)}$; i.e., σ^2 estimated without item *i*) and the log of the overall score variance estimate:

$$Pseudo_i = k \times \log(\sigma^2) - (k-1) \times \log(\sigma^2_{(i)})$$

The average of these pseudovalues is the jackknifed estimate of the variance. Finally, the pseudovalues and jackknifed estimates were transformed back to the original scale by taking the antilog.

Item-Total Correlations

Item-total correlations are commonly used to assess item properties. However, they are limited in an HLM context because they do not examine item performance on multiple levels unless aggregated. However, in order to make adequate comparisons, classical corrected item-total correlations were calculated at the student-level and analogous corrected item-total correlations at the classroom level were calculated using classroom item means. These results were then compared with the item pseudovalues at each of the two levels.

Results

The resulting variance component estimates are given in Table 2 and in Figures 2 and 3. The solid horizontal line in each figure represents the HLM variance estimate, whereas the dotted line represents the jackknifed estimate. The largest difference between the HLM estimate and the jackknifed estimate occurred at the student level. A small amount of difference was also evident between the school-level HLM and jackknifed estimates. These results support the results of Study 1.

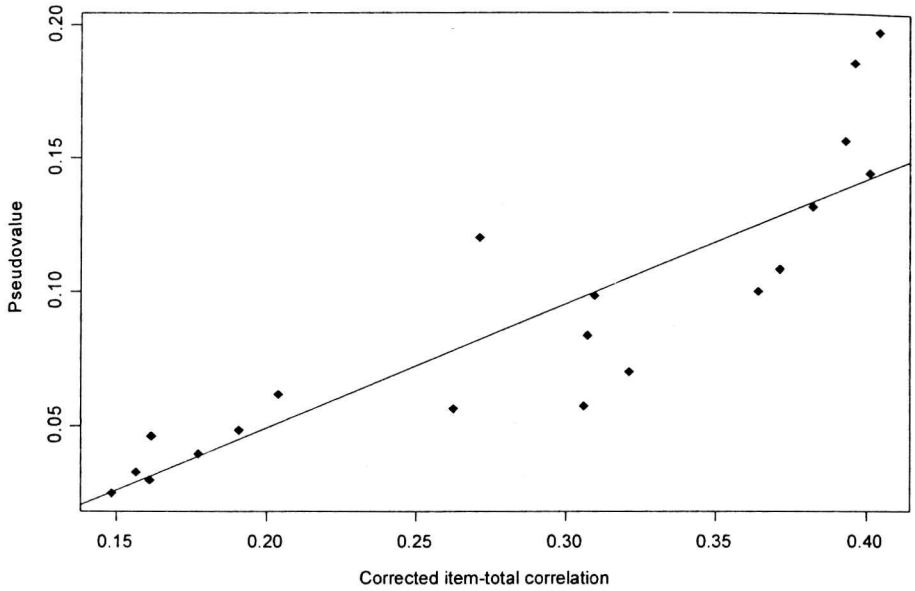

Figure 4. The relationship between item pseudovalues and corrected item-total correlation at Level 1 (student).

The student- and school-level item pseudovalues are also displayed in Figures 2 and 3 respectively. The student-level pseudovalues are much more variable than the school-level variables. Despite the increased variability, the item pseudovalues appear to have a fairly even dispersion.

Further analyses of the item pseudovalues were conducted by comparing them with appropriate corrected item-total correlations. A robust least squares regression was plotted for both teacher- and school-level data. Specifically, the pseudovalue was regressed on the item-total correlation. The pseudovalues and the corrected item-total correlations are displayed in Figures 4 and 5. Although both figures show positive correlations, the relationship at the student level is much stronger (r = .46) than the relationship at the school level (r=.06). The reduced correlation at the school level is appropriate when one considers the reduced item variability that naturally occurs during aggregation.

Discussion

The pseudovalues represent the individual item variance. Therefore, theoretically pseudovalues can be used for item analyses. The results of the second

Table 2
A Comparison of HLM and Jackknife Variance Component Estimates in TIMSS

	HLM variance estimate	*Jackknife variance estimate*
Student (Level 1)	0.2629 (84%)	0.2376 (85%)
School (Level 2)	0.0592 (16%)	0.0587 (15%)

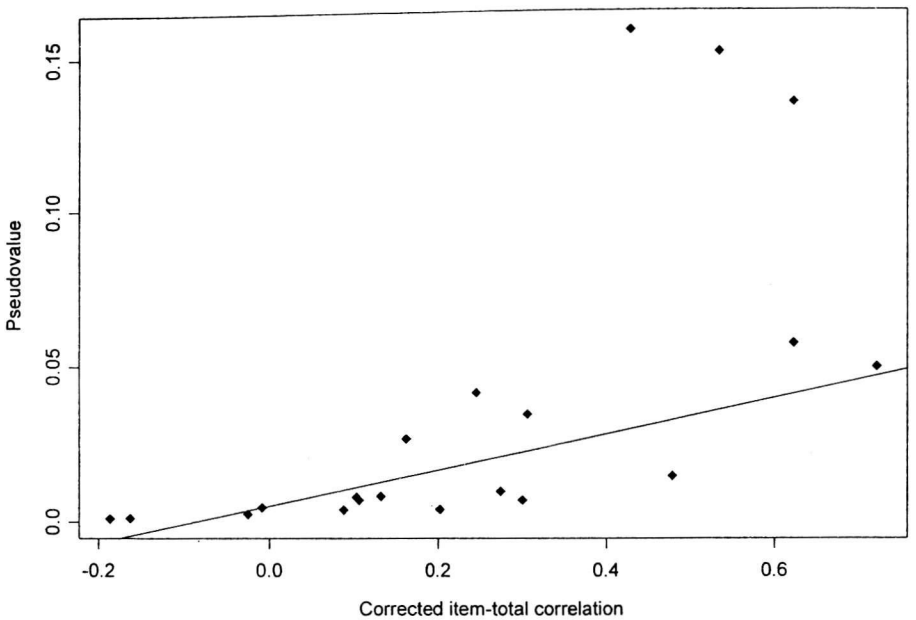

Figure 5. The relationship between item pseudovalues and corrected item-total correlations at Level 2 (school).

study support this notion. If the variance of the scale without one item is smaller than the original variance estimate for the whole scale, then the item that was removed has a large pseudovalue. Items with large pseudovalues have high variance contributions and therefore high covariances with the other items on the scale. According to this principle, Items 1 and 12 appear to be good items at the student level, whereas Items 3 and 9 could be improved. In contrast, Items 16 and 17 appear to be good items at the school level, whereas Items 2 and 19 appear to be having little influence. Item 1 appears to be influencing the variance components on both student and school levels.

Items with high pseudovalues are expected to have a high item-total correlation. In contrast, items with low pseudovalues have low variance contributions, low inter-item covariances, and, therefore, low item-total correlations. If the item pseudovalue is small, then the item has a low amount of variation and does not provide much information. If the pseudovalue is high, then the item is providing a lot of information. Examining the pseudovalues generated when using the jackknife on HLM estimates offers a unique opportunity for multilevel item analysis.

Conclusion

Jackknifing over the items of composite measures seems to offer enhancements and insights to conventional HLM analyses. It appears to reduce the bias in the estimates, both of slopes and of variances. Future research should examine the consistency of these results with larger datasets and different variance components. Specifically, the approach should be tested when a larger proportion of the variance lies at the second level. The studies should also be expanded to include models with more than two levels and repeated measure designs.

Finally, a simulation study aimed at quantifying the amount of bias reduction and verifying the item analysis results should be conducted.

References

Brennan, R.L., Harris, D.J., & Hanson, B.A. (1987). *The bootstrap and other procedures for examining the variability of estimated variance components in testing contexts.* Iowa City, IA: American College of Testing.

Burstein, L. (1993). *The IEA study of mathematics III: Student growth and classroom processes.* Oxford, NY: Pergamon.

Byrk, A.S., & Raudenbush, S.W. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Huitema, B.E., & McKean, J.W. (1994). Reduced bias autocorrelation estimation: Three jackknife methods. *Educational and Psychological Measurement, 54,* 645-665.

Kifer, E., & Wolfe, R.G. (1986, April). *Generalized analysis of the components of variance and covariance in the Second IEA Mathematics Study.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Kifer, E., Wolfe, R.G., & Schmidt, W. (1993). The identification and description of student growth in mathematics achievement. Understanding patterns of student growth. In L. Burstein (Ed.), *The IEA study of mathematics III: Student growth and classroom process in early secondary school* (pp. 107-127). Oxford, NY: Pergamon.

Miller, R.G. (1964). A trustworthy jackknife. *Annals of Mathematical Statistics, 35,* 1594-1605.

Miller, R.G. (1974). The jackknife—A review. *Biometrika, 61,* 1-15.

Mosteller, F., & Tukey, J.W. (1977). *Data analysis and regression: A second course in statistics.* Don Mills, ON: Addison-Wesley.

Pandey, T.N., & Hubert, L. (1975). An empirical comparison of several interval estimation procedures for coefficient alpha. *Psychometrika, 40,* 169-81.

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika, 43,* 353-60.

Raudenbush, S.W, Byrk, A.S., & Congdon, R. (2000). *Hierarchical linear and nonlinear modelling* [Computer software]. Chicago, IL: Scientific Software International.

Rogers, W.T. (1976). Jackknifing disattenuated correlations. *Psychometrika, 41,* 121-33.

Schmidt, W.H., & Wolfe, R. (1983, February). *Estimating variance components.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Wilcox, R.R. (1997). *Introduction to robust estimation and hypothesis testing.* San Diego, CA: Academic Press.

White, A.E. (2000, January). *Result generalizability and detection of discrepant data points: Illustrating the jackknife method.* Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas.