Title

Estimating reliability of school-level scores using multilevel and generalizability theory models

Journal
Asia Pacific Education Review, 10(2)

ISSN
1876-407X

Authors
Jeon, Min-Jeong
Lee, Guemin
Hwang, Jeong-Won
et al.

Publication Date
2009-06-01

DOI
10.1007/s12564-009-9014-3

Peer reviewed

Asia Pacific Educ. Rev. (2009) 10:149–158
DOI 10.1007/s12564-009-9014-3

Estimating reliability of school-level scores using multilevel and generalizability theory models

Min-Jeong Jeon · Guemin Lee · Jeong-Won Hwang · Sang-Jin Kang

Received: 29 April 2007 / Revised: 13 December 2007 / Accepted: 14 March 2008 / Published online: 6 May 2009

Abstract The purpose of this study was to investigate the methods of estimating the reliability of school-level scores using generalizability theory and multilevel models. Two approaches, 'student within schools' and 'students within schools and subject areas,' were conceptualized and implemented in this study. Four methods resulting from the combination of these two approaches with generalizability theory and multilevel models were compared for both balanced and unbalanced data. The generalizability theory and multilevel models for the 'students within schools' approach produced the same variance components and reliability estimates for the balanced data, while failing to do so for the unbalanced data. The different results from the two models can be explained by the fact that they administer different procedures in estimating the variance components used, in turn, to estimate reliability. Among the estimation methods investigated in this study, the generalizability theory model with the 'students nested within schools crossed with subject areas' design produced the lowest reliability estimates. Fully nested designs such as (students:schools) or (subject areas:students:schools) would not have any significant impact on reliability estimates of school-level scores. Both methods provide very similar reliability estimates of school-level scores.

Keywords Reliability · Generalizability theory · Multilevel model

M.-J. Jeon
University of California at Berkeley, Berkeley, USA

G. Lee (✉) · J.-W. Hwang · S.-J. Kang
Yonsei University, Seoul, Korea
e-mail: guemin@yonsei.ac.kr

Introduction

School performance assessment programs have been implemented for the purpose of evaluating and monitoring the quality of school systems in many countries. Various achievement tests have been commonly used as a primary indicator in assessing school performance. In general, achievement test scores of students are aggregated into school-level scores such as school mean scores or PAACs: percentages of students at or above cutscores. Aggregated school-level scores obtained from student scores have been examined in many previous studies of various fields to investigate topics related to school quality and educational policies (Hill and Hurley 1984; Ingelhart 1977, 1985a, 1985b; Rohrschneider 1988; Dalton 1984; Sabatier et al. 1987; Wright et al. 1985; Brennan 2001a, b; Kane a Staiger 2002).

Before school-level scores are used, it is necessary to examine their fitness from the perspectives of reliability and validity. This confirmation is critical to making accurate, substantial inferences based on those scores (Dunbar et al. 1991; Gao et al. 1994; Linn et al. 1991). It is required for researchers who use school-level scores to gather and provide information regarding the quality of those measures.

Unfortunately, many previous studies have reported individual-level reliability estimates such as Cronbach's alpha, even though aggregated school-level scores were used (Jones and Norrander 1996). These studies failed to recognize the fact that the reliability estimates for individual-level scores differed from those for school-level scores. This might lead to the misinterpretation or misuse of scores, resulting from the application of inappropriate levels of score consistency. In addition, many researchers believe that school-level scores are more reliable than individual-level scores. However, this kind of conventional opinion on the reliability of school-level scores does not necessarily hold true. For example, if the number of persons within groups goes to infinite, it is reasonable to assume that error variance for persons is likely to be larger than error variance for groups. However, for small number of persons within groups, this is not necessarily true (Brennan 1995, 2001a, b). More precise investigation of the reliability of school-level scores should be conducted before those scores are used as primary measures.

In this study, several methods of estimating the reliability of school-level scores were conceptualized using multilevel and generalizability theory models. Generalizability theory has been commonly used for this purpose (Brennan 1995; Gao et al. 1994; Jones and Norrander 1996; O'Brien 1991), as it enables investigators to explore reliabilities for various circumstances by fixing or randomizing measurement conditions (Brennan 2001a, b). Though multilevel models have not been frequently used for this purpose, they do offer many advantages in examining the relationships among individual-level and school-level measures (Raudenbush and Bryk 1986; Teddlie and Reynolds 2000). Snijders and Bosker (1999) provided several multilevel model procedures for estimating the reliability of school-level scores, utilizing data involving individuals nested within schools. However, there are relatively few multilevel model studies that address this issue.

The main purposes of this study were to conceptualize possible methods for estimating the reliability of school-level scores, using generalizability theory and multilevel models, and to investigate the relative fitness of the estimates derived from both models. Similarities and differences among those estimates were examined and discussed in relation to the model specifications and estimation procedures. The following were the specific research objectives:

1. To estimate the reliability of school-level scores incorporating a 'students within schools' approach using generalizability theory and multilevel models.
2. To estimate the reliability of school-level scores incorporating a 'students within schools and subject areas' approach using generalizability theory and multilevel models.
3. To evaluate and contemplate the similarities and differences in reliability estimates of school-level scores from the four estimation methods.

School-level score estimation methods

Two different approaches in estimating the reliability of school-level scores were differentiated in this study. In the first approach, students are nested within schools and the students' test scores are averaged into a school-level score (the 'students within schools' approach). In the second approach, students are nested within schools and the students take several tests in various subject areas (the 'students within schools and subject areas' approach). These two different approaches are combined with both generalizability theory and multilevel models, respectively, and constitute several reliability estimation methods as shown in Table 1. In general, the lower reliability estimates of school-level scores could be expected for the "students within schools and subject areas" approach than for the "students within schools" approach in both models, because the former addresses one more source of errors, the 'subject areas' in addition to the 'students', in the generalization of test scores.

Estimation methods using generalizability theory models

A generalizability theory design (p:s) can be used to estimate the reliability of school-level scores, in which students (p) are nested within schools (s),

Table 1 Methods of estimating reliability of school-level scores used in this study

Approach	Generalizability theory	Multilevel model
Students within schools	Model	Model
	$X_{\mathrm{ps}} = \mu + \mu_{\mathrm{s}} \sim + \mu_{\mathrm{p:s}} \sim$	Two-level
		$Y_{\mathrm{ij}} = \mu + U_{\mathrm{j}} + R_{\mathrm{ij}}$
	Reliability	Reliability
	$E\rho^2 = \frac{\sigma^2(\mathrm{s})}{\sigma^2(\mathrm{s})+\sigma^2(\mathrm{p:s})/\mathrm{n}'_{\mathrm{p}}}$	$\lambda_{\mathrm{j}}= \frac{\sigma^2(\mathrm{s})}{\sigma^2(\mathrm{s})+\sigma^2(\mathrm{p:s})/n_{\mathrm{j}}}$
Students within schools and subject areas	Model	Model
	$X_{\mathrm{pst}} = \mu + \mu_{\mathrm{s}} \sim + \mu_{\mathrm{p:s}} \sim + \mu_{\mathrm{st}} \sim + \mu_{\mathrm{t}} \sim + \mu_{\mathrm{pt:s}}$	Three-Level
		$Y_{\mathrm{ijk}} = \mu + U_{\mathrm{k}} + R_{\mathrm{jk}} + e_{\mathrm{ijk}}$
	Reliability	Reliability
	$E\rho^2 = \frac{\sigma^2(\mathrm{s})}{\sigma^2(\mathrm{s})+\left[\sigma^2(\mathrm{st})/\mathrm{n}'_{\mathrm{t}}+\sigma^2(\mathrm{p:s})/\mathrm{n}'_{\mathrm{p}}+\sigma^2(\mathrm{pt:s})/\mathrm{n}'_{\mathrm{p}}\mathrm{n}'_{\mathrm{t}}\right]}$	$\lambda_{\mathrm{k}} = \frac{\sigma^2(\mathrm{s})}{\sigma^2(\mathrm{s})+\sigma^2(\mathrm{p:s})/n_{\mathrm{k}}+\sigma^2(\mathrm{t:p:s})/n_{\mathrm{k}}n_{\mathrm{jk}}}$

$$\begin{aligned} X_{\mathrm{pst}} = {} & \mu\ [\text{grand mean}] + (\mu_{\mathrm{s}} - \mu)\ [\text{school effect}] \\ & + \left(\mu_{\mathrm{ps}} - \mu_{\mathrm{s}}\right)\ [\text{student within school effect}] \end{aligned} \tag{1}$$

where the last term of students within schools effects is compounded by unexplained sources of error (O'Brien 1991). Suppose that schools are the objects of measurement, in this case, the universe of generalization consist of a random p facet. For this design, the reliability of school-level scores (called the generalizability coefficient) is

$$E\rho^2 = \frac{\sigma^2(\mathrm{s})}{\sigma^2(\mathrm{s}) + \sigma^2(\mathrm{p{:}s})/n'_{\mathrm{p}}} \tag{2}$$

where $\sigma^2(\mathrm{s})$ is the variance of schools, $\sigma^2(\mathrm{p{:}s})$ is the variance of students within schools, and n'_{p} is the number of students within a school.

The linear model for data including one additional subject areas facet (t), in which students within schools are crossed with subject areas, is expressed as

$$\begin{aligned} X_{\mathrm{pst}} = {} & \mu\ [\text{grand mean}] \\ & + (\mu_{\mathrm{s}} - \mu)\ [\text{school effect}] \\ & + (\mu_{\mathrm{t}} - \mu)\ [\text{subject area effect}] \\ & + \left(\mu_{\mathrm{ps}} - \mu_{\mathrm{s}}\right)\ [\text{student within school effect}] \\ & + \left(\mu_{\mathrm{st}} - \mu_{\mathrm{s}} - \mu_{\mathrm{t}} + \mu\right)\ [\text{school by subject area} \\ & \qquad \textit{interactioneffect}] \\ & + \left(X_{\mathrm{pst}} - \mu_{\mathrm{st}} - \mu_{\mathrm{ps}} + \mu_{\mathrm{s}}\right)\ [\text{student by subject} \\ & \qquad \textit{areawithinschooleffect}] \end{aligned} \tag{3}$$

where the last term of students by subject areas within schools effects is compounded by unexplained sources of error (Brennan 1995).

Schools are the objects of measurement, and the universe of generalization consists of p and t facets. The reliability of school-level scores, then, is

$$E\rho^2 = \frac{\sigma^2(\mathrm{s})}{\sigma^2(\mathrm{s}) + \sigma^2(\mathrm{st})/n'_{\mathrm{t}} + \sigma^2(\mathrm{p{:}s})/n'_{\mathrm{p}} + \sigma^2(\mathrm{pt{:}s})/n'_{\mathrm{p}}n'_{\mathrm{t}}}, \tag{4}$$

where $\sigma^2(\mathrm{s})$ is the variance of schools, $\sigma^2(\mathrm{st})$ is the variance of schools by subject areas interaction, $\sigma^2(\mathrm{p{:}s})$ is the variance of students within schools, $\sigma^2(\mathrm{pt{:}s})$ is the variance of students by subject areas within schools, n'_{p} is the number of students within a school, and n'_{t} is the number of subject areas.

If subject areas were treated as fixed, a different conceptualization of the universe of generalization and a different formula to estimate the reliability of school-level scores should be considered. That is, in this case, only limited subject areas (e.g., language, mathematics, and English) are used in computing school-level scores, and the researcher is not interested in generalization over other subject areas in assessing school-level performances. The reliability of school-level scores, where subject areas are treated as fixed, is

$$E\rho^2 = \frac{\sigma^2(\mathrm{s}) + \sigma^2(\mathrm{st})}{\sigma^2(\mathrm{s}) + \sigma^2(\mathrm{st}) + \sigma^2(\mathrm{p{:}s})/n'_{\mathrm{p}} + \sigma^2(\mathrm{pt{:}s})/n'_{\mathrm{p}}}, \tag{5}$$

where the definition of each term is the same as in Eq. (4).

Estimation methods using multilevel models

For the data structure for students (p) nested within schools (s), a two-level multilevel model is appropriate and is expressed as

$$Y_{\mathrm{ij}} = \mu + U_{\mathrm{j}} + R_{\mathrm{ij}}, \tag{6}$$

where μ is the population grand mean, U_{j} is the specific effect of school j, which is to say the deviation of school j's mean from the grand mean, andR_{ij} is the residual effect for student i within school j.

The two-level model partitions the total variability of an observed score into between-school variance and within-school variance. Applying the general definition of reliability, the two-level model provides the reliability of the aggregate scores

$$\lambda_{\mathrm{j}} = \frac{\sigma^2(\mathrm{s})}{\sigma^2(\mathrm{s}) + \sigma^2(\mathrm{p{:}s})/n_{\mathrm{j}}}, \tag{7}$$

where $\sigma^2(\mathrm{s})$ is the variance between schools, $\sigma^2(\mathrm{p{:}s})$ is the variance of students within schools, and n_{j} is the student sample size of school j (Snijders and Bosker 1999).

In the second approach, students within schools take several tests in certain subject areas. Unlike generalizability theory, multilevel models view subject areas nested within students within schools treating different subject areas as multiple data points where students' test scores are observed. In this case, the linear model of observed scores with the three-level multilevel model is appropriate, and is expressed as

$$Y_{\mathrm{ijk}} = \mu + U_{\mathrm{k}} + R_{\mathrm{jk}} + e_{\mathrm{ijk}}, \tag{8}$$

where μ is the grand mean of the population, U_{k} is the school effects, which is to say the deviation of school k's mean from the grand mean, R_{jk} is the student effects, or the deviation of student jk's mean from the school mean, and e_{ijk} is the residual effect for subject area i within student j within school k.

The reliability of the school-level scores in this model can be expressed as

$$\lambda_{\mathrm{k}} = \frac{\sigma^2(\mathrm{s})}{\sigma^2(\mathrm{s}) + \sigma^2(\mathrm{p{:}s})/n_{\mathrm{k}} + \sigma^2(\mathrm{t{:}p{:}s})/n_{\mathrm{k}}n_{\mathrm{jk}}}, \tag{9}$$

where $\sigma^2(\mathrm{s})$ is the variance of schools, $\sigma^2(\mathrm{p{:}s})$ is the variance of students within schools, $\sigma^2(\mathrm{t{:}p{:}s})$ is the variance of subject areas among students, n_k is the number of students in school k, and n_jk is the number of subject areas for student j in school k.

Methods

Data sources

The data used in this study were taken from the Korean Education and Employment Panel (KEEP) administered to grade three high school students in 2002. The data was obtained from a representative sample by applying nationwide survey procedures. In this study, the data of 1,477 students in 90 high schools were used for the final analyses. In addition, three subject areas, Korean Language, Mathematics, and English, were used to measure the students' achievement on the Korean College Scholastic Ability Tests (similar to the SAT or ACT tests in the United States). Since the data structure was unbalanced, with varying numbers of students across schools, for the purpose of comparing results from both balanced and unbalanced data sets, balanced data were created in which the number of students within schools was set to 10. The same estimation procedures were applied to both the balanced and the unbalanced data. Summary statistics describing the balanced and unbalanced data used in this study are presented in Table 2.

Analyses

To estimate the reliabilities of school-level scores, (p:s) and (p:s) × t univariate generalizability theory designs (p, students; s, schools; t, subject areas) and the two- and three-level multilevel models were employed. The computer application program HLM 6.0 (Raudenbush et al. 2005) was used with the multilevel models; GENOVA (Brennan 2001a) for balanced data and urGENOVA (Brennan 2001b) for unbalanced data were used with the generalizability theory models. The variance components of the score effects were estimated, and reliability estimates were obtained and compared for each method. Methods based on mean-squares are applied to the GENOVA and urGENOVA programs in estimating variance components for the generalizability theory models (Lee 2002; Lee and Frisbie 1999). We used default estimation methods of HLM 6.0 in this study: restricted maximum likelihood method (REML) for two-level models and full information maximum likelihood method (FIML) for three-level models.

Results

'Students within schools' approach

Estimation of variance components

Table 3 presents the variance component estimates using the generalizability theory models (G-Model) and multilevel models (M-Model) for the 'students within schools' approach, with balanced and unbalanced data. In both models, students' average scores for subject areas were used as inputs. The school effects (s) and the students within schools effects (p:s) were considered to constitute the total variability of the observed scores.

The result shows that the G-Model and the M-model produced exactly the same variance component estimates for the school effects (s) as for the students within schools effects (p:s) for the balanced data. The percentages of variance component estimates for schools and students within schools were 22.7% and 77.3%, respectively.

For the unbalanced data where the numbers of students per school varied across schools, and ranged from 10 to 20, the variance component estimates for students within schools effects (p:s) in both models were similar, although not

Table 2 Descriptive statistics for balanced and unbalanced data

	Balanced data			Unbalanced data		
	Mean	SD	Range	Mean	SD	Range
Korean language	101.50	8.81	65.00–123.00	100.80	8.43	65.00–121.80
Mathematics	97.20	8.21	78.80–119.40	97.76	7.50	78.80–121.33
English	100.03	9.70	71.40–123.30	99.54	9.34	71.40–122.73
Average score	99.58	8.30	73.30–121.90	99.37	7.89	73.30–121.96
Number of students per school	10	–	–	16.41	2.72	10–20

Note: *SD* standard deviation

Table 3 Variance component estimates of students within schools approach

Data	Effect	G-Model		M-Model	
		Variance component	*df*	Variance component	*df*
Balanced data	School (s)	51.396 (22.7%)	89	51.396 (22.7%)	89
	Student:school (p:s)	174.972 (77.3%)	810	174.973 (77.3%)	810
	Total	226.368 (100.0%)		226.368 (100.0%)	
Unbalanced data	School (s)	43.897 (20.1%)	89	47.967 (21.6%)	89
	Student:school (p:s)	174.214 (79.9%)	1387	174.517 (78.4%)	1387
	Total	218.111 (100.0%)		222.484 (100.0%)	

Notes: The numbers in parentheses represent the percentage of each score effects relative to the total variance

G-model generalizability theory model, *M-model* multilevel model

identical. The school variance component in the M-Model was somewhat greater than that in the G-Model.

Estimation of reliability

Table 4 shows the reliability estimates of school-level scores for the four methods under the G-Model and the M-Model, where the student sample sizes in a school varied from 10 to 100 in increments of 10.

The reliability of school-level scores increased as the student sample size per school increased, though the degree of increase gradually diminished. For the balanced data, the reliability estimates for Method (A) and Method (B) were the same. For the unbalanced data, the reliability estimates of Method (D) were somewhat higher than those of Method (C). The difference ranged from 0.003 to 0.02. The reliability estimates for the unbalanced data were somewhat lower than those for the balanced data. Using the four methods, at least 20 students within a school were required in order to obtain a reliability level of 0.8, whereas at least 40 students were required for a reliability level of 0.9.

‘Students within schools and subject areas’ approach

Estimation of variance components

Table 5 provides the variance component estimates for the G-Model and the M-Model, incorporating students within schools and subject areas for balanced and unbalanced data. In this case, the two models applied different designs and decomposed the total score variance into different sources of score effects. That is, the G-Model considered five variance components including school effects (s), students within schools effects (p:s), subject area effects (t), schools by subject area interaction effects (st), and students by subject area interaction effects within schools (pt:s), whereas the M-Model considered three variance components including school effects (s), students within schools effects (p:s), and subject areas among students within schools effects (t:p:s).

For the balanced data, the variance component estimates for school effects (s) and students within schools effects (p:s) were similar in the two models. The percentages of schools and students within schools variance components

Table 4 Reliability estimates of school-level scores using four estimation methods of ‘students within schools’ approach

Number of students per school	Balanced data		Unbalanced data	
	Method (A) (G-Model)	Method (B) (M-Model)	Method (C) (G-Model)	Method (D) (M-Model)
10	0.746	0.746	0.716	0.733
20	0.855	0.855	0.834	0.846
30	0.898	0.898	0.883	0.892
40	0.922	0.922	0.910	0.917
50	0.936	0.936	0.926	0.932
60	0.946	0.946	0.938	0.943
70	0.954	0.954	0.946	0.951
80	0.959	0.959	0.953	0.957
90	0.964	0.964	0.958	0.961
100	0.967	0.967	0.962	0.965

Notes: *G-Model* generalizability theory model, *M-Model* multilevel model

Table 5 Variance component estimates in 'students within schools and subject areas' approach

Data	Effect	G-Model		M-Model	
		Variance component	*df*	Variance component	*df*
Balanced data	s	49.679 (16.1%)	89	50.631 (16.5%)	89
	p:s	137.992 (44.8%)	810	134.743 (44.0%)	810
	t:p:s	–	–	120.690 (39.4%)	1800
	t	4.595 (1.5%)	2	–	–
	st	5.152 (1.7%)	178	–	–
	pt:s	110.943 (40.0%)	1620	–	–
	Total	308.362 (100.0%)		306.064 (100.0%)	
Unbalanced data	s	42.005 (13.9%)	89	47.280 (15.5%)	89
	p:s	135.525 (44.9%)	1387	133.101 (43.7%)	1387
	t:p:s	–	–	124.263 (40.8%)	2954
	t	2.521 (0.8%)	2	–	–
	st	5.676 (1.9%)	178	–	–
	pt:s	116.066 (38.5%)	2774	–	–
	Total	301.792 (100.0%)		304.644 (100.0%)	

Notes: The numbers in parentheses represent each score's effects as a percentage relative to the total variance

G-Model generalizability theory model, *M-Model* multilevel model, *s* school, *p* students, *t* subjects

were about 16% and 44%, respectively. In addition, in the G-Model, the sum of the variance components of subject area effects (t), schools by subject area interaction effects (st) and subject area by students within schools effects (pt:s) was 120.690 which was exactly the same value as the variance component estimate in the M-Model for subjects within students within schools effects (t:p:s).

For the unbalanced data, the variance component estimates for school effects (s) and students within schools effects (p:s) were somewhat different between the G-Model and M-Model. However, the sum of the variance components of subject area effects (t), schools by subject area interaction effects (st) and subject area by students within schools effects (pt:s) was 124.263 in the G-Model, which was the same as the variance estimate of subjects within students within schools effects (t:p:s) in the M-Model.

Estimation of reliability

Table 6 presents the reliability estimates of school-level scores based on school sample sizes ranging from 10 to 100 in increments of 10.

The reliability of school-level scores in the four methods gradually increased as the student sample size per school increased. For the balanced data, the reliability estimates of the school scores of Method (A) were consistently lower than those of Method (B). The difference between the two methods was about 0.03. For the unbalanced data, the reliability estimates of Method (C) were also consistently lower than those of Method (D). The difference between the two methods was about 0.05 which was greater than that for the balanced data. In the case of the 'students within schools' approach, the reliability estimates for the unbalanced data were lower than those for the balanced data. The reliability estimates of Method (C) for the unbalanced data were the lowest, whereas those of Method (B) for the balanced data were the highest among the methods.

Comparison among specified methods

One of the research objectives of this study was to investigate the similarities and differences between several G- and M-Model methods in estimating the reliabilities of school-level scores. To that end, the variance component estimates of several methods, according to different specifications, are presented in Tables 7 and 8. The reliability estimates of the school-level scores are also presented. To enhance the utility of the comparison of the methods, one additional method, that of the G-Model (t:p:s) design, was analyzed. The variance component estimates of this method were obtained by analyzing the balanced and unbalanced data that were used for the three-level multilevel model of (t:p:s) design.

The variance component estimates and the related reliability estimates of the six methods, for the balanced data, are presented in Table 7. For the purpose of estimating the reliability, a student sample size of 10 was used in all of the methods.

Method (A), Method (B), and Method (E) produced the same reliability estimate, 0.746, which was the highest value among the proposed methods. They also produced the same variance component estimate for school effects (s). The reliability estimate of Method (D) was similar to

Table 6 Reliability estimates of school-level scores using four estimation methods of 'students within schools and subject areas' approach

Number of students per school	Balanced data		Unbalanced data	
	Method (A) (G-Model)	Method (B) (M-Model)	Method (C) (G-Model)	Method (D) (M-Model)
10	0.721	0.743	0.685	0.730
20	0.826	0.853	0.799	0.844
30	0.868	0.897	0.845	0.890
40	0.891	0.920	0.871	0.916
50	0.905	0.935	0.887	0.931
60	0.915	0.946	0.898	0.942
70	0.922	0.953	0.906	0.950
80	0.927	0.959	0.912	0.956
90	0.931	0.963	0.917	0.961
100	0.935	0.967	0.920	0.964

Notes: *G-Model* generalizability theory model, *M-Model* multilevel model

Table 7 Variance components and related reliability estimates of five methods with balanced data

Effect	Students within schools approach		Students within schools and subject areas approach		
	Method (A) G-Model	Method (B) M-Model	Method (C) G-Model (p:s) × t	Method (D) M-Model (t:p:s)	Method (E) G-Model (t:p:s)
s	51.396 (22.7%)	51.396 (22.7%)	49.679 (16.1%)	50.631 (16.5%)	51.396 (16.8%)
p:s	174.972 (77.3%)	174.972 (77.3%)	137.992 (44.8%)	134.743 (44.0%)	134.743 (43.9%)
t:p:s	–	–	–	120.690 (39.4%)	120.690 (39.3%)
t	–	–	4.595 (1.5%)	–	–
st	–	–	5.152 (1.7%)	–	–
pt:s	–	–	110.943 (36.0%)	–	–
Total	226.368 (100.0%)	226.368 (100.0%)	308.362 (100.0%)	306.064 (100.0%)	306.830 (100.0%)
Reliability estimates (10)	0.746	0.746	0.721	0.743	0.746

Notes: The numbers in parentheses represent each score's effects as a percentage relative to the total score variance. Reliability estimates (10) are reliability estimates when the student sample size within a school is 10

G-Model generalizability theory model, *M-Model* multilevel model, *s* school, *p* students, *t* subjects

Table 8 Variance components and related reliability estimates of five methods with unbalanced data

Effect	Students within schools approach		Students within schools and subject areas approach		
	Method (A) G-Model	Method (B) M-Model	Method (C) G-Model (p:s) × t	Method (D) M-Model (t:p:s)	Method (E) G-Model (t:p:s)
s	43.897 (20.1%)	47.967 (21.6%)	42.005 (13.9%)	47.280 (15.5%)	43.904 (14.6%)
p:s	174.214 (79.9%)	174.517 (78.4%)	135.525 (44.9%)	133.101 (43.7%)	132.611 (44.1%)
t:p:s	–	–	–	124.263 (40.8%)	124.451 (41.4%)
t	–	–	2.521 (0.8%)	–	–
st	–	–	5.676 (1.9%)	–	–
pt:s	–	–	116.066 (38.5%)	–	–
Total	218.111 (100.0%)	222.484 (100.0%)	301.792 (100.0%)	304.644 (100.0%)	300.966 (100.0%)
Reliability estimates (10)	0.716	0.733	0.685	0.730	0.716

Notes: The numbers in parentheses represent each score's effects as a percentage relative to the total score variance. Reliability estimates (10) are reliability estimates when the student sample size within a school is 10

G-Model generalizability theory model, *M-Model* multilevel model, *s* school, *p* students, *t* subjects that of Method (A), Method (B), and Method (E), while that of Method (C) was the lowest at 0.721.

The total variance of the observed scores in the 'students within schools and subject areas' approach was greater than that in the 'students within schools' approach. It was evident that additional consideration of subject area effects could lead to a considerable increment of the total score variance, since the subject area variance was not considered in the 'students within schools' approaches, in which the students' average scores were used as inputs instead of the individual test scores for the several subject areas.

It is meaningful to note that the variance component estimates in Method (D) and Method (E), under different estimation procedures, were very similar. Given the fact that Method (A) and Method (B) also produced the same variance component estimates, it is reasonable to expect that for balanced data, using either the G-Model or the M-Model with the same design could lead to the same or very similar variance component estimates and, consequently, to the same or very similar reliability estimates of school-level scores.

The variance component estimate for the (t:p:s) effects in the (t:p:s) designs was the same value as the sum of the t, (st), and (pt:s) effects in the (p:s) × t design. That is, for the balanced data, the variance component for the (t:p:s) effects could be decomposed into the three variance components for the t, (st), and (pt:s) effects. In addition, the methods of the fully nested (p:s) and (t:p:s) designs produced very similar reliability estimates, although the (p:s) × t mixed design produced lower reliability estimates than the other methods.

Table 8 presents the variance component estimates and related reliability estimates of the six methods for the unbalanced data. Even though the actual number of students per school varied across schools for the unbalanced data, the number of students in a school was set to 10 in order to estimate the reliability of school-level scores. The number of students in a school was fixed at 10 in order to yield results comparable to those from the analysis for the balanced data.

Method (A) and Method (E) produced the same reliability estimate of 0.716, and Method (B) and Method (D) produced similar estimates. The reliability estimate of Method (B) was the highest among the presented methods. Method (A) and Method (B), as well as Method (D) and Method (E), had the same designs under different models but produced different variance component estimates and reliability estimates, owing, as previously indicated, to the different estimation procedures used in the G-Model and M-Model for the unbalanced data. As was the case for the balanced data, the total variance of observed scores was larger in the 'students within schools and subject areas' approach than in the 'students within schools' approach.

As was the case for the balanced data, for the unbalanced data the variance component estimate for the (t:p:s) effects in the (t:p:s) design M-Model was decomposed into three variance components for the t, (st), and (pt:s) effects in the (p:t) x s design. In addition, the fully nested (p:s) and (t:p:s) designs produced the same or similar reliability estimates, and the (p:s) × design produced the lowest reliability estimate.

Discussion

This study was designed to address issues related to the estimation of the reliability of school-level scores. Two approaches were conceptualized according to generalizability theory and multilevel models, a 'students within schools' approach and a 'students within schools and subject areas' approach. Several methods, being combinations of the approaches and measurement models, were applied to both the balanced and unbalanced data.

In the 'students within schools' approach for balanced data, the G-Model and the M-model produced exactly the same variance components and reliability estimates. The linear equations of the score effects for the G- and M-Models were mathematically the same, and the reliability estimation procedures in both models seemed comparable. These results suggest that the different estimation procedures employed by the G-Model and M-Model (EMS in the G-Model and REML in the M-Model, respectively) made no difference in estimating the variance components for the balanced data. Consequently, for the 'students within schools' approach with balanced data, it does not matter to use either the G-Model or the M-Model in estimating reliability of school-level scores.

However, for the unbalanced data in the 'students within schools' approach, the M-Model and the G-Model produced somewhat different variance component and reliability estimates. As Brennan (2001a, b) and Searle et al. (2006) indicated, using the G-Model while implementing analogous-ANOVA procedures for unbalanced data could lead to different estimates from those yielded by the REML in the M-Model. In turn, different estimation procedures implemented by the two models can lead to different variance components and reliability estimates for school-level scores. We found slightly larger variance component estimates with the M-Model (from HLM) than with the G-Model (from urGENOVA).

There could be several explanations about discrepancy among variance component estimates from two models. For example, the HLM uses EM approach where complete sufficient statistics are estimated in each of the iteration of estimation. The iteration of estimation procedures might influence on the variance component estimates. In another perspective, the estimation procedures of HLM and urGENOVA are so complicated and the differences among variance component estimates might come from accumulated rounding errors in the long computation processes. It is not clear to the authors, however, what causes such differences among variance component estimates at this point. This question cannot be answered within the scope of this study and can be more thoroughly investigated by additional simulation studies.

The 'students within schools and subject areas' approach led to different reliability estimates of school-level scores in both the G-Model and the M-Model. Treating 'subject areas' as a nested facet does not seem to have significant impact on the reliability estimates of school-level scores. That is, incorporating 'subject areas' as a nested facet under a fully nested design such as (subject areas:students:schools) would not have any significant influences on reliability estimates of school-level scores.

However, treating 'subject areas' as a crossed facet leads to lower reliability estimates due to the consideration of additional sources of errors. If considering 'subject areas' as an important source of variation of school-level scores, it would be recommended to involve this facet in the models of estimating reliability of school-level scores. The G-Model would be appropriate for this purpose, because it can incorporate any facets as crossed or nested factors with great flexibility under fixed, random, and/or mixed effects models (Brennan 2001a, b; Hox and Maas 2006). For example, the G-Model can easily specify a [(students:schools) × subject areas] design that treats subject areas as a crossed facet.

If subject areas are crossed with students within schools, variance components for multilevel models cannot be estimated by any current commercial software that can handle just fully nested designs. However, there are several solutions to this limitation of multilevel models. Hox and Maas (2006) explained the method of implementing the lowest level to estimate the residual variance by using fixed "dummy" levels. Variance components for the two-way and *n*-way crossed designs can be also estimated under random effects models (Kang 1992; Kang et al. 2004; Raudenbush 1993; Rasbash and Goldstein 1994). However, if data sets with a large number of crossed facets, the current multilevel software such as HLM 6.0 does not handle this well.

References

Brennan, R. L. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement, 32*, 385–396. doi:10.1111/j.1745-3984.1995.tb00473.x.

Brennan, R. L. (2001a). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L. (2001b). *urGENOVA 2.1*. Iowa City, IA: The University of Iowa.

Dalton, R. J. (1984). Cognitive mobilization and partisan dealignment in advance industrial democracies. *The Journal of Politics, 46*, 264–284. doi:10.2307/2130444.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289–303. doi:10.1207/s15324818ame0404_3.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science. *Applied Measurement in Education, 7*, 323–342. doi:10.1207/s15324818ame0704_4.

Hill, K. Q., & Hurley, P. A. (1984). Estimating congressional district attributes with national election study data: A reliability assessment. *Political Methodology, 10*, 447–463.

Hox, J., & Maas, C. (2006). Multilevel models for multimethod measurements. In M. Eid & E. Diener (Eds.), *Multimethod measurement in psychology* (pp. 269–281). Washington, DC: American Psychological Association.

Ingelhart, R. (1977). *The silent revolution: Changing values and political styles among western publics*. Princeton, NJ: Princeton University Press.

Ingelhart, R. (1985a). Aggregate stability and individual-level flux in mass belief systems: the level of analysis paradox. *The American Political Science Review, 79*, 97–116. doi:10.2307/1956121.

Ingelhart, R. (1985b). New perspectives on value change: responses to Lafferty and Knotsen, Savage, and Boltken and Jagodzinski. *Comparative Political Studies, 17*, 485–532. doi:10.1177/0010414085017004004.

Jones, B. S., & Norrander, B. (1996). The reliability of aggregated public opinion measures. *American Journal of Political Science, 40*(1), 295–309. doi:10.2307/2111703.

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *The Journal of Economic Perspectives, 16*(4), 91–114. doi:10.1257/089533002320950993.

Kang, S.J. (1992). *A mixed linear model with two-way crossed random effects and estimation via the EM algorithm.* Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Kang, S. J., Rowan, B. P., & Raudenbush, S. W. (2004). Estimating the effects of academic departments on organic design in high schools: A crossed-multilevel analysis. In W. Hoy & C. Miskel (Eds.), *Educational administration, policy, and reform: Research and measurement* (pp. 123–152). Greenwich, CN: Information Age Publishing.

Lee, G. (2002). The influence of several factors on reliability for complex reading comprehension tests. *Journal of Educational Measurement, 39*, 149–164.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under ageneralizability theory models for test scores composed of testlests. *Applied Measurement in Education, 12*, 237–255. doi:10.1207/S15324818AME1203_2.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21.

O'Brien, R. M. (1991). Correcting measures of relationship between aggregate-level variables. *Sociological Methodology, 21*, 125–165. doi:10.2307/270934.

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational Statistics, 19*(4), 337–350.

Raudenbush, W. S. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics, 18*(4), 321–350. doi:10.2307/1165158.

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59*, 1–17. doi: 10.2307/2112482.

Raudenbush, W. S., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2005). *HLM: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International, Inc.

Rohrschneider, R. (1988). Citizens' attitudes toward environmental issues: Selfish of selfless? *Comparative Political Studies, 21*, 347–367. doi:10.1177/0010414088021003002.

Sabatier, P., Hunter, S., & Mclaughlin, S. (1987). The devil shift: Perceptions and misperceptions of opponents. *The Western Political Quarterly, 50*, 449–476. doi:10.2307/448385.

Searle, R. S., Casella, G., & McCulloch, C. E. (2006). *Variance components*. Hoboken, NJ: Wiley.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis*. London: Sage Publications.

Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research*. New York: Falmer Press.

Wright, G. C., Erikson, R. S., & McIver, J. P. (1985). Measuring state partisanship and ideology with survey data. *The Journal of Politics, 47*, 469–489. doi:10.2307/2130892.