Journal of Educational Measurement
Winter 1991, Vol. 28, No. 4, pp. 338–354

Multilevel Factor Analysis of Class and Student Achievement Components

Bengt O. Muthén
University of California, Los Angeles

This article analyzes mathematics achievement data from the Second International Mathematics Study (SIMS; Crosswhite, Dossey, Swafford, McKnight, & Cooney, 1985) in which U.S. students are measured at the beginning and end of eighth grade. The aim of the article is to address some substantive analysis questions in the SIMS data and show the potential of multilevel factor analysis methodology. Issues related to between- and within-class decomposition of achievement variance and the change of this decomposition over the course of the eighth grade are studied. As a starting point, random effects ANOVA is considered for each achievement score. Each score contains a large amount of measurement error. The effects of unreliability on variance decomposition are shown with the help of a multilevel factor analysis model. Unreliability has severely distorting effects on this type of ANOVA while multilevel factor analysis gives results corresponding to what would be obtained with perfectly reliable scores.

Educational research often depends on multivariate statistical methods like confirmatory factor analysis and other covariance structure techniques (e.g., see Bollen, 1989; Joreskog & Sorbom, 1979) to study underlying dimensions of systematic variation in data collected on students and to assess reliability and invariance of measurement instruments (e.g., see Bohrnstedt, 1983). Standard analysis methods make the simplifying assumption that the data have been obtained as a simple random sample from a given population. This involves an assumption of independently and identically distributed observations (IID). Much educational data is, however, obtained through a complex, multistage sample design involving clustered observations where the IID assumption is unrealistic. Typical examples are large-scale surveys like the National Longitudinal Study (NLS) of the high-school graduates of 1972 and the National Longitudinal Study of 1988 (NELS: 88; National Center for Education Statistics [NCES], 1989), both of which employ stratified sampling of schools along with random sampling of students within schools. This article considers another large-scale survey, the Second International Mathematics Study (SIMS; Crosswhite et al., 1985), which has a similar nested or hierarchical data structure of students observed within classes, classes within schools, and

This research was supported by Grant SES-8821668 from the National Science Foundation and by Grant OERI-G-86-003 from the Office of Educational Research and Improvement, Department of Education. I would like to thank Ginger Nelson, Jin-Wen Yang Hsu, Kathleen Wisnicki, and Tammy Tam for valuable research assistance and my fall, 1990, research seminar group for stimulating discussions. Linda Muthén, David Rindskopf, and Albert Satorra have also provided helpful comments.

#37

schools within school districts. Standard analysis methods are adversely affected by such deviation from simple random sampling (e.g., see Skinner, Holt, & Smith, 1989). For achievement tests in schools, the violation of the assumption of independent observations may be particularly important because students of the same class are likely to produce sizable intraclass correlations due to strong common sources of variation. In the last few years, appropriate techniques have begun to be employed for univariate response models such as multiple regression through the use of random coefficient or multilevel regression models (e.g., see Bock, 1989; Raudenbush & Bryk, 1988). Extensions of techniques for multivariate response models are, however, just emerging. Some new developments of this type will be demonstrated in this article.

From the previously described sampling perspective, design features are viewed as complications to the statistical analysis. On the other hand, the complex sample design of educational studies can be viewed as an opportunity for more informative modeling of substantive phenomena. There is often an explicit interest in relating the variation in the data to the multiple stages of the sampling—such as school, class, and student—and data is often gathered on such multiple levels with an interest in studying the interaction among these levels. For example, in NELS: 88 (NCES, 1989) there is not only an interest in student-level data but also data obtained for these students' teachers, school principals, and parents. SIMS (Crosswhite et al., 1985) has information on 8th- and 12th-grade mathematics achievement, where effects of differing amounts of "tracking" in different educational systems in different countries and provinces can be studied. In SIMS, it is, therefore, of interest to separate within-class and between-class variation of student achievement, to relate between-class achievement variation to class-level information on teacher and teaching characteristics, and to contrast different educational systems (e.g., see Burstein, in press).

Whereas the statistical concerns about multivariate modeling in complex samples and the emerging solutions are rather new, the substantive concerns about "multilevel" modeling are relatively old, including issues related to the proper unit of analysis, aggregation effects, and contextual effects. For overviews in educational and sociological contexts, see Cronbach (1976) and Burstein (1980). The Cronbach reference is particularly relevant to this article because he discusses issues of factor analysis on multiple levels. Cronbach reanalyzes Bond-Dykstra data (Bond & Dykstra, 1967) from the Cooperative Reading Study using separate factoring of within-class and between-class covariance matrices for ability measures. Harnqvist (1978) uses a similar approach to factor analyze mental ability scores of students observed within classrooms. These analyses point to different structures for the between, within, and overall covariance matrices. However, these ideas do not appear to have had a large impact on factor analysis practice when it comes to hierarchical data such as students within classes. One reason is, perhaps, that the statistical methodology and software development has lagged behind. Appropriate theory, however, was developed already in Schmidt (1969).

Relevant statistical methodology is now becoming generally available for efficient multivariate analyses of the kind that Cronbach and others envisioned (for an overview, see Muthén, 1989). The aim of this article is to address some substantive analysis questions in the SIMS data (Crosswhite et al., 1985) and let these analyses indicate the considerable potential of this new methodology. A set of mathematics achievement test scores for U.S. eighth graders will be studied. These students are to some extent selected into different types of eighth-grade math classes based on previous performance. Typically, arithmetic content is well covered, but there are major differences in how much algebra and geometry is taught. In this way, the classes can be characterized in broad categories like remedial, typical, algebra, and enriched, although differences in emphases across classes remain even within these categories. This article studies issues related to between- and within-class decomposition of achievement variance and the change of this decomposition over the course of the eighth grade. In the univariate response case, random effects ANOVA is used for such variance decomposition. For applications of variance decomposition to educational test scores, see, for example, Wiley and Bock (1967) and Rakow, Airasian, and Madaus (1978). For related SIMS analyses, see Schmidt, Wolfe, and Kifer (in press).

As a starting point, random effects ANOVA will be considered for each achievement score. Each score contains a large amount of measurement error, however, and an important point of the article is to study the effects of unreliability on variance decomposition with the help of a multilevel factor analysis model. It will be shown that unreliability has severely distorting effects on this type of ANOVA and that the multilevel factor analysis gives results corresponding to what would be obtained with perfectly reliable scores. In the section entitled, "The Data and the Substantive Research Questions," the SIMS data is described, and the general substantive research questions are formulated. This gives a background for the conventional ANOVA analyses in the section, "Conventional Analysis: Random Effects ANOVA," and the multilevel factor analysis methods development in the section, "Multilevel Factor Analysis." The section entitled, "Multilevel Analysis," gives the MFA results and compares them to ANOVA. The section, "Unreliability Sensitivity Analyses," investigates the MFA-ANOVA comparison under varying degrees of unreliability. The final section concludes.

The Data and the Substantive Research Questions

In the Second International Mathematics Study (Crosswhite et al., 1985), a national probability sample of school districts was selected proportional to size; a probability sample of schools was selected proportional to size within school district, and two classes were randomly drawn within each school. I will consider a subset of the U.S. eighth-grade data of 3,724 students who took the core test at both the pretest in the fall of 1982 and the posttest in the spring of 1983. These students were observed in 197 classes from 113 schools. The class sizes vary from 2 to 38 with a typical value of around 20.

The core test consisted of 40 items in the areas of arithmetic, algebra,

geometry, and measurement. The topics covered in these items were broken down into eight subscores where each subscore was the sum of binary items (an item classification is given in the appendix of Muthén, Kao, & Burstein, 1991). One item in the core had a very low item-test correlation and was excluded. The subscore RPP consisted of eight ratio, proportion, and percent items. FRACT consisted of eight common and decimal fraction items. EQEXP consisted of six algebra items involving equalities and expression. INTNUM consisted of two items involving integer number algebra manipulations. STESTI consisted of five items dealing with measurement items involving standard units and estimation. AREAVOL consisted of two measurement items dealing with area and volume determination. COORVIS consisted of three geometry items involving coordinates and spatial visualization. PFIGURE consisted of five geometry items involving properties of plane figures.

Although the subscores consisted of relatively few items and thus might have been rather unreliable, it was of interest to be able to study these variables separately because to some extent they corresponded to different emphases in eighth-grade mathematics curricula. At the end of this article, however, I compare the analysis of these eight variables with the analysis of more reliable subscores, using the four aggregated variables ARITHMETIC (RPP and FRACT), ALGEBRA (EQEXP and INTNUM), MEASUREMENT (STESTI and AREAVOL), and GEOMETRY (COORVIS and PFIGURE).

Teacher-reported opportunity-to-learn (OTL) information was also recorded for these items. This will be used as auxiliary information and will not be included in the analysis. For each item, the value 0 or 1 was recorded, where 1 was given if the mathematics needed to solve the item had been taught during eighth grade or in prior years and 0 was given otherwise. The achievement subscore averages and corresponding OTL averages for these eight variables are given in Table 1 for both the pretest and posttest occasions. One sees that OTL varies considerably over subscores. For the arithmetic topics of RPP and FRACT, the value of the OTL variable is close to the maximum score of 8 while, for the algebra topic of EQEXP and the geometry topic of PFIGURE, the OTL value is only about two thirds of the possible maximum. For EQEXP and PFIGURE, OTL also has a relatively large standard deviation, corresponding to the tracking effect of different classes' putting different emphasis on algebra and geometry training.

The substantive questions of interest in this article are the variance decomposition of the subscores with respect to within-class student variation and between-class variation and the change of this decomposition from pretest to posttest. In the SIMS (Crosswhite et al., 1985), such variance decomposition relates to effects of tracking and differential curricula in eighth-grade math. On the one hand, one may hypothesize that effects of selection and instruction tend to increase between-class variation relative to within-class variation, assuming that the classes are homogeneous, have different performance levels to begin with, and show faster growth for higher initial performance level. On the other hand, one may hypothesize that eighth-grade exposure to new topics will increase individual differences among students within each class so that

Table 1
Performance and opportunity to learn for eight math achievement subscores

Subscore	Number of items	Pretest		Posttest		Opportunity to learn	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
RPP	8	3.41	2.12	4.19	2.30	7.48	1.19
FRACT	8	3.28	1.96	4.09	2.14	7.52	0.83
EQEXP	6	2.38	1.42	2.98	1.63	3.96	1.82
INTNUM	2	0.65	0.70	1.04	0.79	1.88	0.41
STESTI	5	2.91	1.30	3.14	1.37	4.09	1.18
AREAVOL	2	0.64	0.74	0.91	0.81	1.74	0.54
COORVIS	3	1.14	0.91	1.51	0.98	1.61	0.95
PFIGURE	5	1.63	1.27	2.32	1.47	2.91	1.45

posttest within-class variation will be sizable relative to posttest between-class variation.

Conventional Analysis: Random Effects ANOVA

Because the SIMS (Crosswhite et al., 1985) data hierarchy involves school, class, and student, I will consider the following three-level decomposition using a standard random effects nested analysis of variance model (e.g., see Winer, 1971):

$$y_{ghi} = \mu + \alpha_g + \beta_{gh} + \gamma_{ghi}, \quad (1)$$

for individual i observed within the hth class within the gth school. Here, μ is the overall mean, and α, β, γ are independent random normal variables with zero means and variances to be estimated. The variance estimates can be obtained for each of the achievement subscores at both pretest and posttest using BMDP3Vs maximum likelihood estimator for unbalanced, nested data (Dixon, 1983).

Table 2 gives the ANOVA variance decomposition in terms of variance estimates and percentages, following the model in (1). The estimates show that the within-class, student-level percentage of the variance clearly dominates the subscore variability. The within variation is about 60%–80% whereas the between variation is divided into about 20%–30% for classes and about 3%–13% for schools. Both within variation and between variation increase over time.

The two right-most columns of Table 2 give the difference of the posttest and pretest value relative to the pretest value for between (school and class) and

Table 2
Analysis of variance decomposition of achievement scores
(percentages of total variance in parenthesis)

	Number of items	Pretest				Posttest				% Increase	
		School	Class	Student	Prop. between	School	Class	Student	Prop. between	Between	Within
RPP	8	.189[a] (4.2)	1.353 (29.9)	2.990 (66.0)	.34	.638 (11.8)	1.446 (26.7)	3.326 (61.5)	.38	35	11
FRACT	8	.337[a] (8.8)	1.123 (29.4)	2.366 (61.8)	.38	.557 (11.9)	1.349 (28.9)	2.767 (59.2)	.41	31	17
EQEXP	6	.089[a] (4.4)	.454 (22.5)	1.473 (73.1)	.27	.260 (9.7)	.781 (29.1)	1.646 (61.3)	.39	92	18
INTNUM	2	.020[a] (4.0)	.107 (21.2)	.358 (70.9)	.29	.053 (8.3)	.142 (22.3)	.442 (69.4)	.31	54	24
STESTI	5	.159 (9.1)	.421 (24.2)	1.163 (66.7)	.33	.179 (9.3)	.485 (25.2)	1.258 (65.5)	.34	15	8
AREAVOL	2	.017[a] (3.1)	.077 (14.1)	.451 (82.8)	.17	.062 (9.6)	.094 (14.6)	.490 (75.9)	.24	66	9
COORVIS	3	.028[a] (3.4)	.145 (17.5)	.656 (79.1)	.21	.073 (7.6)	.202 (21.2)	.680 (68.3)	.32	59	4
PFIGURE	5	.062[a] (3.9)	.301 (19.0)	1.224 (77.1)	.23	.274 (12.7)	.437 (30.1)	1.451 (67.1)	.33	96	19

[a] Not significant at 5% level

within. This shows that the between components increase much more than the within components. The between variation increase is particularly large for the algebra content of EQEXP and the geometry content of PFIGURE. This is in line with the OTL variation across classes discussed in conjunction with Table 1. In terms of percentage of total variation, however, the within variation decreases only an average of about 5%–10% over time. This suggests that the heterogeneity within classes remains very large. The influence of tracking on between-class variation in math achievement results in a strong increase in between variation over eighth grade.[1] Within variation still clearly dominates, although it does not increase much over time.

One should, however, note that the within variation includes individual-level measurement error variance which would inflate the contribution of within variation. Also, the relative size of the measurement error is presumably larger at pretest than at posttest because less learning has taken place at pretest. This would confound comparisons over time of relative variance contributions. The issue of how measurement error can be taken into account will be considered next.

Multilevel Factor Analysis

Each of the achievement subscores is created by summing a small number of dichotomous items. As a result, the subscores are likely to contain a sizable amount of measurement error. At the same time, the achievement subscores pertain to various aspects of central eighth-grade math topics. This suggests the use of a multivariate measurement model in the form of a factor-analytic, multiple-indicator model for the eight subscores.

A Multilevel Factor Model

In line with (1), consider a variance component decomposition of the multidimensional observation vector $\mathbf{y}$ for individual i in group (class) g

$$\mathbf{y}_{gi} = \mathbf{v} + \mathbf{y}_{Bg} + \mathbf{y}_{Wgi}, \tag{2}$$

where $\mathbf{v}$ is an overall mean and the between component $\mathbf{y}_{Bg}$ and the within component $\mathbf{y}_{Wgi}$ are independent as in conventional random effects analysis of variance. To simplify the exposition in this analysis, the between component represents both class and school contributions to the individual's score. The within component still represents the within-class contribution of the student. In this way the variance of $\mathbf{y}_{gi}$ can be decomposed into a between and a within part,

$$\Sigma = \sum_{B} + \sum_{W}. \tag{3}$$

In this article, I will assume that a one-factor model holds for both the between and the within components of (2) and (3). For a given y variable indexed j in the vector $\mathbf{y}_{gi}$, one can then use the following decomposition,

$$y_{gij} = \nu_j + \lambda_{Bj}\eta_{Bg} + \epsilon_{Bgj} + \lambda_{Wj}\eta_{Wgi} + \epsilon_{Wgij}, \tag{4}$$

where the λs are loading parameters and the four random variables are independent.

The within part of (4) can be interpreted in line with conventional factor analysis in that the within factor η_{Wgi} and the within residual ϵ_{Wgi} refer to individual-level variation. In the multilevel model, this is within-group variation. The single factor accounts for all covariation among the individual-level achievement scores, representing a general math achievement trait for these eighth graders. While other research on these data suggests several minor factors on the item level (e.g., see Muthén, 1988), these factors largely vanish in the aggregated scores. The residuals are viewed as measurement errors—that is, variable-specific individual variation not accounted for by the factor. These errors are independent of the factor and are independent of each other.

The between part of (4) departs from conventional analysis in that it addresses across-group variation rather than across-individual variation. Here, the factor η_B is interpreted in terms of selection effects due to tracking and differences in curricula. The single factor represents a single dimension on which selection is made with differing λ_B coefficients giving different weights to different topics. One may, for example, hypothesize that, on entering eighth-grade, selection is dominated by previous arithmetic performance so that the between loadings are relatively higher for RPP and FRACT. During eighth grade, curricula vary greatly in terms of both algebra and geometry topics. This means that at posttest the variance of the between residual ϵ_B may increase for these topics and the unidimensionality of the between factor model may be called into question.

Consider now the variance components of (4) for observed variable j,

$$\begin{aligned}\sigma^2_{ygij} &= \lambda^2_{Bj}\sigma^2_{\eta B} + \sigma^2_{\epsilon Bj} + \lambda^2_{Wj}\sigma^2_{\eta W} + \sigma^2_{\epsilon Wj} \\ &= \mathrm{BF} + \mathrm{BE} + \mathrm{WF} + \mathrm{WE},\end{aligned} \tag{5}$$

where B and W stand for between and within while F and E stand for factor and error.

The conventional factor analysis definition of reliability uses the R^2-like ratio of variance due to the factor divided by total variance in y. Because this refers to individual-level reliability, the analogous reliability definition for y_j in multilevel data in our factor model is:

$$\text{Within reliability } (y_j) = \mathrm{WF}/(\mathrm{WF} + \mathrm{WE}) \tag{6}$$

while the reliability in the across-group variation is:

$$\text{Between reliability } (y_j) = \mathrm{BF}/(\mathrm{BF} + \mathrm{BE}). \tag{7}$$

In line with conventional factor analysis, these reliabilities may be viewed as lower limits since the WE and BE measurement error components also include variable-specific variation.

Analogous to random effects analysis of variance (Winer, 1971), one notes that the model implies the following correlation between two individuals i and i' in group g for variable y_j,

$$\begin{aligned}\mathrm{Corr}\,(y_{gij}, y_{gi'j}) &= \mathrm{Cov}\,(y_{gij}, y_{gi'j})/\sigma^2_{ygij} \\ &= (\mathrm{BF} + \mathrm{BE})/(\mathrm{BF} + \mathrm{BE} + \mathrm{WF} + \mathrm{WE}).\end{aligned} \tag{8}$$

This intraclass correlation (e.g., see Fisher, 1958; Haggard, 1958; Koch, 1983) is also the proportion of between variance in y_j in the present model. The larger it is, the further one deviates from the conventional assumption of all observations' being independent. If BF and BE are both zero for all variables, there is no need for a multilevel analysis, and the independence assumption of a conventional analysis is fulfilled. One notes that the intraclass correlation is influenced by measurement error.

Given the decomposition in (5) into factor and error variance, one may consider an error-free version of the variance ratio in (8)—namely, the error-free proportion of between variance, or *true intraclass correlation coefficient,* for each variable,

$$\mathrm{BF}/(\mathrm{BF} + \mathrm{WF}). \tag{9}$$

One may also consider the error-free increase in both between and within variance from pre- to posttest:

$$(\mathrm{BF}_{\mathrm{post}} - \mathrm{BF}_{\mathrm{pre}})/\mathrm{BF}_{\mathrm{pre}}, \tag{10}$$

$$(\mathrm{WF}_{\mathrm{post}} - \mathrm{WF}_{\mathrm{pre}})/\mathrm{WF}_{\mathrm{pre}}. \tag{11}$$

Multilevel Factor Analysis Estimation

The multilevel factor modeling just presented leads to a covariance structure model for two-level data which uses a conventional factor analysis covariance structure on both the between and within level. Muthén and Satorra (1989) and Muthén (1989, in press-a) consider variations of multilevel models leading to these and related covariance structures. These articles also show the relationship of these models to random parameter models that have become popular in educational applications of regression analysis (e.g., see Raudenbush & Bryk, 1988). Essentially, the above modeling may be viewed as a random factor means, random measurement intercepts model. Random slopes are not involved. The modeling can be extended to three-level data for school, class, and student, but that will not be considered here. Maximum likelihood (ML) estimation of multilevel covariance structure models was studied already by Schmidt (1969). (Also see Schmidt & Wisenbaker, 1986.) But these techniques do not appear to have come into practical use. More recent contributions are from Goldstein and McDonald (1988), McDonald and Goldstein (1989), Longford and Muthén (1990), and Muthén (1989, in press-a). I showed that multiple-group structural equation modeling software can be modified for multilevel factor analysis (MFA) ML analysis. In line with this idea, I proposed a simpler ML-based MFA estimator which can be used with already existing multiple-group structural equation software such as LISREL, LISCOMP, and EQS. This estimator uses the customary between and pooled-within sample covariance matrices. In the balanced case, it is equivalent to the MFA ML estimator. In the unbalanced case, the estimator is consistent and, despite the fact that it uses less information than ML, has given similar results in the analyses to date (Muthén, in press-a). Approximate values for chi-square tests of model fit and standard errors of estimates are also obtained. For technical details, see Muthén (in press-a). A pedagogical introduction is given in Muthén (in press-b).

Multilevel Analysis

Of particular interest in the achievement analysis is the change from pretest to posttest in variance contributions. An efficient way to use the data is to perform a simultaneous analysis of pretest and posttest data. Such a longitudinal model also makes it possible to study change in variance contributions due to the between and within factors, ensuring that these factors are measured in comparable metrics over time.

As a first step, preliminary analyses of the pretest and posttest were carried out following the multilevel analysis steps suggested by Muthén (in press-a, in press-b). Drawing on these analyses, the longitudinal model specifies one between factor and one within factor as in (4) for each of the two time points. The two between factors are allowed to be correlated, and so are the two within factors. The variables may also be allowed to correlate over time via correlated measurement errors. The need for correlated individual-level errors is often found in conventional covariance structure analyses where the same instrument is repeatedly administered. Here, this is extended to correlation of

between errors. For example, if at pretest classroom differences in algebra were beyond what could be explained by the general level of the pretest (a proxy for the between factor), this algebra difference might prevail at posttest leading to a between error correlation.

To be able to study change in factor-related variances over time, it is necessary to specify the same metric for both the between and the within factor over time. To this aim, one restricts the loadings to be equal over time at both the between and within level. It is known a priori, however, that some math topics become much more familiar to the students over the course of the eighth grade and therefore may lead to different measurement properties of subscores over time. Coverage of other topics does not change as much over time. This is also supported by the OTL differences and the pre–post differences in estimated reliabilities. For such reasons, I will allow the variables EQEXP and PFIGURE to have different loadings over time, whereas the other loadings will be held equal, apart from the fixed loading for RPP.

Table 3 shows the different steps of MFA model fitting. For comparison, results from analyzing the conventional (total) covariance matrix S_T and the pooled-within covariance matrix S_{PW} are given in addition to MFA results. In the baseline model 1, there is no loading invariance imposed, and no error correlations are included. The fit is poor for this model. Adding loading invariance over time as in model 2 also gives a poorly fitting model. Model 3, using partial loading invariance, improves the fit considerably with a loss of only a few degrees of freedom. Adding correlated within errors as in model 4 gives a

Table 3
Longitudinal factor analysis model tests

Model	Chi-square* (d.f.)		
	S_T	S_{PW}	MFA
1. Baseline	1,041.38 (103)	687.43 (103)	1,101.90 (206)
2. 1 + loading invariance	1,160.47 (110)	734.26 (110)	1,183.59 (220)
3. 1 + partial loading invariance	1,072.98 (108)	697.39 (108)	1,122.42 (216)
4. 3 + correlated within errors	380.51 (100)	222.51 (100)	595.24 (208)
5. 4 + correlated between errors	-	-	461.55 (200)

* For MFA, approximate chi-square values are reported

dramatic improvement in fit. It is interesting to note the large difference in fit between using S_T and using S_{PW}. Using S_{PW} gives consistent estimates of the within part of the multilevel model (Muthén, in press-a). This analysis points to a good individual-level fit.[2] Model 5 adds correlated between errors to the MFA model resulting in a significant improvement in fit. Comparing the model 4 result for S_{PW} with the model 5 result for MFA shows that doubling the number of degrees of freedom by addition of the between structure does not result in much more than a doubling of the chi-square value. We conclude that the MFA model 5 fits reasonably well in both its within and between part, given the large sample size.

Although all are significant, the within error correlations are rather small, in the range 0.07–0.21. The between error correlations are considerably larger, in the range 0.25–0.78. The factor correlations point to a strong linear relationship over time, .80 for within and .93 for between.

Table 4 gives estimated variance ratios for the eight subscores in the form of reliabilities, error-free proportion between variance, and error-free increases from pre- to posttest (see Equations 6–11). The MFA within-reliabilities (see definitions in "Multilevel Factor Analysis Estimation") are very low, as should be expected given the small number of items comprising each subscore. As expected, the highest reliability values occur for the arithmetic topics of RPP and FRACT, perhaps not only because these subscores consist of more items than the others but also because these topics have higher eighth-grade OTL as shown in Table 1. There is a strong increase over time, particularly for EQEXP and PFIGURE. These correspond to new topics at pretest for many eighth graders, whereas they have been better covered at posttest.

The between reliabilities are very high and do not change much from pretest to posttest. It is interesting to note that the largest reliability increase occurs for the algebra content of EQEXP, meaning that EQEXP becomes a better measurement at posttest of the dimension that makes classrooms differ. On the whole, however, the indicators of the between factor are very homogeneous, adding very little variation around the general dimension.

Table 4
Item characteristics estimated from the eight variable longitudinal multilevel factor analysis model

	Reliability				Error-free prop. between		Error-free % increase	
	Pre		Post					
	Between	Within	Between	Within	Pre	Post	Between	Within
RPP	.97	.43	.96	.53	.54	.52	29	41
FRACT	.96	.40	.97	.48	.60	.58	29	41
EQEXP	.82	.17	.93	.32	.65	.64	113	117
INTNUM	.82	.19	.88	.23	.63	.61	29	41
STESTI	.85	.26	.89	.34	.58	.56	29	41
AREAVOL	.83	.17	.82	.23	.54	.52	29	41
COORVIS	.91	.19	.80	.26	.57	.55	29	41
PFIGURE	.77	.16	.77	.32	.60	.54	87	136

It is interesting to return to the question of variance decomposition discussed in the sections, "The Data and the Substantive Research Questions" and "Conventional Analysis: Random Effects ANOVA," and compare the results of conventional random effects analysis of variance with those of the multilevel factor analysis. Table 4 also includes the error-free proportions of between variation, or true intraclass correlations. The error-free proportion between for each variable is calculated as in (9) using the error-free variance ratio BF/(BF + WF) in the notation of (5). The values are around 0.6 with little difference between the pre- and posttest results. This value should be compared to the observed variable intraclass correlations, or proportion between variation, of Table 2 which were in the range 0.2–0.4. In this way, between-class variation becomes relatively more important when purging the measurement error in the scores. This is in line with the expectation that measurement error inflates the within variation. While Table 2 shows a slight increase over time in the proportion between for all variables, the error-free proportion between of Table 4 shows a slight decrease for several variables.

An additional indication of the inflation of within variation due to measurement error is seen in the MFA estimates given in the two right-most columns of Table 4. As in the ANOVA results of Table 2, these columns display the increase in variance from pretest to posttest relative to the pretest value, although they are now purged of measurement error (see Equations 10 and 11). The ANOVA and MFA results show very different pictures. Overall, MFA shows the between variance increase over time to be smaller than for ANOVA and the within increase to be much larger. In fact, with MFA the within increase is the largest with one exception. The ANOVA results are distorted by individual-level measurement error. Despite an increase over time in true within variance due to the factor, the decrease in the measurement error variance over time substantially dampens the total within variance increase as judged by ANOVA. Using MFA, one finds that the error-free within variance increases dramatically over time, or, in other words, within-class student heterogeneity increases dramatically. This would support the hypothesis that increasing learning opportunities leads to increasing individual differences. Taken together with the finding that the within variance increases more than the between variance, this may suggest that the major effect of the U.S. eighth-grade tracking system for math is not so much that classes become more different but that students within classes become more different. This is in sharp contrast to the ANOVA findings.

Unreliability Sensitivity Analyses

The multilevel factor analysis results provide estimates of the error-free proportions of between variance for each variable. They also provide an error-free assessment of change from pre- to posttest in between and within variance. The corresponding observed variable quantities from analysis of variance presented in Table 2 are quite different. The Table 2 results point to a larger share of within variation and a smaller increase over time in within variation. The differences are hypothesized to be due to measurement error. Using more reliable scores might not make for such a large difference in

conclusions. More reliable scores can be obtained by the summing of more dichotomous items. Because the assumption of unidimensionality of the items has been supported in the analysis, one may contemplate the use of more aggregated subscores. It is of practical interest to get a feeling for how different amounts of aggregation and reliabilities affect the results. Also, the use of different aggregation levels gives a check of the robustness of the MFA results. It is of practical interest to know if the eight variable factor analysis, using variables with very low reliability, gives trustworthy results for the error-free estimates.

To study influence of unreliability, RPP and FRACT were combined into a single ARITHMETIC score based on 16 items; EQEXP and INTNUM were combined into an ALGEBRA score based on 8 items; STESTI and AREAVOL were combined into a MEASUREMENT score based on 7 items, and COORVIS and PFIGURE were combined into a GEOMETRY score based on 8 items. Also, a total score based on all 39 items was used.

Table 5 gives analysis of variance estimates for these new scores. Consider first the total score. This score may be viewed as a proxy for the factor in the MFA.

The proportion between is .52 at pretest and .53 at posttest. These values are not too far off from the Table 4 average values in the columns "Error-free prop. between," reflecting the reliability of the total score. The percent increase in Table 5 compares reasonably well to the average value of the Table 4 column "Error-free % increase." It is clear, however, that the total score cannot capture the differences in increase for the different subscores exhibited in Table 4. In sum, using the total score does not give misleading ANOVA results, but it certainly gives undifferentiated ones.

Consider next the use of the four aggregated scores ARITHMETIC, ALGEBRA, MEASUREMENT, and GEOMETRY. A comparison of Table 5 with Table 4 shows ANOVA biases similar to those that were observed for the eight less aggregated variables. The general conclusion for ANOVA is still that

Table 5
Analysis of variance decomposition of four aggregated achievement scores (percentages of total variance in parenthesis)

Score	Number of items	Pretest				Posttest				% Increase	
		School	Class	Student	Prop. between	School	Class	Student	Prop. between	Between	Within
TOTAL	39	4.404[a] (7.8)	24.377 (43.3)	27.576 (48.9)	.52	12.532 (15.2)	31.039 (37.7)	38.750 (47.1)	.53	51	41
ARITHMETIC	16	1.016[a] (7.5)	4.919 (36.3)	7.607 (56.2)	.44	2.349 (13.7)	5.589 (32.7)	9.173 (53.6)	.46	34	21
ALGEBRA	8	.172[a] (5.3)	.964 (29.3)	2.122 (65.1)	.35	.502 (10.9)	1.571 (34.0)	2.543 (55.1)	.45	83	20
MEASUREMENT	7	.229[a] (7.8)	.815 (27.9)	1.882 (64.3)	.36	.430 (12.2)	.949 (26.9)	2.155 (61.0)	.39	32	15
GEOMETRY	8	.131[a] (4.1)	.842 (26.1)	2.247 (69.8)	.30	.585 (13.3)	1.129 (25.6)	2.701 (61.2)	.39	76	20

[a] Not significant at 5% level

the proportion between is underestimated, the percent change in between is overestimated, and the percent change in within is underestimated. From a practical point of view, it is interesting to note that the ANOVA biases are quite large even for the 16-item score of ARITHMETIC. Going from the 8-item subscores of RPP and FRACT in Table 2 to the 16 items of ARITHMETIC in Table 5 decreases the ANOVA bias considerably but not sufficiently. The effects of unreliability make it impossible in the analysis of variance to distinguish math topic differences between subscores from differences in the number of items used to create the score.

Table 6 gives the results of the longitudinal MFA using the new set of four achievement scores. A model analogous to model 5 in Table 3 is used. The four-variable MFA results of Table 6 give a picture similar to the eighth-variable results of Table 4. In this data set, it is clear that, unlike ANOVA, MFA is not sensitive to the level of variable aggregation and reliability. I conclude that, while ANOVA needs long subtests for trustworthy results, MFA can use variables consisting of few items.

Conclusions

Multilevel factor analysis has been shown to give new types of useful information on educational test scores. Using the structure of the sample design, the effects of clustered (nested) observations is not only taken into account but also modeled in interesting ways to shed light on within- and between-class variance components and their changes over time.

From a substantive point of view, it was found that the strong elements of tracking in eighth-grade math classes result in between-class variation in the achievement scores which is about as large as the within-class student variation. At the same time, however, within-class variability increases much more than between-class variation over the course of eighth grade. Individual differences appear to be sharply increased by instructional exposure. Increases in both between and within variation are particularly dramatic for algebra topics related to equations and expressions and geometry topics related to plane figures.

From a methodological point of view, several interesting findings emerge. Due to unreliability in the observed scores, the results obtained by analysis of

Table 6
Item characteristics estimated from the four-variable longitudinal multilevel factor analysis model

	Reliability				Error-free prop. between		Error-free % increase	
	Pre		Post					
	Between	Within	Between	Within	Pre	Post	Between	Within
ARITHMETIC	.98	.58	.96	.65	.57	.55	27	36
ALGEBRA	.85	.26	.94	.45	.64	.63	99	110
MEASUREMENT	.93	.38	.95	.46	.57	.55	27	36
GEOMETRY	.87	.27	.88	.46	.59	.55	71	102

variance are quite different from those of MFA. ANOVA substantially underestimates the intraclass correlation, or the proportion of between-class variation, and substantially underestimates the increase over time in within-class variation. For trustworthy ANOVA results on sums of dichotomously scored items, large sets of items are needed which may preclude differentiation of subtopics. MFA can achieve trustworthy results using subscores created from only a few items. MFA is a readily available technique because it can be carried out with standard structural equation software. Given this, it is hoped that educational researchers quickly adopt these exciting new analysis tools. However, MFA is not a small sample technique. In particular, MFA calls for data that have a sizable number of groups, preferrably at least about 50–100. As was pointed out by Cronbach (1976), if cost permits, it may be better to observe fewer students per class in favor of including more classes.

Several extensions of the MFA models studied here are possible. In addition to class-level components of student-level variables, one may include class-level variables. For example, the class-level OTL information can be incorporated to explain the class-level student achievement variation (see Muthén, in press-a). The modeling is not limited to factor analysis, but structural equation models can also be analyzed (see Schmidt & Wisenbaker, 1986; Muthén, 1989). More than two levels of nesting can be incorporated. All of these extensions fit into conventional software using the estimator of Muthén (1989, in press-a). The MFA techniques are, of course, not limited to educational data with students observed within classes and schools but can be used in any situation where cluster sampling has been employed, such as surveys with geographically determined groups of households.

Notes

[1]Table 2 shows that the increase over time in proportion between is largely due to an increase in the school component. Notable class component increases are seen for variables corresponding to new topics in eighth-grade math—namely, EQEXP and PFIGURE. In the further analyses, no distinction will be made between school and class variation, but together they will be viewed as between-class variation.

[2]Nominally, this outcome should lead to rejection of the model. The large sample size may, however, give rise to a large power of model tests so that trivial departures from the model are detected. It should be noted, however, that little if anything is known about the power of rejection for between structures in multilevel factor models, since this is influenced both by the number of groups and the number of individuals.

References

Bock, R. D. (1989). *Multilevel analysis of educational data.* San Diego, CA: Academic.

Bohrnstedt, G. W. (1983). Measurement. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of survey research* (pp. 69–121). New York: Academic.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bond, G. L., & Dykstra, R. (1967). The Cooperative Research Program in first-grade reading instruction. *Reading Research Quarterly, 2,* 5–142.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. *Review of Research in Education, 8,* 158–233.

Burstein, L. (Ed.). (in press). *The IEA study of Mathematics III: Student growth and classroom process.* London: Pergamon.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design, and analysis.* Unpublished manuscript, Stanford University, Stanford Evaluation Consortium.

Crosswhite, F. J., Dossey, J. A., Swafford, J. O., McKnight, C. C., & Cooney, T. J. (1985). *Second international mathematics study: Summary report for the United States.* Champaign, IL: Stipes.

Dixon, W. J. (1983). *BMDP statistical software.* Berkeley: University of California Press.

Fisher, R. A. (1958). *Statistical methods for research workers* (13th ed.). New York: Hafner.

Goldstein, H. I., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53,* 455–467.

Haggard, E. A. (1958). *Intraclass correlation and the analysis of variance.* New York: Dryden.

Harnqvist, K. (1978). Primary mental abilities of collective and individual levels. *Journal of Educational Psychology, 70,* 706–716.

Joreskog, K. G., & Sorbom, D. (1979). *Advances in factor analysis and structural equation models.* Cambridge, MA: Abt Books.

Koch, G. G. (1983). Intraclass correlation coefficient. *Encyclopedia of Statistical Sciences, 4,* 212–217.

Longford, N. T., & Muthén, B. (1990). *Factor analysis for clustered observations* (UCLA Statistics Series No. 71). Los Angeles: University of California.

McDonald, R. P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology, 42,* 215–232.

Muthén, B. (1988). *Instructionally sensitive psychometrics: Applications to the Second International Mathematics Study.* Los Angeles: University of California, Graduate School of Education.

Muthén, B. (1989). Latent variable modeling in heterogenous populations [Presidential address to the Psychometric Society]. *Psychometrika, 54,* 557–585.

Muthén, B. (in press-a). Mean and covariance structure analysis of hierarchical data. *Journal of Educational Statistics.*

Muthén, B. (in press-b). Multilevel covariance structure analysis. *Sociological Methods & Research.*

Muthén, B., Kao, C-F., & Burstein, L. (1991). Instructional sensitivity in mathematics achievement test items: Applications of a new IRT-based detection technique. *Journal of Educational Measurement, 28,* 1–22.

Muthén, B., & Satorra, A. (1989). Multilevel aspects of varying parameters in structural models. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 87–99). San Diego: Academic.

National Center for Education Statistics. (1989). *Policy/analysis agenda for base year NELS: 88 and teacher follow-ups.* Washington, DC: Author.

Rakow, E. A., Airasian, P. W., & Madaus, G. F. (1978). Assessing school and program effectiveness: Estimating teacher level effects. *Journal of Educational Measurement, 15,* 15–21.

Raudenbush, S., & Bryk, A. (1988–89). Methodological advances in studying effects of schools and classrooms on student learning. In E. Z. Roth (Ed.), *Review of Research in Education* (423–475). Washington, DC: American Educational Research Association.

Schmidt, W. H. (1969). *Covariance structure analysis of the multivariate random effects model.* Unpublished doctoral dissertation, University of Chicago.

Schmidt, W., & Wisenbaker, J. (1986). *Hierarchical data analysis: an approach based on structural equations* (Tech. Rep. No. 4.). East Lansing, MI: Michigan State University, Department of Counseling Educational Psychology and Special Education.

Schmidt, W., Wolfe, R. G., & Kifer, E. (in press). The identification and description of student growth in mathematics achievement. In L. Burstein (Ed.), *The IEA Study of Mathematics III: Student Growth and Classroom Processes.* London, England: Pergamon.

Skinner, C. J., Holt, D., & Smith, T. M. F. (1989). *Analysis of complex surveys.* Chichester, England: Wiley.

Wiley, D. E., & Bock, R. D. (1967). Quasi-experimentation in educational settings: Comment. *The School Review, 75*(4), 353–366.

Winer, B. J. (1971). *Statistical principles in experimental design.* New York: McGraw-Hill.

Author

BENGT O. MUTHÉN is Professor, Graduate School of Education, University of California—Los Angeles, 405 Hilgard Ave., Los Angeles, CA 90024-1521. *Degree:* PhD, University of Uppsala. *Specializations:* categorical data and latent variable modeling.