Education Policy Analysis Archives

Volume 9 Number 34 **September 14, 2001** **ISSN 1068-2341**

A peer-reviewed scholarly journal
Editor: Gene V Glass, College of Education
Arizona State University

Articles appearing in **EPAA** are abstracted in the *Current Index to Journals in Education* by the ERIC Clearinghouse on Assessment and Evaluation and are permanently archived in *Resources in Education*.

Predicting Variations in Mathematics Performance in Four Countries Using TIMSS

Daniel Koretz
Harvard University

Daniel McCaffrey
RAND Education

Thomas Sullivan
RAND Education

Citation: Koretz, D., McCaffrey, D., and Sullivan, T. (2001, September 14). Predicting variations in mathematics performance in four countries using TIMSS. *Education Policy Analysis Archives*, *9*(34). Retrieved [date] from http://epaa.asu.edu/epaa/v9n34/.

Abstract
Although international comparisons of average student performance are a staple of U.S. educational debate, little attention has been paid to cross-national differences in the variability of performance. It is often assumed that the performance of U.S. students is unusually variable or that the distribution of U.S. scores is left-skewed – that is, that it has an

unusually long ‘tail' of low-scoring students – but data from international studies are rarely brought to bear on these questions. This study used data from the Third International Mathematics and Science Study (TIMSS) to compare the variability of performance in the U.S., Australia, France, Germany, Hong Kong, Korea, and Japan; investigate how this performance variation is distributed within and between classrooms; and explore how well background variables predict performance at both levels. TIMSS shows that the U.S. is not anomalous in terms of the amount, distribution, or prediction of performance variation. Nonetheless, some striking differences appear between countries that are potentially important for both research and policy. In the U.S., Germany, Hong Kong, and Australia, between 42 and 47 percent of score variance was between classrooms. At the other extreme, Japan and Korea both had less than 10 percent of score variance between classrooms. Two-level models (student and classroom) were used to explore the prediction of performance by social background variables in four of these countries (the U.S., Hong Kong, France, and Korea). The final models included only a few variables; TIMSS lacked some important background variables, such as income, and other variables were dropped either because of problems revealed by exploratory data analysis or because of a lack of significance in the models. In all four countries, these sparse models predicted most of the between-classroom score variance (from 59 to 94 percent) but very little of the within-classroom variance. Korea was the only country in which the models predicted more than 5 percent of the within-classroom variance in scores. In the U.S. and Hong Kong, the models predicted about one-third of the total score variance, and almost all of this prediction was attributable to between-classroom differences in background variables. In Korea, only 19 percent of total score variance was predicted by the model, and most of this most of this was attributable to within-classroom variables. Thus, in some instances, countries differ more in terms of the structure and prediction of performance variance than in the simple amount of variance. TIMSS does not provide a clear explanation of these differences, but this paper suggests hypotheses that warrant further investigation.

Introduction

International comparisons of average student performance are widely discussed by policymakers and the press and have had a powerful influence on educational debate and policy in the US. In an era when traditional norm-referenced reporting of student performance ostensibly has gone out of favor, "country norms" have become an increasingly important indicator of the success of US education and the levels of performance to which this country should aspire. The publication of the results of the Third International Mathematics and Science Study (TIMSS) over the past several years (Beaton et al., 1996a, 1996b; Mullis et al., 1997, 1998) has increased further the prominence of international comparisons in the US debate.

Much of the discussion of international comparisons has focused on horse-race

comparisons of means or medians. Although presented in TIMSS reports, information on the *variability* of student performance has usually been ignored in the US debate or has been used in a lopsided and potentially misleading fashion. Typically, the variability in the US has been considered, while the variability in the countries to which the US is compared has been ignored. For example, earlier this decade, the results of the 1991 International Assessment of Educational Progress (IAEP) were projected onto the National Assessment of Educational Progress (NAEP) scale, permitting comparison of countries participating in IAEP to states participating in the 1992 NAEP Trial State Assessment in mathematics. These comparisons, which have been widely cited, showed that the highest-scoring US states, such as Iowa and North Dakota, had mean scores similar to those of the highest-scoring countries, such as Taiwan and Korea (National Center for Education Statistics, 1996, Figure 25). High-scoring regions in Taiwan and Korea, however, were not compared to the US mean.

Underlying some of these comparisons appears to be an expectation that the variability of student performance is atypically large in the US. Indeed, some observers have made this expectation explicit. For example, Berliner and Biddle, in disparaging the utility of international comparisons of mean performance, wrote:

> The achievement of American schools is a *lot* more variable than is student achievement from elsewhere….To put it baldly, American now has some of the finest, highest-achieving schools in the world—and some of the most miserable, threatened, underfunded educational travesties, which would fail by any achievement standard (1995, p. 58, emphasis in the original).

To buttress this assertion, they cited the NCES comparisons of US states and foreign nations noted above, which displayed no information about the variation of performance in other countries and included no information about the variation of performance among schools within any country.

Research Questions

This study was undertaken to explore the variability of performance in the US and several other countries using TIMSS data. Specifically, the study explored two primary questions:

1. How large is the performance variation in our sample countries, and how is this variation distributed between and within classrooms?
2. How well do background variables predict performance variation in our countries, both within and between classrooms?

The results reported here are limited to mathematics in the higher grade in Population 2 (grade 8). We focused on Population 2 rather than Population 1 (elementary grades) because of doubts about the validity and utility of self-report data from elementary school students.(Note 1) Population 3 (end of high school) presented formidable difficulties of sample non-equivalence. The analyses focused on mathematics because the TIMSS sample design which selected students based on the mathematics classes they attended rather than the science classes (Foy, Rust, and Shleicher, 1996, p. 4-7). This precluded decomposition of score variation and hierarchical modeling in science.

Methods

To answer these research questions, our analyses proceeded in two steps:

1. We compared the distributions of student-level performance across all the countries in the Population 2 sample.
2. We used a smaller, purposive subsample of countries to analyze the variability in student performance between and within classrooms and to explore the contributions of student background characteristics to both of these sources of variability.

The performance measure used in all analyses was BIMATSCR, the "international mathematics achievement score" (Gonzalez, Smith et al., 1997) used in TIMSS published reports for Population 2. Technically, BIMATSCR is not a score in the traditional sense, but it is labeled a score here for simplicity. TIMSS was designed to provide aggregate estimates but not scores for individual students. In lieu of scores, TIMSS provides for each student five plausible values, which are "random draws from the estimated ability distribution of students with similar item response patterns and background characteristics" (Gonzalez, Smith et al., 1997, p. 5-1). In this respect, TIMSS followed a variant of the procedures NAEP has used since 1984. In the case of Population 2, however, scores were conditioned only on country, gender, and class mean, not on background variables (Gonzalez, 1998). In theory, the variance of repeated estimates using different plausible values should be added to the sampling variance to obtain an estimate of error variance for statistics calculated with plausible values. However, Gonzalez, Smith et al. (1997, p. 5-8) report that the intercorrelations among TIMSS plausible values are so high that this error component can be ignored. It was not calculated for statistics reported in this paper.

The step 1 analyses are purely descriptive and use data available in TIMSS publications (Beaton et al., 1996a and 1996b; Mullis, et al., 1997, Martin et al., 1997).

Our initial purposive subsample for the more detailed analyses in step 2 included seven counties: Australia, France, Germany, Japan, Hong Kong, Korea, and the US. Japan and Korea were selected because they are often used as examples of high-performing countries in comparisons with the US. Germany was included because it is often noted in discussions of the competitiveness of the US workforce. Hong Kong was included because it has both parallels with and interesting differences from Japan and Korea. France was included because in eighth-grade mathematics, it showed an unusually small variance of performance. Australia was considered primarily for methodological reasons. Although we present some results for all seven countries, we limited modeling of the predictors of variance to four: the US, France, Hong Kong, and Korea. Students in Japan did not complete the survey items used in the modeling. Response patterns for students in Germany made us suspicious of that country's data. Since Australia was included more for methodological than for substantive reasons, we dropped it from the modeling because of similarities in the preliminary results from Australia and other countries.

In our second stage analyses we decomposed the variance among students scores from each of the countries into the variance within classrooms and the variance between classrooms, and in the four primary countries, we explored the predictors of variance at each of these levels. Ideally one would want to decompose the variance into at least three levels: within classrooms, between classrooms within schools, and between

schools. The school and classroom levels of aggregation are not exchangeable. For example, a decision to track students on the basis of ability would increase the variance between classrooms within schools while decreasing the variance within classrooms, but it would not directly affect the variance between schools. Conversely, residential segregation on the basis of social class would increase performance variance between schools, but it could decrease the variance between classrooms within schools by making schools more homogeneous with respect to achievement.

In all countries other than the US, Australia, and Cyprus, however, the TIMSS Population 2 sample consisted of a single classroom per school. Therefore, in most countries, one can only specify a two-level model in which variations in performance between schools and between classrooms within schools are completely confounded. Accordingly, we decomposed the variability in math scores from each of the four countries into within classroom variability and between classroom variability. The between classroom variability includes contributions from both the variation of classrooms within schools and the variation between schools.

To fit these models we sacrificed some of the richness of the US data in order to obtain comparable to the results from all four countries. We did this by creating a subsample of the US samples that consisted of a single classroom per school, randomly selected from the multiple classrooms in the original sample. We modified the sample weights and jackknife replicates used in variance estimation accordingly.

Our step 2 analyses followed the same course in each country and extended from simple exploratory data analysis (EDA) to hierarchical modeling. Extensive EDA was used to explore individual-level and classroom-level variations in performance and background variables, to determine whether background variables showed sufficient variability to be usable in analysis, to determine whether the relationships between background variables and performance appeared sensible, and to decide whether and how to categorize variables. The patterns uncovered by this EDA substantially constrained our analyses in several instances.

Simple bivariate relationships between performance and background variables were examined for all of the variables considered for the hierarchical models. When necessary, variables were recoded so that a positive relationship with scores would be expressed as a positive correlation. The bivariate analyses were carried out three ways because of the inherently hierarchical nature of the data: (1) student-level uncentered (i.e., simple student-level analyses without regard to classrooms); (2) student-level, centered on classroom means (corresponding to the within-classroom component of variance); and (3) classroom-level (corresponding to the between-classroom component of variance).

Hierarchical modeling using multiple background variables followed bivariate analyses. The models include the classroom mean for each background variable and the individual student-level values, centered on classroom means. With centering, the coefficients produced by the model separately measure each variable's contribution to both the between- and within-classroom variability.

TIMSS used a complex sampling plan with unequal probability of selection among schools from each country's sample. To account for this disproportionate sampling, all analyses reported here are weighted unless noted. Weighted analyses produce consistent estimates of model parameters even if the sample design is disproportionate or more technically nonignorable (see, e.g., Pfefferman, 1996 for discussion on the use of weights in model fitting). We used the methods of Pfefferman et al. (1998) to fit our weighted hierarchical models using specially written SAS macros. (For the macro and more detail on methods, see Koretz, McCaffrey, and Sullivan, 2000.)

Distributions of Student-Level performance in TIMSS

Basic information about the size of the performance variation in participating countries, analyzed at the level of students without regard to aggregation, is provided in TIMSS publications. Appendices to the reports provide standard deviations and selected percentiles (5th, 25th, 50th, 75th, and 95th) of the performance distributions (Beaton et al., 1996a and 1996b, Appendix E; Mullis et al., 1997, Appendix C; Martin et al., 1997, Appendix C).

At the level of individual students, the eighth-grade mathematics performance of US students was near the median of the 31 countries that met the TIMSS sampling requirements for the eighth grade (see Beaton, et al., 1996a, Tables 2.1 and E.3). The country-level standard deviations varied greatly, from 58 to 110, but half were clustered in the narrow range from 84 to 92. The median standard deviation across the 31 countries was 88. The standard deviation of the US sample was 91, only slightly above the international median. Among these 31 countries, the country-level standard deviation of eighth-grade mathematics performance was strongly predicted by country means: the higher the mean, the larger the standard deviation (r=.71; see Figure 1). Seen this way, the standard deviation of mathematics performance in the US was about nine percent higher than the value that would be predicted from the US mean. Numerous other countries, however, had standard deviations that deviated comparably from those predicted by their means. For example, clustered tightly around the US in Figure 1 are England and New Zealand, and Germany would be as well if it were included in Figure 1. Germany does not appear in Figure 1 because it did not meet all sampling requirements. (In eighth-grade science, the standard deviation in the US was indeed one of the largest, but it is not an outlier; see Koretz, McCaffrey, and Sullivan, 2000.)

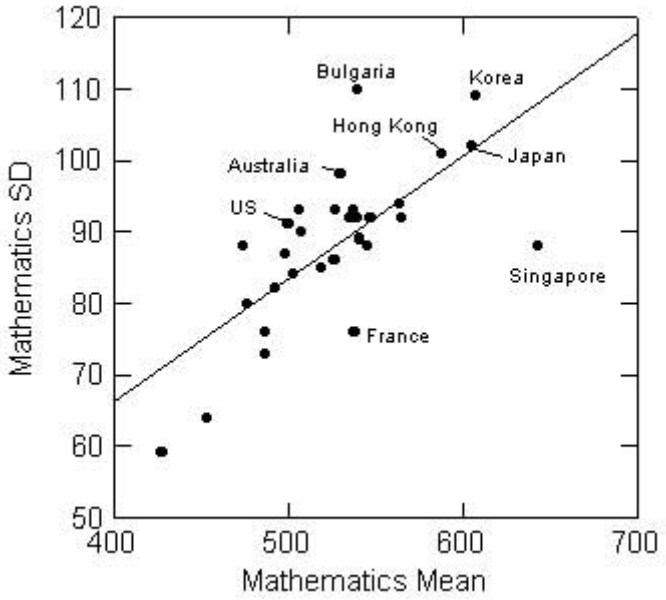

Figure 1. Plot of Mathematics Standard Deviation by Mathematics Mean, Grade 8, 31 Countries Meeting Sampling Requirements (based on Beaton et al., 1996a)

Figure 1 rebuts the common notion that high-scoring Asian countries have a more equitable (i.e., narrower) dispersion of performance, at least in eighth-grade

mathematics. All three of the Asian countries in our sample have larger standard deviations than does the US: Hong Kong's and Japan's standard deviations are roughly 10% larger than that in the US, and Korea's is approximately 20% larger. Among our sample of seven countries, only France has an unusually small standard deviation of eighth grade mathematics performance, either in absolute terms or relative to its mean.

In grade 8 mathematics, TIMSS also calls into question the view that the US mean is pulled downward by a distribution with an unusually long left-hand (low-scoring) tail. As shown in Figure 2, the US distribution shows a slight right-hand skew rather than a left-hand skew. The US mean is not pulled downward because of a small number of low scoring students. Figure 2 compares the US distribution to the data from Korea. The Korean distribution is substantially wider, as its larger standard deviation indicates. The right-hand tails of the distributions in the two countries are nearly parallel. The left-hand side of the distribution is much shorter in the US, however, pulling the US tail closer to the Korean tail. (Note 2)

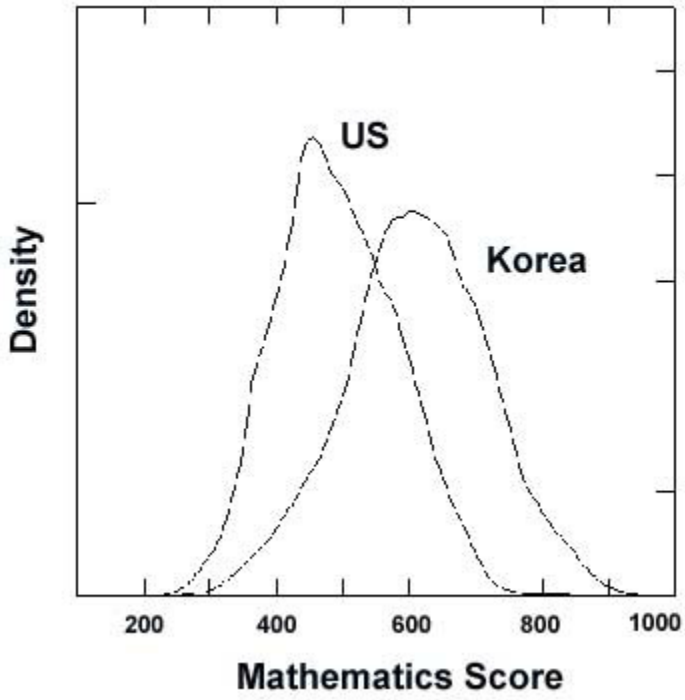

Figure 2. Distributions of Mathematics Scores, Grade 8, Korea and US.
This plot is unweighted. Weighting has virtually no effect on the distribution of scores in Korea and only a trivial effect on the distribution in the US.

Simple Decomposition of Performance Variance in Four Countries

The previous discussion demonstrates that the overall distribution of student level performance in the US is not anomalous. However, looking only at the overall variability might miss important differences between performance in the US population compared to that of other countries. For example, the extent to which the variability is clustered, e.g., within classrooms or schools, might vary across countries. In addition, the possible sources of the variance might also differ across countries, which would suggest different interpretations of the variability of performance and different policy responses to low mean performance in the US. We used data from all seven countries to determine the clustering of variability within and between classrooms. As noted above,

we focus on the classroom rather than the school because the TIMSS sample makes it impossible to distinguish clustering within schools from clustering within classrooms.

The decomposition of mathematics score variance into within- and between-classroom components is sufficient to reveal striking differences among the seven countries in our sample. In the US, Hong Kong, Germany, and Australia, a bit over half of the total variance in eighth-grade mathematics scores lies within classrooms (Table 1). In contrast, in Japan and Korea, over 90 percent of the variance lies within classrooms. France is intermediate, with about three-fourths of the total variance lying within classrooms.

Table 1
Percent of Score Variance Within and Between Classrooms

Country	Percent Between	Percent Within
Australia	47%	53%
France	27	73
Germany	45	55
Hong Kong	46	54
Japan	8	92
Korea	6	94
US	42	58

Similarities among some countries in this decomposition of variance, however, might mask important differences that would be come apparent if TIMSS made it possible to distinguish between-school from between-classroom variance. For example, Schmidt, Wolfe, and Kifer (1993) partitioned the variance of eighth grade mathematics scores in six countries using data from the Second International Mathematics Study, which had two classrooms per school in a number of countries. They found striking differences among countries in the partitioning of aggregate variance. In France, for example, they found that two-thirds of the aggregate variance lay between schools, while in the US, only 9 percent of the aggregate variance lay between schools (with the remainder lying between classrooms within schools).

The average classrooms in our sample of seven countries differ strikingly in their heterogeneity of performance, with the US showing relatively little variability within classrooms. The heterogeneity of performance within classrooms depends on both the total variance of performance in each nation and the breakdown of this variance into within- and between-classroom components. Japan and Korea have slightly larger national standard deviations than the US in Population 2 mathematics and also have a much larger share of their total variance lying within classrooms than does the US. Therefore, the typical within-classroom standard deviation in mathematics is considerably larger in Japan (96) and Korea (102) than in the US (74). (See Table 2.) The average classrooms in France, Germany, Hong Kong, and Australia are more similar to that in the US in heterogeneity.

Table 2
Within-Classroom Standard Deviations

Country	Standard Deviation
Australia	83
France	63
Germany	64
Hong Kong	73
Japan	96
Korea	102
US	74

Multilevel Models of Performance Variation

As noted, we used data from four countries, the US, France, Hong Kong, and Korea, to explore the relationships between performance variation and background variables.

Based on research showing which background characteristics predict student performance in the US, we chose to examine parental education, other measures of socioeconomic status and family composition, measures of academic press in the family and community, and a few measures of student attitudes. We also examined the effect of student age, which could predict performance in at least two ways. Through maturational effects, older students might be expected to perform better than others do. On the other hand, to the extent that students who do poorly in school are held back in grade, older students in a given grade might be expected to perform more poorly than others, particularly in the higher grades. Variations in age at entry could also affect later scores in several ways.

We did not examine curricular variables. As measured, these will not predict variation within classrooms, and research in the US has generally shown variations in schooling to be less powerful predictors of performance than background factors. However, curricular differences may be important predictors of performance variation between classrooms within schools (for example, when students are tracked by ability) and between schools (when schools differ substantially in curriculum). Moreover, important curricular variables are likely to be correlated with background variables. Thus, the results we report here should not be interpreted as clear effects of background variables. Rather, they are likely joint effects of the measured background factors, educational factors collinear with them, and other omitted variables correlated with the measured variables.

Selecting Variables for Inclusion

As noted, exploratory data analysis revealed limitations in some variables that constrained their use in formal models. The few examples presented here illustrate that EDA has particular importance in comparative, international studies because variables may behave differently in different countries.

Although TIMSS includes numerous attitude and press variables, we focused on a set of 15 Likert variables that asked students how strongly they disagreed or agree with statements that the student's mother, the student's friends, and the student herself considered it important to do well in mathematics, do well in the language of the test, do

well in sports, be in a high-achieving class, and have time to have fun. EDA showed these press and attitude variables to be problematic in several respects. In some instances, responses showed little variation. Some relationships with scores were not what one would anticipate if the variables were measuring the intended constructs. In several instances, data showed suggestions of response bias.

For example, several problems can be seen in the responses of eighth-grade students to the BSBMMIP2 press for achievement variable, "My mother thinks it is important for me to do well in mathematics at school" (Figure 3). Each of the six panels arrayed across Figure 3 represents the results from a different country. In the figure we include the four countries in our analysis sample as well as Australia and Germany; this item was not administered in Japan. The common vertical axis, labeled BIMATSCR, is the final TIMSS mathematics score. The four categories of responses to the survey question are arrayed on the X-axis of each panel: SD = strongly disagree, D = disagree, A = agree, and SA = strongly agree. The vertical position of each plotted circle indicates the mean score of the students in that country who gave that particular response to the background question. The radius of each circle is proportional to the percent of students within each country who provided that particular response. The range of sizes is constrained to make the graphic intelligible, however, and in the case of variables with extreme differences in cell counts, including some cells in Figure 3, the relative sizes of the circles understate the actual differences in cell counts.

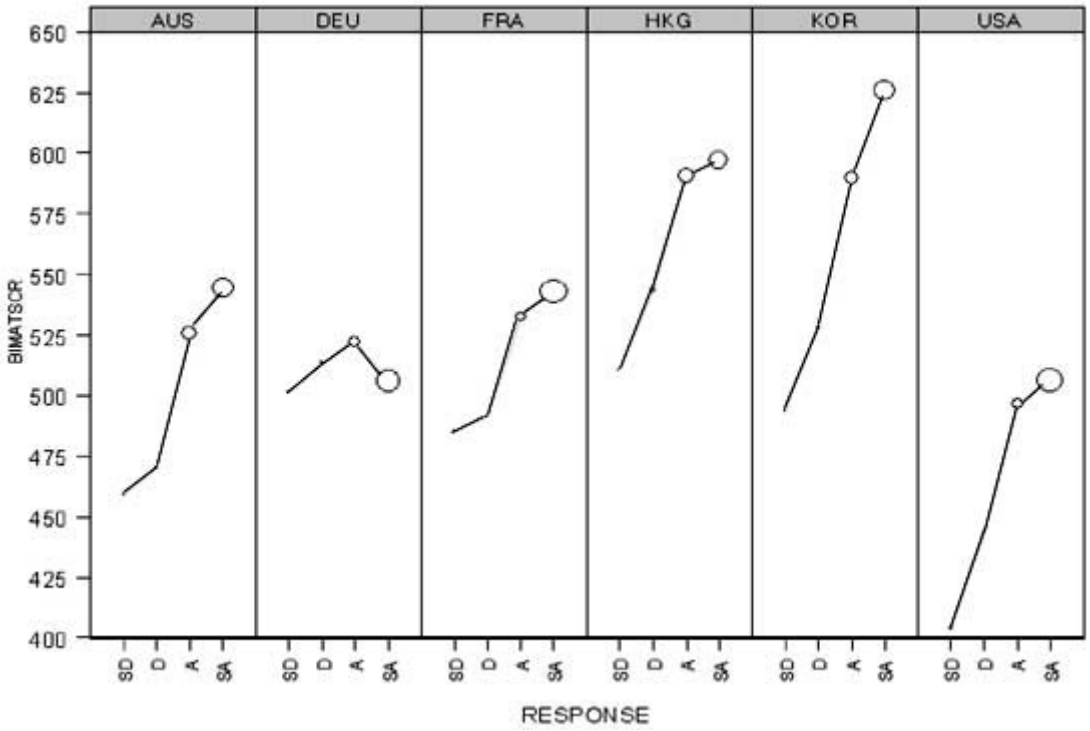

Figure 3. Mathematics Scores and Responses to BSBMMIP2 Press Variable

In all the six countries other than Germany, the relationship between scores and responses to the "My mother thinks it is important for me to do well in mathematics at school" variable was in the anticipated direction: the more strongly students agreed with this statement, the higher their average scores. In most countries, however, this relationship stemmed in large measure from very small groups of students who "disagree" or "strongly disagree" with this statement, and the group that included most students showed only weak relationships. In the US, for example, 97 percent of all students are in the "strongly agree" and "agree" categories, the mean mathematics scores

of which differed by only 10 points. The "disagree" and "strongly disagree" categories had markedly different score means but contained only 2 and 1 percent of students, respectively. This variable is likely to have relatively little utility in predicting score variability in the sampled countries, even if maternal press for achievement is an important influence.

The data from Germany in Figure 3 show an unusual pattern and demonstrate the value of EDA. The relationship between this press variable and scores is not monotonically positive in Germany; the strongly agree and strongly disagree groups had approximately the same mean scores. This pattern, which appeared repeatedly across the TIMSS press and attitude variables in the German data, calls the validity of the responses into question. Because of patterns such as these and the less than optimal sampling in Germany, we did not model the relationships between background variables and scores in Germany. The extremely strong positive relationship in Korea, which also appeared repeatedly, was also grounds for concern. For example, the strong very strong positive relationship appearing in Korea extended to "I think it is important be placed in the high achieving class," even though eighth-grade classes are not tracked by achievement in Korea (Hyung Im, 1998). However, the response patterns in Korea to the variables we used in modeling were not sufficiently suspect in our judgment to warrant excluding Korea from modeling.

The relationships between some other press variables and student performance varied markedly, sometimes dramatically, among countries. These differences among countries could have several causes. There might be response biases, either consistent or item-specific, that vary among countries. Translation problems could engender misleading response differences. There might be substantive reasons for these differences as well; for example, press variables might in fact have stronger relationships with student performance in some countries than in others, perhaps because of differences in the correlations between press variables and school characteristics or between press variables and ethnicity.

TIMSS also includes press variables that one would expect to show weak or even negative relationships with scores. One set, for example, asks students how strongly they agree with the statements that mother, friends, and the student herself think it is important to have time to have fun. One might expect that students who think it particularly important to save time for fun might be less willing to put long hours into study and would therefore score lower. T>wo of the strongest *positive* predictors of mean scores from this set of variables, however, are the strength of agreement with the statements "I think it is important to have time to have fun" (BSBGSIP4) and "My friends think it is important for me to have time to have fun" (BSBGFIP4).

In response to these findings, we used only two of these 15 press variables in our models: the strength with which the student agreed that the mother and the student herself consider it important to do well in mathematics. We pooled these two variables for each subject, creating a single "press for mathematics variable" variable from the students' responses pertaining to themselves and their mothers. These composites were the mean of the two variables for the subject when both were present and whichever was present when one was missing. The decision to pool these two variables, which is consistent with the logic of Likert scales, was made because the two press variables taken individually had only insubstantial relationships with scores, while the composite showed stronger relationships with scores.

We also examined the quality of data for 10 student and family background variables: whether the student was born in the country of testing; mother's and father's educational attainment; number of people in the home; whether the father, mother, and

any grandparents lived with the student; how many books were in the home; and whether the home had a study desk and a computer. Fewer problems appeared with background variables than with press and attitude variables. Missing data and "I don't know" responses, however, posed serious difficulties, particularly in France.

In all of our countries, responses to the questions about parents' educational attainment were missing for a substantial percentage of students. This problem was particularly severe in France (where 17 percent were missing for fathers and 16 percent for mothers). More important, of the students who responded to these question, many answered "I don't know." This problem was particularly severe in France, where 34 percent responded "I don't know," so that a total of 50 percent of respondents provided no informative answer (Table 3). Efforts to impute values were unsuccessful. Thus, we had to choose between omitting parental educational attainment from models in France in order to use most of the sample, or including parental education and using a substantially reduced sample. We opted to include mother's education at the cost of using a reduced sample. Comparisons of preliminary models indicated that the choice between these options probably affected parameter estimates but did not to have a major on the overall prediction of score variance. Although our interpretation focuses on the latter, any interpretation of the results from France should be taken with caution because of this limitation of the data.

Table 3
Percent of Students with No Response or Response of "I don't Know" to Question About Mother's Educational Attainment

	Missing	I don't know	Total
Australia	4%	15%	18%
France	13	34	47
Germany	9	21	30
Hong Kong	5	9	14
Korea	0	9	9
USA	3	7	11

Two variables, mother's educational attainment and number of books in the home, illustrate another issue that can arise in comparative studies – that is, it may be desirable or necessary to treat variables differently in different countries. Both variables showed substantial but not always monotonic positive relationships with achievement. For example, the mean mathematics scores of students whose mothers were in the "finished secondary" and "some vocational" categories were not in the same order in all countries. We combined these categories in all countries except Hong Kong. In Hong Kong, small samples and a different pattern of means suggested collapsing the "some vocational" category of maternal education with "finished university." Similarly, in France only, the mean scores of the students reporting the largest number of books was lower than that of the category below, so we collapsed those two categories into a single category for the French model. We did not collapse these groups for other countries.

The variables used in the final modeling are noted in Appendix A.

Specifying Multilevel Models

The multilevel models reported here are simple "fixed coefficients" models (Kreft and DeLeeuw, 1998). That is, the coefficients estimating the level-one relationships between background factors and achievement (student-level relationships within classrooms) are held constant across classrooms within countries. Between-classroom effects were thus limited to differences in intercepts. In general form, this model is:

$$y_{ij} = \alpha + \beta(x_{ij} - \bar{x}_j) + \gamma \bar{x}_j + \varepsilon_{ij}$$

...where the subscript *i* indicates individuals, *j* indicates classrooms, an underscore indicates a vector, and a bar over a variable indicates a mean. That is, a student's score reflects a vector of background variables weighted by a vector of regression coefficients, a vector of classroom means of those same background characteristics weighted by a second vector of coefficients, and random error. The coefficients applied to individual characteristics are unaffected by classroom characteristics. (That is, there are no cross-level interactions.) Equivalently, this can be expressed in terms of two levels as follows:

$$y_{ij} = \alpha_j + \beta(x_{ij} - \bar{x}_j) + \varepsilon^*_{ij}$$
$$\alpha_j = \alpha + \gamma \bar{x}_j + \eta_j$$
$$\varepsilon_{ij} = \eta_j + \varepsilon^*_{ij}$$

In other words, the intercept in each classroom is the sum of the overall intercept and the sums of the classroom aggregate variables weighted by the classroom-level regression coefficients, plus error. The score of each individual student is then the sum of that student's classroom intercept and the sum of the student-level background variables weighted by the student-level regression coefficients, plus error. Preliminary analysis indicated that little would be gained by allowing the within-classroom slopes to vary randomly or by modeling their variation.

These models center observations around classroom means. Without group-mean centering, the predictor variance within and between classrooms would be confounded. Centering eliminates confounding of the predictor variance between and within classrooms. Centering also makes the model's coefficients straightforward estimates of the within-classroom and between-classroom effects (e.g., Bryk and Raudenbush, 1992).

We began with the assumption that all variables that survived screening by EDA would be included in the models. Including some that survived the EDA, however, resulted in numerous small and statistically non-significant parameter estimates. We therefore constructed models based on what could be called a ‘judgmental stepwise' procedure, in which we began with a null model (i.e., a model including nothing but an intercept), built up to a more complex model, and then pared back to a more parsimonious model based on the size and significance of coefficients.(Note 3) In general, we opted to include variables that were only marginally significant or that failed to reach significance by a modest amount, leaving it to the reader to discount them, provided that their inclusion did not markedly change the coefficients of other variables. In addition, because our classroom-level variables are aggregates of student-level

variables, we included at both levels any variable that was significant at either level.

The statistics normally reported from hierarchical models—intercepts and regression coefficients at each level of aggregation—are sufficient for predicting means but not for comparing variance of performance across countries. For example, at the classroom level, the estimated effect of the proportion of students living with their fathers indicates how much, on average, the classroom mean score would increase if the proportion increased from 0 to 1, but it does not indicate how much of the variability among classroom mean scores is attributable to this factor. Therefore, we also present a summary of the variance accounted for by the predictors at each level, expressed as the absolute value of the predicted variance, the percentage of variance predicted within level, and the percentage of total variance predicted.

Decomposing Performance Variation

We first give a detailed discussion of the models for the US. This discussion serves as a template for evaluating the results of the other models. We then compare the results from the four countries.

The final two-level model of mathematics scores in the US contained only five variables at each level: the number of books in the home, the presence of a computer in the home, the presence of the father in the home, the academic press variable, and student age. The square of age was included because of nonlinearities in the relationships between age and scores that became apparent in the exploratory data analysis. Each of these variables was at least marginally significant at one of the two levels.

The importance of these predictors can be evaluated several ways. One can look at the significance and impact of the individual coefficients within each level, the relative significance or impact of the coefficients across levels, and the total predictive power of the coefficients at each level. These three views are each described in turn.

Within classrooms in the US, the strongest effects were those of the number of books, the academic press variable, and students' age (Table 4). The effects of having a computer and the father living at home were both smaller and non-significant. Comparisons of these parameter estimates, however, is clouded by their imprecision. Confidence bands around most of these estimates were wide (see Appendix B).

Table 4
Two-Level Models of Mathematics Scores

Variable	United States	France	Hong Kong	Korea
Intercept	-351.7	592.6	-424.8	27.9
Within class (b)				
Number books	7.9**		0.3	20.2**
Computer present	4.4		-3.8	10.9**
Father present	1.7	8.9*	-7.4	
Mother's education		4.6		
Father's education				9.5**

Press	9.6**	8.6*	10.3**	36.2**
Age	-14.4**	-18.2**		-6.0
Age^2	-6.9	-0.6		-14.8**
Born in Country			-19.1**	
Between-class (c)				
M Books	45.5**		44.1**	16.2*
M Computer	37.2*		89.8*	44.5**
M Father present	90.3**	59.5**	326.9**	
M Mother's education		26.4**		
M Father's Education				18.8**
M Press	43.2**	45.0**	174.5**	47.4**
M Age	33.9	-23.0*		20.5
M Age^2	-149.4	-23.3		-26.2*
M Born in Country			-44.7	
Residual variances				
r^2 (within)	4570.4	4040.8	5485.0	9290.6
t (between)	766.2	554.7	1406.2	48.0

NOTE. All estimates of significance reflect jackknifed estimates:
* p<.05 ** P<.01

The effects of these estimates can be compared to the distribution of scores to provide a concrete estimate of their size. For example, in the US, the estimated student-level effect of the number of books was 7.9. This variable had five categories. The model predicts that holding constant the other variables, the mean difference between students in the lowest and highest categories would be 32 points, roughly one-third of the standard deviation of mathematics scores, which was 89 points in this subsample. The press coefficient was larger, but most students were concentrated within two categories of either of the press variables, and the effect of being in the higher of these two categories, relative to the lower of them, was only about one-tenth of a standard deviation. The age coefficient was significant and negative, suggesting that either retention or late entry of slower learners have a larger impact than maturational effects.

At first glance, the estimated effects at the between-classroom level (preceded by an "M," for "mean," in all tables) appear much larger than the coefficients at the within-classroom level. However, the standard errors of the estimated between-class coefficients are generally large, and the t statistics of the between-class coefficients are on average only modestly larger than those of the within-class estimates.

Nonetheless, in the US, there are some striking differences between the within- and between-class estimates. The presence of the father in the home had a non-significant and near-zero relationship to scores within classrooms, but the percentage of fathers in the home showed a substantial relationship to classroom mean scores. On average, the estimated within-classroom effect of having the father present

was less than 2 points, roughly 2 percent of a standard deviation. Classrooms in our grade 8 mathematics model sample ranged from 15 to 100 percent of fathers present. Holding other variables constant, going from one standard deviation below the mean to one standard deviation above on the scale of proportion of fathers present (from .50 to .82) would predict an increase in mean scores of about one-third of a standard deviation.

The difference in predictive power at the within- and between-classroom levels in the US becomes clearer if one compares the variance accounted for by variables at each level. In this model, 59 percent of the total variance in scores in the US was within classrooms, while the remaining 41 percent was between classrooms (Table 5). The five variables in the model predicted about 77 percent of the between-classroom variance but only 4 percent of the within-classroom variance. The predicted between-classroom variance was 2,532, while the predicted within-classroom variance was only 198. Thus, the five between-classroom variables accounted for 31 percent of the total variance of mathematics scores [2532/(3299+4769)], while the five within-classroom variables accounted for only 2 percent of the total variance.

Table 5
Total and Predicted Variance in Mathematics Scores at Each Level

	United States		France		Hong Kong		Korea	
Share of Variance	Between	Within	Between	Within	Between	Within	Between	Within
Total at level	3299	4769	1356	4232	4543	5557	799	10722
Percent at level	41	59	24	76	45	55	7	93
Predicted by variables at level	2532	198	801	191	3137	73	751	1431
Percent at level predicted by variables at level	77	4	59	5	69	1	94	13
Percent of total predicted by variables at level	31	2	19	3	31	1	7	12

One surprising finding in the multilevel model for the US was the lack of importance of mother's and father's education, which are generally considered to be among the strongest predictors of student performance in the US. Parental education did not have large enough effects to warrant keeping either variable in the model. Alternative models (for example, one in which the TIMSS parental education categories were entered as dummies) produced the same result. To explore this, we conducted additional analyses of TIMSS and the base year of the National Education Longitudinal Study (NELS-88), modifying our TIMSS model in several ways to make it as

comparable as possible to the model we analyzed in NELS. This comparison suggested that several factors contributed to the unimportance of maternal education in our TIMSS model, including the use of a single classroom per school and the inclusion of the academic press variable. However, much of the difference remained unexplained and appears to be a result of unknown characteristics of the TIMSS database. When nearly identical models were analyzed in TIMSS and NELS, in both cases using schools rather than classrooms as the level 2 unit, the level 1 and level 2 parameters for maternal education were both less than half the size in TIMSS as in NELS.

None of the final models fully matched any other in terms of the variables included (Table 4). Only a single variable, academic press, appeared in the final models for all countries. The final models for Hong Kong, Korea, and the US all included variables for number books in the home, a computer in the home, and academic press. The model for Hong Kong, however, included a variable for father present in the home but excluded age, which was included in both the US and Korea. The model for Korea included age but excluded presence of a father, which was included in the other two countries. The model for Hong Kong included a variable for born in country, and the model for Korea included a variable for father's education; neither of these variables was included in the models for any other countries. The model for France was was the only model that excluded variables for the number of books or computer present and was the only one to include mother's education.

Although some of the coefficients were similar in magnitude across countries, others differed markedly. For example, the student-level (within-classroom) coefficients for press were similar in the US, France and Hong Kong: 9.6, 8.6 and 10.3, respectively. The between-classroom coefficients for this variable were 43.2, 45.0 and 47.4 for the US, France and Korea. In contrast, the between-classroom coefficient for the press variable in Hong Kong was 174.5, several times as large as the coefficients for the same variable in the other models. However, as explained below, we do not place great confidence on specific parameter estimates, and this estimate in Hong Kong may be seen as implausible.

Although the variables in the models and the effects of those variables differed across countries, the models in all countries were consistent in predicting most of the variance between classrooms but little of the variance within classrooms (Table 5). This prediction of between-classroom variance ranged from 59 percent in France to 94 percent in the Korea, and the prediction of within-classroom variance ranged from 1 percent in Hong Kong to 13 percent in Korea. The prediction of within-classroom variance in Korea, while a modest 13 percent, is several times as strong as in any other country; the next strongest prediction was 5 percent of the within-classroom variance in France.

The consistency of this strong prediction of between-classroom variance is all the more striking in the light of the sparseness of the models and the weak measurement of social background. Our models included few predictors. The variables available in TIMSS do not necessarily include those that researchers in participating countries would suggest are the most important predictors of achievement. For example, TIMSS does not include income, race/ethnicity, or inner-city location, all three of which are known to be important predictors of performance in the US. Similarly, the National Research Coordinator for Korea indicated that income, type of community (urban, suburban, rural) and geographic region are all somewhat correlated with performance in Korea (Im, 1998). In addition, the selection of variables for use in the models was constrained in some instances by problems with the data.

Thus, the variables included in the models were a potentially weak proxy for those

that would best show the relationships between score variance and background variables in each country. It is possible that the use of a stronger set of predictors would have substantially increased the percentage of variance predicted at one or both levels, particularly the within-classroom level, at which our prediction was very weak. We cannot determine whether this is the case, however. In the general case, the degree of prediction may not be substantially lessened by the weakness of collinear predictors if enough of them are used in the model (e.g., Berends and Koretz, 1996).

We have less confidence in the specific parameter estimates we obtained, particularly in cases in which the estimates varied markedly among countries. There are several reasons for this caution. First, as noted earlier, parameter estimates in multi-level models are often quite sensitive to specification differences (Kreft and DeLeeuw, 1998), and our selections of variables were necessarily somewhat happenstance, constrained as they were by the limitations of the TIMSS database. Models that included additional variables (such as family income) or better-measured constructs might have yielded substantially different estimates of the parameters in our models. Second, EDA showed that some variables behaved quite differently across countries. Other operationalizations of these constructs might have altered these differences and might therefore have produced different parameter estimates.

To test the importance of the particular selections of variables in our final models, we ran a constant, minimal model in each of the four countries, including the individual and aggregate values of number of books, computer present, press, age, and age squared. This fixed model predicted almost as much of the variance in performance as did our final models, which were selected to optimize prediction in each country and subject (Table 6; compare Table 5). This suggests that predicted variability is somewhat invariant to the variables included in the model.

Table 6
Percent of Variance at Each Level Predicted by Fixed Model

	Mathematics	
	Between Classroom	Within Classroom
United States	72%	4%
France	54	4
Hong Kong	67	1
Korea	86	12

Differences in the strength of prediction across the four countries therefore may be substantively more important than differences in parameter estimates. One striking difference in prediction becomes apparent when one looks at the prediction of total variance rather than within-level variance. In the US and Hong Kong, roughly one third of the total variance is predicted by the models, in both cases largely because of variation in between-classroom predictors (Table 7). The models predict much less of the variance in France (18 percent) and Korea (19 percent).

Table 7
Percent of Total Variance Predicted by Predictors at Each Level, Final Models

	Between Classroom	Within Classroom	Both Levels
United States	31%	2%	34%
France	14	3	18
Hong Kong	31	1	32
Korea	7	12	19

NOTE: Entries may not sum to totals because of rounding.

The four countries also differ in terms of the relative predictive power of the models between the student and classroom levels. Again, the US and Hong Kong are very similar: almost all of the predicted variance in each country is attributable to between-classroom variation in the predictors (Table 7). France and Korea, however, differ in this respect, even though the percentage of total variance predicted at both levels is nearly identical in the two countries. In France, most of the predicted variance is attributable to the classroom-level predictors, and France differs from the US and Hong Kong in that the prediction is much weaker at the classroom level. In Korea, in contrast to all three other countries, more of the total prediction is due to within-classroom variation in predictors. This can be seen as a reflection of two factors. First, even though the model predicted only a modest percentage of the within-classroom variance in Korea, the predicted percentage was considerably larger than in the other three countries. Second, a larger percentage of the total variance lies within classrooms in Korea (93 percent) than in France (76 percent), the US (59 percent), or Hong Kong (55 percent). The product of these two percentages, which is the percent of total variance predicted by within-classroom predictors, is therefore much larger in Korea than in the other countries.

There are several possible non-exclusive explanations for these cross-national differences in predicted variance. First, the fixed model and our final models may be a better selection of variables for some countries than for others. Changing to a fixed set of variables drawing from the variables in our set did not have much of an impact, but it is possible that including other variables would have. Second, taking our models as a given, stronger prediction in one country than in another could stem from larger estimated effects of some variables in the model, greater variability in the predictors themselves, or both.

Stronger prediction of scores could reflect stronger partial relationships, greater variance in the predictors themselves, or both. To explore this, we partitioned the variance in the predictors themselves into within- and between classroom components. We then compared the amount of variance in the predictors to the amount of predicted variance in scores.

The greater prediction of score variance within classrooms in Korea compared to the US appears not to stem from differences in the variability of predictors. Within classrooms, all of the predictors other than age (which matters less because it is a weak predictor of scores) showed roughly similar variance in the US and Korea. This, in conjunction with the larger parameter estimates reported for Korea earlier, indicate that the stronger within-classroom prediction in Korea stems from stronger partial relationships within classrooms between background variables and scores.

The contribution of predictor variance to the difference between France and the US in the prediction of between-classroom score variance, however, is ambiguous. France shows less between-classroom variance in two predictors, number of books and

computer present, and the former is a relatively powerful predictor of score variance in France. On the other hand, France shows much more between-classroom variance in age, and age is also a strong predictor of score variance.

Recall that although Hong Kong is similar to Japan and Korea in terms of its overall mean and standard deviation, it is similar to the US – and strikingly different from Japan and Korea – in terms of the decomposition of variance into within- and between-school components. Hong Kong is also very similar to the US in terms of the predictive power of the models both within and between classrooms. Hong Kong and the US are also similar in terms of the within- and between-classroom variance of the predictors themselves, with the exception of age.

Conclusions

This study was prompted in part by a widespread view that performance variance in the US is unusual. This view has sometimes been made explicit – for example, in Berliner and Biddle's assertion that "The achievement of American schools is a *lot* more variable than is student achievement from elsewhere" (1995, p. 58). In other instances, this view of variability is implicit, as when the scores for US states or districts are compared to national averages from other countries. In response, we asked whether the distribution of performance in the US is anomalous, how the variance in performance is distributed in the US and other countries, and how well background factors can predict that variation.

TIMSS suggests strongly that the variation in performance in the US is not anomalous. In Population 2, the US variance is large but not exceptional in science and more nearly average in mathematics. Contrary to some expectations, the distribution of scores is not particularly skewed in the US, and in eighth-grade mathematics, it is right- rather than left-skewed. Moreover, differences among countries in the variance of performance do not clearly follow stereotypes about their homogeneity. Socially homogeneous Japan, for example, shows a bit more variation than the US in mathematics, while socially heterogeneous France shows considerably less.

When performance variance is broken into within- and between-classroom components, however, the story becomes more complex. The US, Australia, Germany and Hong Kong show one pattern, in which nearly half of the variance lies between classrooms. Japan and Korea lie at the other extreme; most of their variance lies within classrooms, while very little lies between. The result is that classrooms in Japan and Korea resemble each other in terms of mean performance much more than do classrooms in the US, Germany, Hong Kong, and Australia. France falls between these two poles. By the same token, students in the typical classrooms in Japan and Korea show much greater variability in performance than do their counterparts in the US, Germany, Hong Kong, and Australia.

While the US is similar to many other countries in the overall variability of student performance in mathematics and is similar to several others we investigated in the decomposition of performance variation within and between classrooms, TIMSS does not fully address the reasonableness of Berliner and Biddle's (1995) assertion that US schools are far more variable than are schools elsewhere. Of the countries we considered, only the US and Australia provided samples that allow one to separate between-classroom and between-school variance. For example, if tracking is entirely absent in Japan and Korea, classrooms within schools should be randomly equivalent. In this case, much of the between-classroom variance in these countries might lie between schools – in comparison to the US and Australia, where our preliminary analysis found

that most of the between-classroom variance lies within schools. However, only a sample that includes multiple classrooms per school would permit testing this hypothesis.

What do the present findings imply about the reasonableness of comparing means for US states and districts to averages for other nations? We cannot fully answer that question because the TIMSS design does not yield evidence pertaining to districts or states in the US or about similar units in other countries, such as German Länder. However, the wide dispersion of classroom means in Australia and Germany, and the smaller but still substantial dispersion of means in France, suggests that these comparisons may be misleading. Just as some states in the US compare more favorably than do others to means of other countries, some areas in those other countries are likely to score markedly better than the averages for those countries. In contrast, classrooms in Japan and Korea vary much less in average performance, so comparisons between US states and the means in Japan and Korea may be more meaningful. However, even in Korea and Japan, the standard deviations of classroom means are substantial, and the standard deviation of school means, which cannot be estimated from TIMSS, may be sizable as well.

Our analyses cannot identify causes of the cross-national differences we found, but they raise a number of intriguing possibilities that warrant further investigation. One question is what factors might underlie the patterns in Korea: little total variance between classrooms and an unusually large amount of predicted variance within classrooms.

One possible contributor to the differences between the US and Korea is stratification of students in terms of ability. This hypothesis is consistent with the differences between the US and Korea in terms of both the decomposition of variance and the ability of the models to predict the within-classroom variance. We know that Korea's policy is not to track students into classes by ability in eighth-grade mathematics (Im, 1998). If schools as well as classrooms are relatively little stratified in Korea in terms of background factors associated with student performance, then more of the relevant variance of these background variables may lie within classrooms in Korea than in France, the US, or Hong Kong. Note that the total variance in the background factors included in the fixed model is not larger within classrooms in Korea than in the US. However, more of the variance that predicts student performance may lie within classrooms in Korea. In contrast, in countries like the US, the combination of residential stratification and tracking would result in much of the relevant variance of these background variables lying between classrooms rather than within them.

However, other factors, such as instructional differences, might also contribute to the differences between Korea and the other countries examined. For example, instruction might vary less among classrooms in Korea than in Hong Kong or the US. This might help explain the lack of performance variation between classrooms. Instructional factors might also contribute to the greater within-classroom predictive power of background factors in Korea. Although many current US reform efforts aim for both higher standards and greater equity of outcomes, it is possible that all other factors being equal, a very high level of standards could increase score variance, as the more able students might be better able to take advantage of more difficult material. Curriculum differences might also correlate differently with background factors from one country to another. If curriculum differences are less highly correlated with background factors in Korea than in the US, that too could contribute to the patterns we found.

The results for Hong Kong also raise interesting questions. Four Asian countries,

Singapore, Korea, Japan, and Hong Kong, ranked highest in grade 8 mathematics in TIMSS. Hong Kong is also similar to Japan and Korea, but not Singapore, in terms of its simple standard deviation of scores. Our results, however, showed that in both the decomposition and prediction of performance variation, Hong Kong is very similar to the US and strikingly different from Korea and Japan. Hong Kong is also similar to the US in terms of the decomposition of the variance of predictor variables. Further investigation of factors that might cause Hong Kong to resemble other highly developed Asian countries in some respects but the US in other respects could help avoid simplistic explanations of cross-national differences in performance.

Finally, several aspects of performance variation in France – the relatively small overall standard deviation of scores, and the small total and predicted between-classroom variance – could have important implications for policy. As noted earlier, it is not clear from our results whether lesser between-classroom variation in predictors contributed to this, but decompositions of predictor variance did not suggest that this was a major factor. Some observers maintain that the French curriculum is highly standardized, even compared to that of many other countries with national curricula. If so, that uniformity could contribute to both a smaller between-classroom variance. In addition, by weakening any correlations between curricular variables and social background, uniformity of curriculum could also lessen the prediction of score variance by background factors.

Further analysis of TIMSS data may help shed light on these questions. For example, the present analysis could be expanded to incorporate instructional and curriculum variables as well as background factors. The TIMSS data, however, will not be sufficient to address certain key aspects of these questions. They cannot provide useful data about variations in larger aggregates, including schools and states (and their equivalents). Moreover, in most countries, TIMSS collected very little information about stratification, either within or between schools. These gaps could be addressed either by modifications of future international surveys or by the use of smaller, more focused studies in selected countries.

Notes

1. A number of studies have shown that even older students often provide reports of background variables that are inconsistent with those of their parents. For example, Kaufman and Rasinski (1991) showed that only roughly 60 percent of eighth-grade students in the National Education Longitudinal Study (NELS-88) agreed with their parents about their parents' educational attainment (Kaufman and Rasinski, 1991, Table 3.2). A study of Asian and Hispanic students in NAEP found similar results for middle-school students but found that fewer than half of third-grade students agreed with their parents on this variable (Baratz-Snowden, Pollack, and Rock, 1988).
2. Note that the shape of the distributions depend on the mix of items included in the assessment. For example, it is possible that including a larger number of easy items in the assessment would have stretched the left-hand tails of these distributions, particularly the lower tail of the US distribution.
3. This is in contrast to traditional stepwise or other empirical subsets procedures, in which criteria specified *a priori*, such as F-for-inclusion, are applied algorithmically.

Acknowledgements

This work was conducted under Task Order 1.2.77.1 with the Education Statistics Services Institute, funded by contract number RN95127001 from the National Center for Education Statistics. The opinions expressed here are solely those of the authors and do not necessarily represent the views of the Educational Statistics Services Institute or the National Center for Education Statistics.

The authors would like to acknowledge the assistance of several people who contributed to this work. Eugene Gonzalez of Boston College and TIMSS explained numerous aspects of the TIMSS data. Al Beaton and Laura O'Dwyer of Boston College and Laura Salganik of the Educational Statistics Services Institute reviewed this report and provided valuable comments. Any errors of fact or interpretation that remain are solely the responsibility of the authors. Christel Osborn provided secretarial support for the project.

References

Baratz-Snowden, J., J. Pollack, and D. Rock (1988). *Quality of responses of selected items on NAEP special study student survey*. Princeton: Educational Testing Service, unpublished.

Beaton, A. E., Mullis, I. V. S., Martin, M. O,. Gonzalez, E. J., Kelly, D. L., and Smith, T. A. (1996a). *Mathematics Achievement in the Middle School Years.* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Beaton, A. E., Martin, M. O, Mullis, I. V. S., Gonzalez, E. J., Smith, T. A, and Kelly, D. L. (1996b). *Science Achievement in the Middle School Years.* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Berends, M., and Koretz, D. (1996). Reporting minority students' test scores: How well can the National Assessment of Educational Progress account for differences in social context? *Educational Assessment, 3*(3), 249-285.

Berliner, D. C., and Biddle, B. J. (1995). *The Manufactured Crisis: Myths, Fraud, and the Attach on America's Public Schools.* Reading, MA: Addison-Wesley.

Bryk, A. S., and Raudenbush, S. W. (1992). *Hierarchical Linear Models*. Newbury Park, CA: Sage.

Bryk, A. S., Raudenbush, S. W., and Congdon, R. T. (1996). *HLM: Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L Programs.* Chicago: Scientific Software International.

Foy, P. Rust, K., and Schleicher, A. (1996). Sample design. In M. O. Martin and D. L. Kelly (Eds.), *Third International Mathematics and Science Study (TIMSS) Volume I: Design and Development*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Im, Hyung (1998). Personal communication, June 18.

Kaufman, P., and Rasinski, K. A. (1991). *Quality of Responses of Eighth-Grade Student in NELS-88.* Washington, DC: US Department of Education, Office of Educational Research and Improvement (NCES 91-487).

Koretz, D., McCaffrey, D., and Sullivan, T. (2000). *Using TIMSS to Analyze Correlates of Performance Variation in Mathematics.* Santa Monica: RAND, working paper (February).

Kreft, I., and DeLeeuw, J. (1998). *Introducing Multilevel Modeling*. London: Sage.

Martin, M. O., Mullis, I. V. S., Beaton, A. E., Gonzalez, E. J., Smith, T. A., and Kelly, D. L. (1997). *Science Achievement in the Primary School Years: IEA's Third International Science and Science Study* . Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., and Smith, T. A. (1997). *Mathematics Achievement in the Primary School Years: IEA's Third International Science and Science Study* . Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., and Smith, T. A. (1998). *Mathematics and Science in the Final Year of Secondary School: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

National Center for Education Statistics (1996). *Education in States and Nations: Indicators Comparing U.S. States with Other Industrialized Countries in 1991.* Washington: author (Report NCES 96-160).

Pfeffermann D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, **5**, 239-262.

Pfefferman, D., Skinner, C. J., Holmes, D .J., Goldstein, H., and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, B, 60* Part 1, 23-40.

Schmidt, W. H., Wolfe, R. G., and Kifer, E. (1993). 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.* Oxford: Pergamon, 59-100.

About the Authors

Daniel Koretz is a professor at the Harvard Graduate School of Education and Associate Director of the Center for Research on Evaluation, Standards, and Student Testing (CRESST). His work focuses primarily on educational assessment and recently has included studies of the validity of gains in high-stakes testing programs, the effects of testing programs on schooling, the assessment of students with disabilities, and the effects of alternative systems of college admissions.
E-mail: daniel_koretz@harvard.edu

Daniel McCaffrey is a Statistician at RAND. His areas of concentration are education policy and the analysis of hierarchical data and data from complex sample designs.
E-mail: Daniel_McCaffrey@rand.org

Thomas J. Sullivan is a statistical programmer/analyst with RAND and a doctoral

candidate in Computational Statistics at George Mason University.
E-mail: toms@rand.org

Appendix A
Description of Variables

This Appendix describes the source of the principal variables used the models presented in this report.

Name	TIMSS name	Notes
Math score	BIMATSCR	
Father present	BSBGADU2	
Age	BSDAGE	
Books in home	BSBGBOOK	Sometimes entered as a single variable, if test of linearity warranted.
Computer in home	BSBGPS02	
Press	composite	Mean of BSBMSIP2 and BSBMMIP2 when both were present; either variable if only one present
Mother's education	BSBGEDUM	Sometimes recoded as noted in text; sometimes entered as a single variable, if test of linearity warranted
Father's education	BSBGEDUF	Sometimes recoded as noted in text; sometimes entered as a single variable, if test of linearity warranted
Born in country	BSBGBRN1	

Appendix B
Confidence Limits for Parameter Estimates
Two-level Models

Parameter estimates are the same as those reported in the body of the article. Jackknifed estimates of lower and upper 95 percent confidence limits are in parentheses under each parameter estimate.

Variable	United States	France	Hong Kong	Korea
Intercept	-351.7	592.6	-424.8	27.9
	(-884.9, 181.5)	(289.8, 895.4)	(-708.6, -141.0)	(-660.5, 716.3)
Within class (b)				
Number books	7.9**		0.3	20.2**
	(5.6, 10.3)		(-2.0, 2.7)	(16.7, 23.7)
Computer present	4.4		-3.8	10.9**
	(-2.7, 11.4)		(-10.0, 2.5)	(3.2, 18.7)
Father present	1.7	8.9*	-7.4	

	(-4.9, 8.3)	(0.9, 17.0)	(-19.7, 5.0)	
Mother's education		4.6		
		(1.2, 8.0)		
Father's education				9.5**
				(5.5, 13.5)
Press	9.6**	8.6*	10.3**	36.2**
	(4.4, 14.7)	(0.5, 16.7)	(4.6, 16.0)	(27.9, 44.5)
Age	-14.4**	-18.2**		-6.0
	(-21.1, -7.7)	(-24.6, 11.7)		(-18.4, 6.4)
Age^2	-6.9	-0.6		-14.8**
	(-14.2, 0.5)	(-6.0, 4.7)		(-26.0, -3.7)
Born in Country			-19.1**	
			(-30.1, -8.2)	
Between-class (c)				
M Books	45.5**		44.1**	16.2*
	(30.7, 60.2)		(11.7, 76.6)	(1.7, 30.7)
M Computer	37.2*		89.8*	44.5**
	(3.6, 70.9)		(5.4, 174.1)	(12.5, 76.4)
M Father present	90.3**	59.5**	326.9**	
	(47.4, 133.2)	(14.0, 104.9)	(151.8, 502.1)	
M Mother's education		26.4**		
		(16.2, 36.7)		
M Father's Education				18.8**
				(7.7, 30.0)
M Press	43.2**	45.0**	174.5**	47.4**
	(9.0, 77.4)	(14.3, 75.5)	(103.5, 245.4)	(13.1, 81.7)
M Age	33.9	-23.0*		20.5
	(3.2, 64.6)	(-43.1, -2.8)		(-27.6, 68.5)
M Age^2	-149.4	-23.3		-26.2*
	(-223.8, -75.0)	(-54.7, 8.0)		(-50.0, -2.4)
M Born in Country			-44.7	
			(-119.9, 30.4)	
Residual variances				
r^2 (within)	4570.4	4040.8	5485.0	9290.6

t (between)	766.2	554.7	1406.2	48.0

The World Wide Web address for the *Education Policy Analysis Archives* is **epaa.asu.edu**

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-0211. (602-965-9644). The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

EPAA Editorial Board

EPAA Spanish Language Editorial Board