NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

The Working Paper Series was initiated to promote the sharing of the valuable work experience and knowledge reflected in these preliminary reports. These reports are viewed as works in progress, and have not undergone a rigorous review for consistency with NCES Statistical Standards prior to inclusion in the Working Paper Series.

U.S. Department of Education
Office of Educational Research and Improvement

NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

Using TIMSS to Analyze Correlates of Performance Variation in Mathematics

Working Paper No. 2001-05   June 2001

Contact: Patrick Gonzales
                Early Childhood, International and Crosscutting
                Studies Division
                Tel: 202-502-7346
                e-mail: patrick.gonzales@ed.gov

U.S. Department of Education
Rod Paige
Secretary

National Center for Education Statistics
Gary W. Phillips
Acting Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

> National Center for Education Statistics
> Office of Educational Research and Improvement
> U.S. Department of Education
> 1990 K St., NW
> Washington, DC 20006

June 2001

The NCES World Wide Web Home Page is http://nces.ed.gov

Suggested Citation

U.S. Department of Education, National Center for Education Statistics. *Using TIMSS to Analyze Correlates of Performance Variation in Mathematics.* Working Paper No. 2001-05, by Daniel Koretz, Daniel McCaffrey and Thomas Sullivan. Project Officer, Patrick Gonzales. Washington, DC: 2001.

FOREWORD

In addition to official NCES publications, NCES staff and individuals commissioned by NCES produce preliminary research reports that include analyses of survey results, and presentations of technical, methodological, and statistical evaluation issues.

The *Working Paper Series* was initiated to promote the sharing of the valuable work experience and knowledge reflected in these preliminary reports. These reports are viewed as works in progress, and have not undergone a rigorous review for consistency with NCES Statistical Standards prior to inclusion in the Working Paper Series.

Copies of Working Papers can be downloaded as pdf files from the NCES Electronic Catalog (http://nces.ed.gov/pubsearch), or contact Sheilah Jupiter at (202) 502-7444, e-mail: sheilah.jupiter@ed.gov, or mail: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 1990 K St., NW, Room 9048, Washington, DC 20006.

Marilyn M. McMillen	Ralph Lee
Chief Mathematical Statistician	Mathematical Statistician
Statistical Standards Program	Statistical Standards Program

This page intentionally left blank.

Using TIMSS to Analyze
Correlates of Performance Variation in Mathematics

Prepared by:

Daniel Koretz

RAND Education and
Center for Research on Evaluation, Standards, and Student Testing (CRESST)

Daniel McCaffrey
Thomas Sullivan

RAND Education

PREFACE

The National Center for Education Statistics (NCES), through the Education Statistics Services Institute, supported the research in this report to promote further exploration of the 1995 Third International Mathematics and Science Study (TIMSS) data. Specifically, NCES commissioned this work to investigate classroom-level effects associated with differences in assessment performance in TIMSS. Some of the issues concerning the sample design of TIMSS pointed out by the authors were known beforehand; others are brought to light through the analyses documented in this paper.

In 1999, the International Association for the Evaluation of Educational Achievement (IEA) conducted a repeat of TIMSS (known as TIMSS-R), using the same methods, frameworks, and documentation procedures as in the 1995 TIMSS. Some changes were made to better document data collection and quality issues. However, TIMSS-R was essentially a repeat of TIMSS. Thus, many of the recommendations identified by the authors of this paper could not be taken into consideration in the design of TIMSS-R. For example, the TIMSS-R international guidelines continued to specify a minimum of one intact classroom per sampled school despite the possible advantages of sampling a minimum of two intact classrooms per school as recommended by the authors.

Research such as the work presented here will, nonetheless, continue to assist NCES in improving its own studies and in making data-based recommendations to our partners around the world in the best methods for collecting comparable data in future studies.

Finally, 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 Education Statistics Services Institute, the National Center for Education Statistics, or the U.S. Department of Education.

Eugene Owen
Program Director
International Activities Program

Valena Plisko
Associate Commissioner
Early Childhood, International and
Crosscutting Studies Division

TABLE OF CONTENTS

List of Figures

List of Tables

EXECUTIVE SUMMARY

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. and several other countries, investigate how this performance variation is distributed within and between classrooms, and explore how well background variables predict performance at both levels. In addition, the study explored how well suited the TIMSS design is to addressing questions of this sort.

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.

TIMSS reports show that in both grades 4 and 8, U.S. performance variability was near the median of a large sample of nations in mathematics and near the top (but not an outlier) in science (Beaton, et al., 1996a, 1996b; Martin, et al., 1997; Mullis, et al., 1997). Moreover, the U.S. distribution of scores is not left-skewed. For example, in grade 8 mathematics, U.S. scores show a modest right-hand skew and are less variable than those of Korea (Summary Figure 1).

Summary Figure 1.—Distributions of Mathematics Scores, Grade 8, Korea and U.S.: 1995
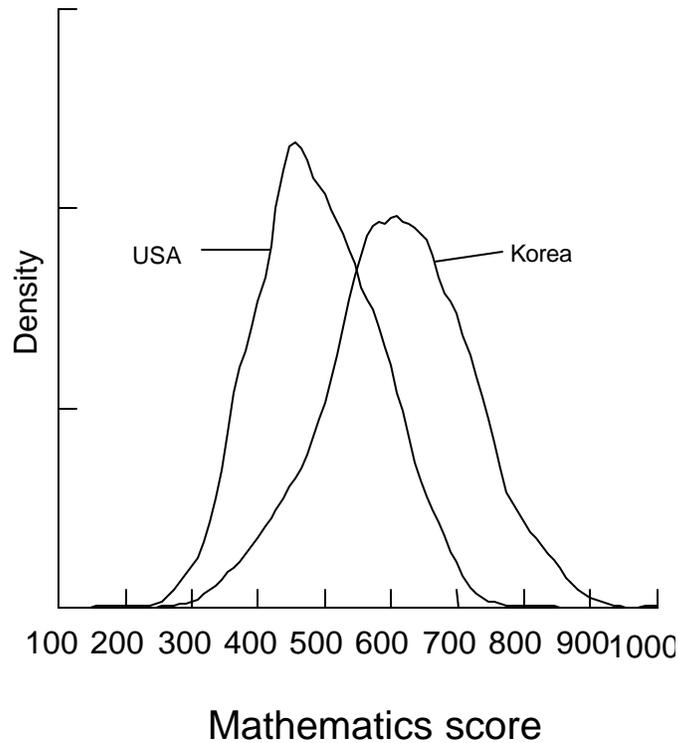
Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

Our analyses focused primarily on mathematics in the upper grade of Population 2 (grade 8) in seven countries: the United States, Australia, France, Germany, Hong Kong, Korea, and Japan. Of these, only France (standard deviation = 76) showed appreciably less variability of performance than the U.S. (standard deviation = 91). Germany was comparable to the U.S. in variability, and the other countries all had more variability of performance, ranging up to Korea's standard deviation of 109.

Although all of our countries other than France had reasonably similar score variance in eighth grade mathematics, the distribution of that variance within and between classrooms differed markedly. In the U.S., Germany, Hong Kong, and Australia, between 42 and 47 percent of score variance was between classrooms (Summary Table 1). At the other extreme, Japan and Korea both had less than 10 percent of score variance between classrooms. France was in between these extremes, with 27 percent of the score variance between classrooms. Thus, Japan and Korea, which had slightly more overall score variation than the U.S. in eighth grade

mathematics, had considerably more heterogeneity within the average classroom than the U.S. but showed substantially less variation among classrooms.

Summary Table 1: Percent of Variance Within and Between Classrooms, Grade 8 Mathematics: 1995

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

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

Predictors of performance in the U.S., France, Hong Kong, and Korea were analyzed using a two-level hierarchical model in which classrooms were the second level. The sampling design of TIMSS included only a single classroom per school in most countries and thus precluded analyzing schools to separate variation between schools from the total variation between classrooms. Predictors at both levels included presence of father in the home, presence of grandparents in the home, number of books in the home, presence of computer in the home, mother's and father's educational attainment, press for achievement, whether the student was born in the country of testing, and student's age. Numerous potentially important predictors, such as family income, ethnicity, region, and urban location, are not available in TIMSS, and others were dropped because exploratory data analysis revealed problems in one or more countries. Educational variables were not included in the models.

In all four countries, the final models, which included only subsets of these variables, predicted most of the between-classroom score variance but very little of the within-classroom variance (Summary Table 2). Korea was the only country in which the models predicted more than 5 percent of the within-classroom variance in scores. The consistently strong prediction of between-classroom variance is all the more striking in the light of the sparseness of our models and the relatively weak measurement of social background in TIMSS.

Summary Table 2.—Percent of Variance at Each Level Predicted by Final Models, Grade 8 Mathematics: 1995

	Between Classroom	Within Classroom
United States	77	4
France	59	5
Hong Kong	69	1
Korea	94	13

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

A somewhat different picture emerges when one asks how well the models predict total score variance in each country. This is affected not only by the within-level prediction (Summary Table 2), but also by the percentage of total variance found at each level (Summary Table 1). In this respect, the U.S. and Hong Kong are similar; in both, about one-third of the total variance in scores is predicted by the model, and almost all of this prediction is attributable to between-classroom differences in background variables (Summary Table 3). Korea shows a very different pattern: only 19 percent of total score variance is predicted by the model, and most of this most of this is attributable to within-classroom variables. The unusually strong prediction of within-classroom variance in Korea appears to reflect stronger relationships between scores and background variables within classrooms; the within-classroom variability of the predictors was similar in Korea and the U.S., and the within-classroom parameter estimates were markedly larger in Korea. France is similar to Korea in terms of the total variance predicted, but as in the U.S. and Hong Kong, the prediction is primarily due to between-classroom variables.

Summary Table 3.— Percent of Total Variance Predicted by Predictors at Each Level, Final Models, Grade 8 Mathematics: 1995

	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.

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

The final selection of variables and the parameter estimates differed among these four countries, but for several reasons, we place less confidence in these specific aspects of the models and focus primarily on the prediction of variance. The fit of the models appeared fairly insensitive to the specific choice of variables, and a more extensive or stronger set of background variables might have yielded substantially different parameter estimates.

Thus, in some instances, countries differ more in terms of the structure and prediction of performance variance than in the simple amount of variance. These differences raise several interesting questions. Why is the partitioning and prediction of score variance so different in Korea than in the other three countries? Why does Hong Kong, which resembles Korea and Japan in terms of its mean level of performance, resemble the U.S. rather than Korea or Japan in terms of the partitioning and prediction of score variance? Why does France show relatively little total and predicted variance between classrooms?

TIMSS does not provide a clear explanation of these differences, but they suggest hypotheses that warrant further investigation. The differences between Korea and the United States, for example, could be caused by differences in stratification. Korea does not track students in grade 8, and the very small percentage of score variance that lies between classrooms suggests that stratification of school attendance areas may also be much less pronounced than in the U.S. As a result, more of the relevant variance in background factors — that is, more of the variance that predicts mathematics performance — might lie within classrooms in Korea than in the U.S. or Hong Kong. Instructional differences among countries, for example, differences in the similarity of curricula across schools, could also contribute to patterns found here. For example, some observers argue that France has an especially uniform curriculum and rigid promotion policies; if that were true, it might contribute to the lesser variation between classrooms and the weaker relationship of that variation to background variables.

This study represents a first step in cross-national study of performance variation. Further analysis of TIMSS might shed additional light on these questions. For example, it might be informative to carry out multi-level models using both background variables and schooling variables as predictors. However, even more extensive analysis of TIMSS will leave many important questions unanswered. The design of a survey necessarily requires tradeoffs among uses of the data, and some aspects of the TIMSS design impose serious constraints on comparative analysis of performance variation. For example, in most TIMSS countries, TIMSS

5

sampled only one mathematics class per school, which makes it impossible to separate mathematics variation between classes within schools (which might be caused by tracking) from variation between schools (which could be caused by residential class segregation or disparities in resources). In science, the TIMSS data do not even permit decomposing variance into within- and between-classroom components. The TIMSS database also includes a relatively weak set of variables pertaining to student background and stratification. Accordingly, adequately some of the questions raised by the current research may require additional data.

ACKNOWLEDGEMENTS

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 U.S. In an era when traditional norm-referenced reporting of student performance has gone out of favor, "country norms" have become increasingly important. International comparisons are widely used to indicate the degree of success of U.S. education and the levels of performance to which this country should aspire. To some degree, international comparisons play the role that trends in the performance of U.S. students played a decade or more ago: they serve as a central indicator of student achievement, providing a framework for understanding U.S. performance and an impetus for education reform. The publication of the results of the Third International Mathematics and Science Study (TIMSS) since 1996 (Beaton, et al., 1996a, 1996b; Mullis, et al., 1997, 1998) has increased further the prominence of international comparisons in the U.S. 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 U.S. debate or has been used in a lopsided and potentially misleading fashion. Typically, when information about the variability of performance has been discussed, the variability in the U.S. has been considered, while the variability in the countries to which the U.S. 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 scale of the National Assessment of Educational Progress (NAEP), 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 U.S. 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 U.S. mean. In 1997, the President and the press gave much attention to the fact that a group of suburban Chicago districts that administered the TIMSS instruments separately, known as the First in the World Consortium,

had means similar to those of some of the highest-scoring countries, such as Korea. Again, high-scoring Korean districts were not compared to the U.S. average.

Underlying some of these comparisons appears to be an expectation that the variability of student performance is atypically large in the U.S. 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 U.S. states and foreign nations noted above, which did not display the variation of performance in other countries.

Depending on the variability of student performance in the nations to which the U.S. is compared, however, such comparisons may be fundamentally misleading. One needs to examine three aspects of the variability of student performance to understand comparisons of this sort:

- How large is the variability of performance in the countries to be compared? Is performance is much more variable in the U.S. than in many other nations?

- Does the student-level distribution of performance differ across countries in other respects? For example, if the distribution of performance were left-skewed (that is, showing a longer tail of unusually low-scoring students) in the U.S. but not in a second country, the U.S. mean would be pulled downward, and a comparison of mean scores in the two countries could be misleading.

- How is the variability of performance distributed across aggregates, such as schools, districts, or states? Are there differences in the clustering of students with similar performance? For example, are low-scoring students in the U.S. more concentrated in certain schools, districts, or states or regions than are similar students in some other nations?

In addition, it would be informative to explore the correlates of performance at both the individual and the aggregate levels. For example, background factors might explain more of the variability among schools or classrooms in one country than in another.

The third issue above, the question of clustering, can greatly influence the inferences that are warranted by international comparisons. For example, assume that the U.S. and several other countries have similar variability of performance at the level of students but that students in the U.S. are more highly clustered with students of similar performance. In that case, despite the similar variability at the level of students, the means of U.S. schools would be more variable than those in other countries, much as Berliner and Biddle (1995) argued. To take another example, if students in a second country are as highly clustered into high- and low-performing districts or states than are students in the U.S., than those areas in the other countries will also vary markedly, and comparisons between U.S. districts and the mean of the second nation would be potentially misleading.

Research Questions

This study was undertaken to explore the variability of performance in several countries. It had two goals: (1) to obtain substantive information about the variability of performance and its correlates, and (2) to explore the strengths and limitations of TIMSS for comparative analysis of performance variability. We anticipated that the study would result in suggestions for possible changes to TIMSS and suggestions for future secondary analysts. It has also resulted in some methodological refinements that can be adapted by other analysts.

The study was initially limited to Population 1 (elementary school) and Population 2 (middle school). Population 3 data were not available when the study was undertaken, and Population 3 presents substantial complications (e.g., because of differences in population definitions and coverage) beyond those posed by Populations 1 and 2. We focused primarily on Population 2 because of doubts about the validity and utility of self-report data from elementary school students.[1] Our analyses focused on the highest grade in each Population (grade 4 in Population 1 and grade 8 in Population 2), although some analyses were replicated in the lower grade as well.

---

[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).

We initially selected seven countries for our analysis: the U.S., Japan, Hong Kong, Korea, Germany, France, and Australia. We selected Japan and Korea because they are often used as examples of high-performing countries in comparisons with the U.S. Germany was included because it is often noted in discussions of the competitiveness of the U.S. 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 included primarily for methodological reasons. (The Australian sample, as noted below, shares one important characteristic with that of the U.S.) Data limitations, however, including missing data and apparent response bias, led us to narrow our focus substantially.

In the end, our most intensive work focused primarily on Population 2 mathematics in the U.S., France, Hong Kong, and Korea. Decomposition of performance variation was not feasible in science in Population 2 because of the construction of the sample. Students were selected for the science assessment based on the mathematics classes they attended rather than the science classes (Foy, Rust, and Shleicher, 1996, p. 4-7); in several of our countries, the number of science classrooms per school was highly variable and often large, making it impossible to partition variance into within- and between-classroom components.

Among the questions addressed by the analyses are these:

- How large is the performance variation in our sample countries, and how is this variation distributed between and within classrooms?
- How adequate are TIMSS background variables for analyzing performance variation?
- How well do background variables predict performance variation in our countries, both within and between classrooms?
- How consistent are the results of these analyses with other findings from U.S. data?

METHODS

One ideally would want at least three levels in a model of the relationships between background variables and test scores: within classrooms, between classrooms within schools, and between schools. Relationships at the between-school and between-classroom, within-school levels could represent different processes and could have substantially different implications for policy and practice. 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 U.S., Australia, and Cyprus, however, the TIMSS 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. In the U.S. and Australia, a second classroom was sampled in most schools, and additional classrooms beyond the second were sampled in a small number of schools. This permits estimating a three-level model that separates between-school relationships from within-school, between classroom relationships. This sampling, however, provides only a limited view of within-school, between-classroom relationships and cannot be compared to results in any other country.

Accordingly, we sacrificed some of the richness of the U.S. and Australia data in order to obtain results from those countries comparable to the results from others.
We did this by creating subsamples of the U.S. and Australia samples that consisted of a single classroom per school, randomly selected from the multiple classrooms in the original sample. For brevity, these samples are labeled "1 CPS" in this report. All results in this report reflect these one-classroom-per-school samples unless noted otherwise.

The reported analyses were weighted using sample weights unless noted otherwise. The procedure we used to weight intermediate statistics (e.g., when we first calculated a weighted mean for a unit and then calculated the weighted distribution of the weighted means) is described

in Appendix F. We modified weights and jackknife replicates for the U.S. and Australia to account for the additional sampling when we used the 1 CPS samples.

Our 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. Illustrative examples of this EDA are presented below.

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 total score variance in each of our seven countries was then decomposed into two components: within-classroom and between-classroom. The within-classroom component represents variation of students' scores around their classroom means. The between-classroom component represents the variation of classroom means around the grand mean for the country. This decomposition was carried out with no modeling of background variables in order to make the decompositions entirely comparable across the seven nations (which had differing numbers of variables available).

Simple bivariate relationships between performance and background variables were examined for all of the variables considered for the models. These 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).

Simple OLS regression analyses predicting performance from background factors were carried out for several purposes. For example, they were used to test the adequacy of imputations of missing values (imputation was rejected as a result) and to help shape and interpret multi-level analyses. The substantive multivariate results presented in this report, however, all reflect multilevel models.

Multilevel models were originally carried out using HLM software (Bryk, Raudenbush, and Congdon, 1996). However, HLM does not properly handle the weights required by the complex TIMSS sample design (see Appendix D). Accordingly, the primary results of multilevel modeling reported here were generated using SAS macros written to use the methods suggested by Pfefferman, et al. (1998). The models are described in more detail in a later section.

STUDENT-LEVEL DISTRIBUTIONS OF PERFORMANCE IN TIMSS

Basic information pertaining to the first of the four research questions noted above—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 TIMSS results in Populations 1 and 2 do not indicate that the achievement of U.S. students is atypically variable. In mathematics in both grades, the variability of performance was near the median. In science, the U.S. variability was near the top of the distribution but was not an outlier. This does not directly address Berliner and Biddle's assertion about the variability of schools; to do so would require a decomposition of the performance variance into within- and between-school components.

Differences across grades and subjects in these results, however, underscore the riskiness of generalizing about cross-national differences in the variability of student performance. The disparity across subjects in the ranking of the U.S. is only one of numerous inconsistencies across the four grades-by-subject combinations considered here. The reasons for these disparities are not clear. They could stem, for example, from characteristics of the tests, in which case a different set of tests could rank countries differently in terms of performance variation.

Among the 31 countries that met the TIMSS sampling requirements for the eighth grade (see Beaton, et al., 1996a, Table 2.1), the variability of mathematics performance in the U.S. was unexceptional. The country-level standard deviations varied greatly, from 58 to 110, but the standard deviations in half of the countries 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 U.S. sample was 91 (see Figure 1).

Figure 1.—Distribution of Country Standard Deviations in Mathematics, Grade 8, 31 Countries Meeting Sampling Requirements: 1995
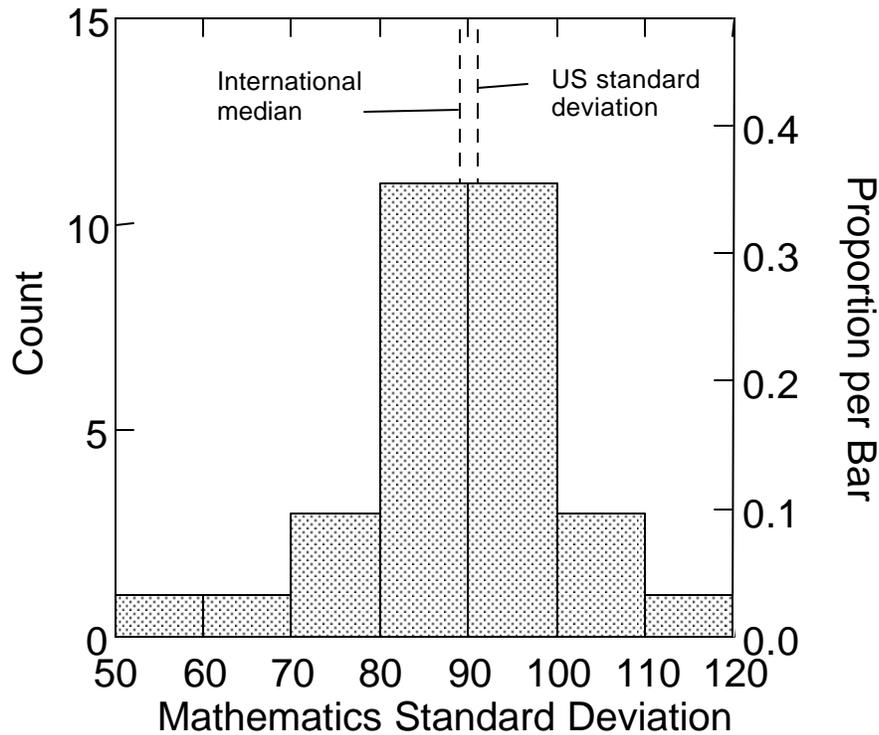
Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

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 2). (Note that Figure 2 has all of the countries in our sample labeled other than Germany. Germany was not included in the 31 countries meeting all sampling requirements because it tested 7th and 8th grades even though students in those grades were older than the TIMSS guidelines specified [Beaton, et al., 1996a, Appendix A].) Seen this way, the standard deviation of mathematics performance in the U.S. (91) was about 9 percent higher than the value of 83.4 that would be predicted from the U.S. mean of 500, which

was somewhat lower than the median of the country level means (522). Numerous other
countries also had standard deviations that deviated comparably from those predicted by their
means. For example, clustered tightly around the U.S. in Figure 2 are England and New
Zealand, and Germany would be as well if it were included in the Figure.

**Figure 2. —Plot of Mathematics Standard Deviation by Mathematics Mean, Grade 8, 31
Countries Meeting Sampling Requirements: 1995**
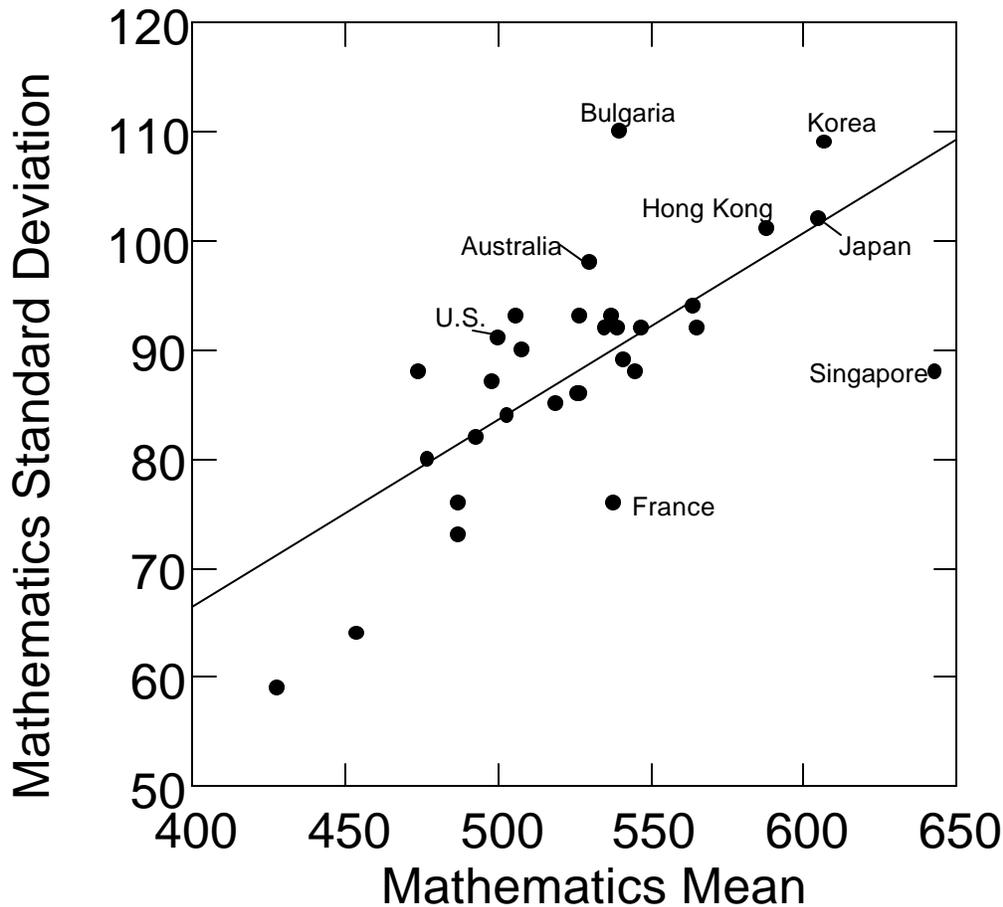
Source: Third International Mathematics and Science Study, Population 2 data set, International
Association for the Evaluation of Educational Achievement (IEA)

Figure 2 also rebuts the notion that high-scoring Asian countries have more equitable
(i.e., smaller) dispersions of performance, at least in eighth-grade mathematics. All three of the
Asian countries in our sample have larger standard deviations than does the U.S.: Hong Kong's

and Japan's standard deviations (101 and 102, respectively) are roughly 10% larger than that in the U.S. (91), and Korea's (109) is approximately 20% larger. The standard deviations of performance in Japan and Korea are roughly those predicted from their mean performance, while Korea, like the U.S., shows modestly more variability than is predicted from its mean. In our sample of seven countries, the only one with an unusually small standard deviation of eighth grade mathematics performance, either in absolute terms or relative to its mean, is France.

The TIMSS mathematics data also call into question the notion that the U.S. mean is pulled downward by a distribution with an unusually long left-hand (low-scoring) tail. None of the distributions of eighth-grade mathematics performance in our seven countries shows substantial skewness (Table 1). The distribution in the U.S. shows slightly more *positive* skewness than those in some other countries, indicating that the right-hand, or high-scoring, tail of the distribution in the U.S. is slightly longer than the left-hand tail. This can be illustrated concretely by comparing the score differences between various percentiles. In the U.S., the distance from the median to the 75th percentile is 69 points, while the corresponding distance between the median and the 25th percentile was a bit smaller, 59 points. Similarly, the median is farther from the 95th percentile than from the 5th. In contrast, in Korea, the distance between the median and the 75th percentile is similar to that between the median and the 25th percentile, while the distance to the 5th percentile is larger than that to the 95th percentile. Estimates of the 5th and 95th percentiles, however, should be interpreted with caution because of small counts.

Table 1. —Scale Points Between Percentiles, Grade 8 TIMSS Mathematics, Seven Countries: 1995

	Skewness	5th Percentile to Median	25th Percentile to Median	(Median)	Median to 75th Percentile	Median to 95th Percentile
Australia	.03	157	69	(529)	71	161
France	.06	119	50	(534)	57	132
Germany	.11	138	58	(506)	66	155
Hong Kong	-.23	180	69	(595)	64	147
Japan	-.08	173	72	(608)	68	163
Korea	-.08	191	69	(609)	73	177
U.S.	.23	138	59	(494)	69	159

SOURCES: Skewness coefficients calculated by RAND, using one-classroom-per-school subsamples in U.S. and Australia. Percentile differences calculated from Beaton, et al., 1996a, Appendix Table E.1.

Because of these differences in the shape of the performance distributions, the gap between the United States and Korea and Japan is larger among high-achieving students than among low-achieving students. For example, Korean students at the 95th percentile (within Korea) outscore U.S. students at the 95th percentile by 133 points, while Korean students at the 5th percentile outscore their U.S. counterparts by 62 points. The differences at the 75th and 25th percentiles are 119 and 105 points, respectively.[2] This can be seen graphically in Figure 3, which shows the distributions of eighth-grade mathematics performance in the U.S. and Korea. 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 U.S., however, pulling the U.S. tail closer to the Korean tail.[3]

---

[2]   The TIMSS reports use a different metric to compare achievement at different points on the distribution: the percentage of students in each nation reaching "international marker levels:" the 50th, 75th, and 90th percentiles of all students in the international sample (Beaton, et al., 1996a, Table 1.4; Beaton, et al., 1996b, Table 1.4). That metric, however, does not directly show differences in the distributions of achievement. If countries differ only in means and have otherwise identical and roughly normal distributions of scores, the low-scoring countries will look progressively worse as the percentile used to define a marker is increased. For example, a country with a below average mean will be more severely underrepresented in the top decile than in the top quartile, simply because of the mean difference.

[3]   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.

Figure 3. —Distributions of Mathematics Scores, Grade 8, Korea and U.S.: 1995
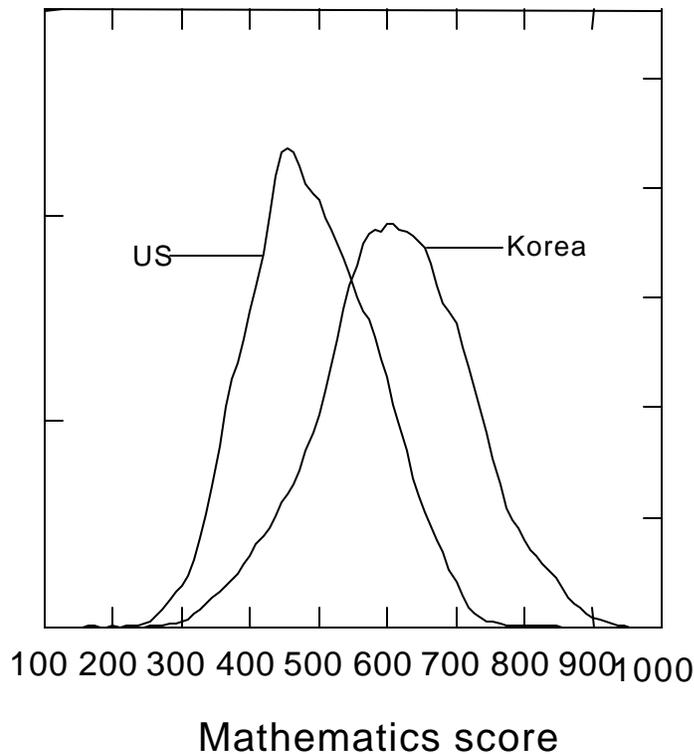
SOURCE: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

In contrast, in eighth-grade science, the variability in the U.S. was large relative to that in other countries. The total range of standard deviations across the 31 countries was slightly smaller in science than in math, ranging from 72 to 111, but the middle half of the countries spanned a wider range, from 81 to 98 points. The median of the 31 countries was 90. The U.S. standard deviation of 106 was well above this international median and was close to the maximum (Figure 4). However, it is important to note that three of the 31 countries (England, Australia, and Bulgaria) had standard deviations equal to or larger than that of the U.S., and four others (Ireland, Austria, Scotland, and New Zealand) were also within 10 percent of the U.S. That is, about one-fourth of the 30 other countries had standard deviations of science performance within 10 percent that of the U.S. Thus, the variability of performance in the U.S., while large, was not exceptional.

Figure 4. —Distribution of Country Standard Deviations in Science, Grade 8, 31 Countries Meeting Sampling Requirements: 1995
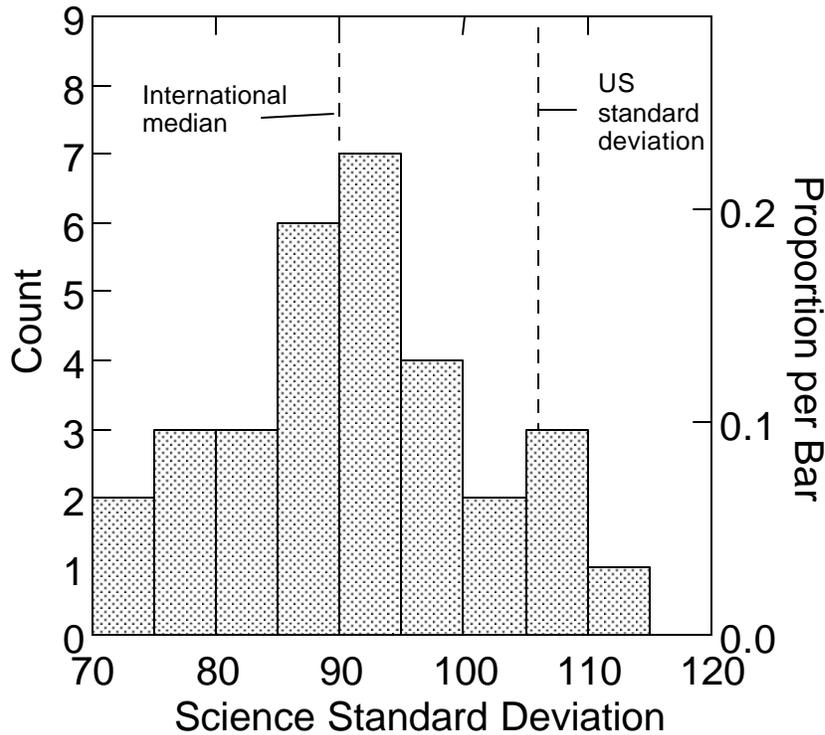

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

The relationship between variability and mean performance was weaker in eighth-grade science (r=.54; see Figure 5) than in eighth-grade mathematics. The standard deviations in the U.S., Germany, and Australia were substantially larger than predicted. (Note that Germany is excluded from Figure 5 as well as Figure 2 because it failed to meet TIMSS sampling requirements. Germany's data point would be very near that of the U.S.: its mean was 3 points lower than that of the U.S., and its standard deviation was 5 points smaller.) The standard deviation was again small in France—it was one of the smallest in eighth-grade science—and was again smaller than predicted by its mean score. In eighth-grade science, unlike mathematics,

the standard deviations of performance in Japan, Hong Kong, and Korea were near the middle of the distribution and were markedly smaller than that in the U.S.

Figure 5. —Plot of Science Standard Deviation by Science Mean, Grade 8, 31 Countries Meeting Sampling Requirements: 1995
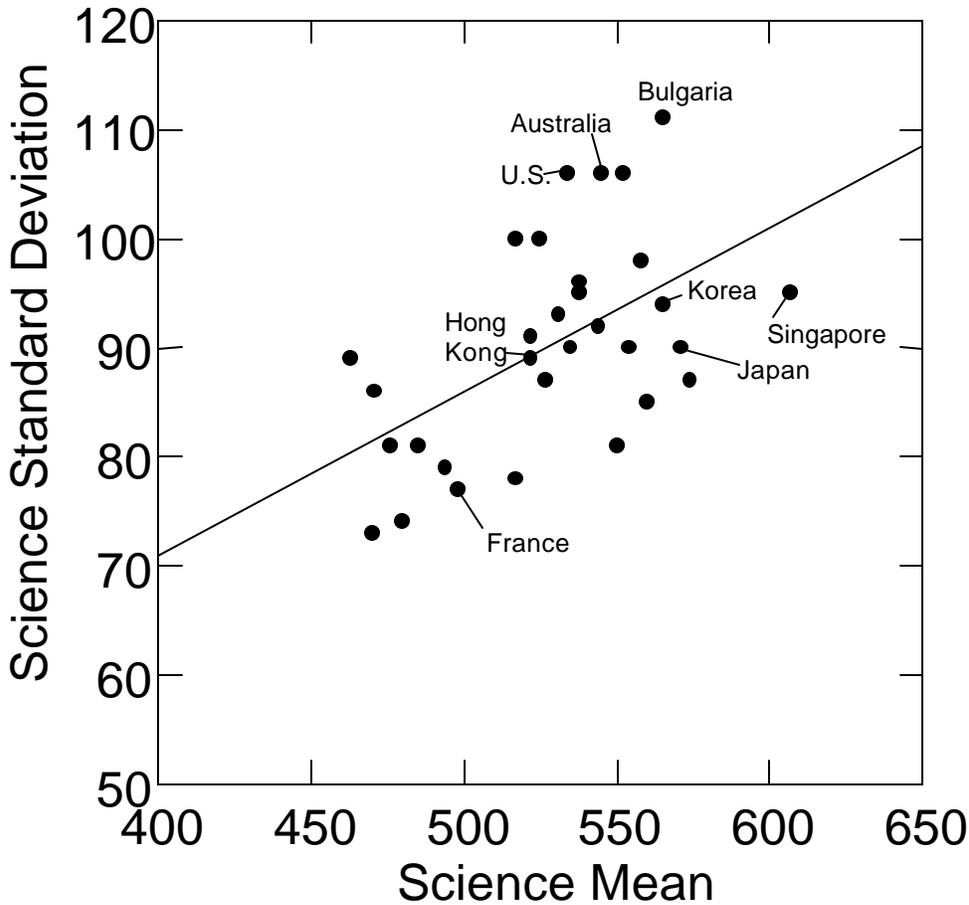
The fourth-grade data provided less information than eighth-grade data about student-level variability. Fewer countries participated in the fourth-grade TIMSS, and only 17 met all of the TIMSS sampling requirements. Two of our seven countries did not participate in the Population 1 assessment, and the Population 1 sample for Australia failed to meet the TIMSS requirement for participation rate before replacement (Mullis, et al., 1997, Appendix A).

In fourth-grade mathematics, half of the 17 countries that met all the TIMSS sampling requirements had standard deviations in the range of 79 to 89 points, and the median was 85 points. The standard deviation in the United States fell exactly at the international median

(Figure 6), while the U.S. mean score was appreciably above average. In this case, Japan, Hong Kong, and especially Korea showed substantially smaller standard deviations than did the U.S. In fourth-grade mathematics, Singapore had by far the largest standard deviation of these 17 countries (Figure 6), even though it had the smallest standard deviation of any of the highest-scoring countries in grade 8 (Figure 2).

Figure 6. —Plot of Mathematics Standard Deviation by Mathematics Mean, Grade 4, 17 Countries Meeting Sampling Requirements: 1995
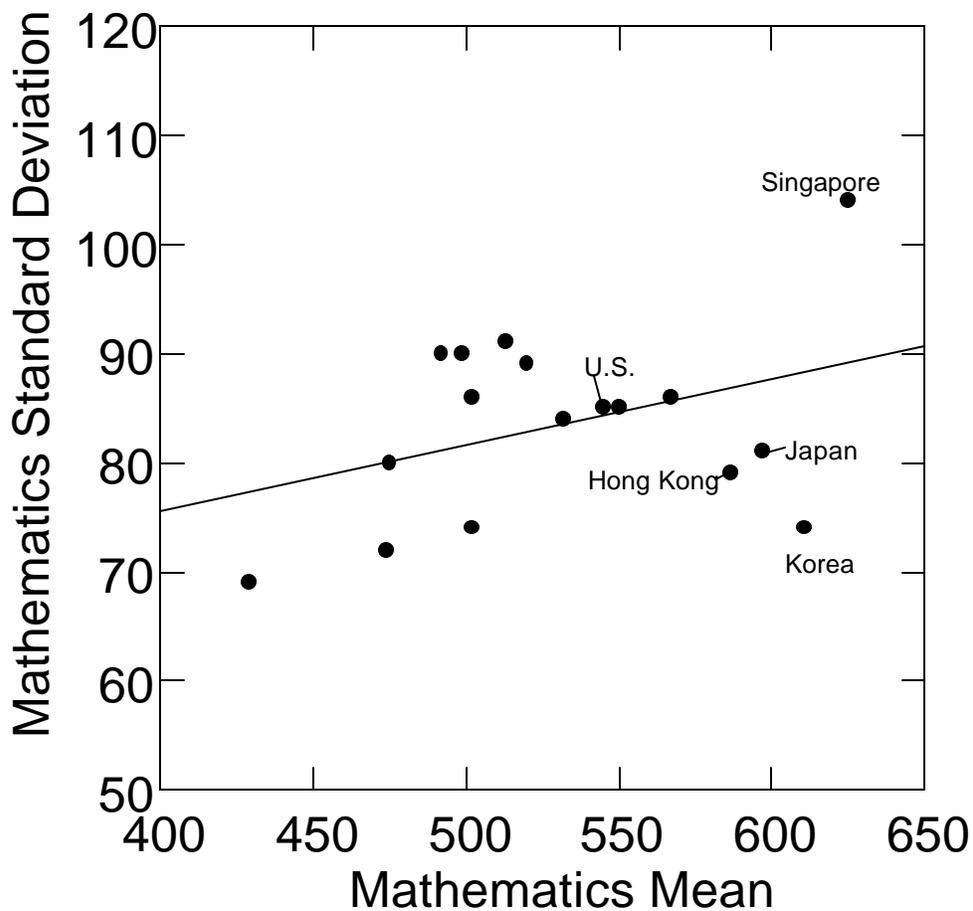
Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

In fourth-grade science, the standard deviations in half of the 17 countries ranged from 78 to 93, and the median was 85 points. As in eighth-grade science, the standard deviation in the

U.S., 95, was near the maximum (97). However, as in the eighth-grade, this did not make the U.S. exceptional; about one third of the 16 other countries had standard deviations within 10 percent of that of the U.S. (Figure 7). In fourth-grade science, even more than in fourth-grade mathematics, the standard deviations of performance were relatively small in Hong Kong, Japan, and especially Korea.

Figure 7. —Plot of Science Standard Deviation by Science Mean, Grade 4, 17 Countries Meeting Sampling Requirements: 1995
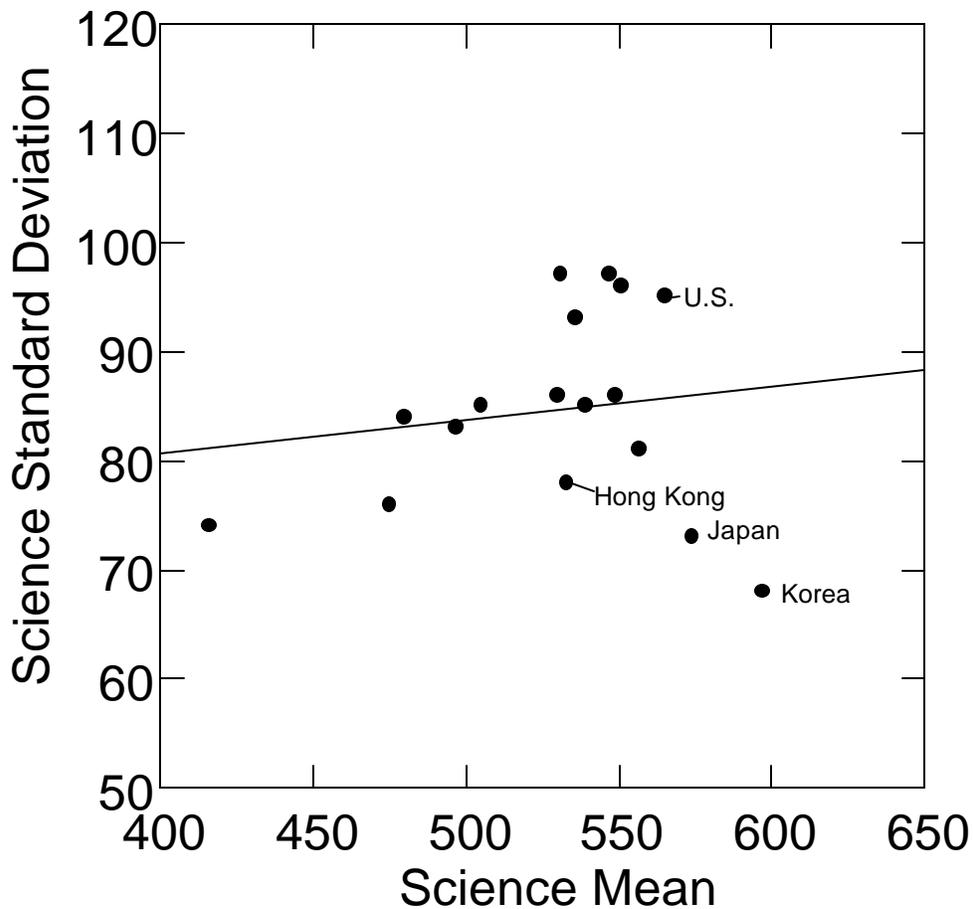
Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

SIMPLE DECOMPOSITION OF VARIANCE IN SEVEN COUNTRIES

To interpret differences in the variability of performance across countries requires an understanding of the extent to which the variability is clustered, for example, within classrooms or schools. As noted earlier, the TIMSS sample makes it impossible to distinguish clustering within schools from clustering within classrooms, and it permits decomposition of score variance only in mathematics. The decomposition of mathematics performance variance between classrooms and students, however, is sufficient to reveal striking differences among the seven countries in our sample.

In the United States, mean eighth grade mathematics scores show a wide dispersion among classrooms (see Figure 8; note that the country label is above each distribution). Classroom means vary roughly as much in Hong Kong, Germany, and Australia as in the U.S. (In Hong Kong, however, the distribution is left-skewed, while it is right-skewed in the U.S.) Japan is very different: classroom means are highly concentrated over a narrow range of scores. Korea's classroom means are nearly as concentrated as those in Japan. The distribution in France appears to fall midway between the concentration shown in Japan and the wide dispersion evident in the U.S.

Figure 8. —Distribution of Classroom Mean Scores, Grade 8 Mathematics: 1995
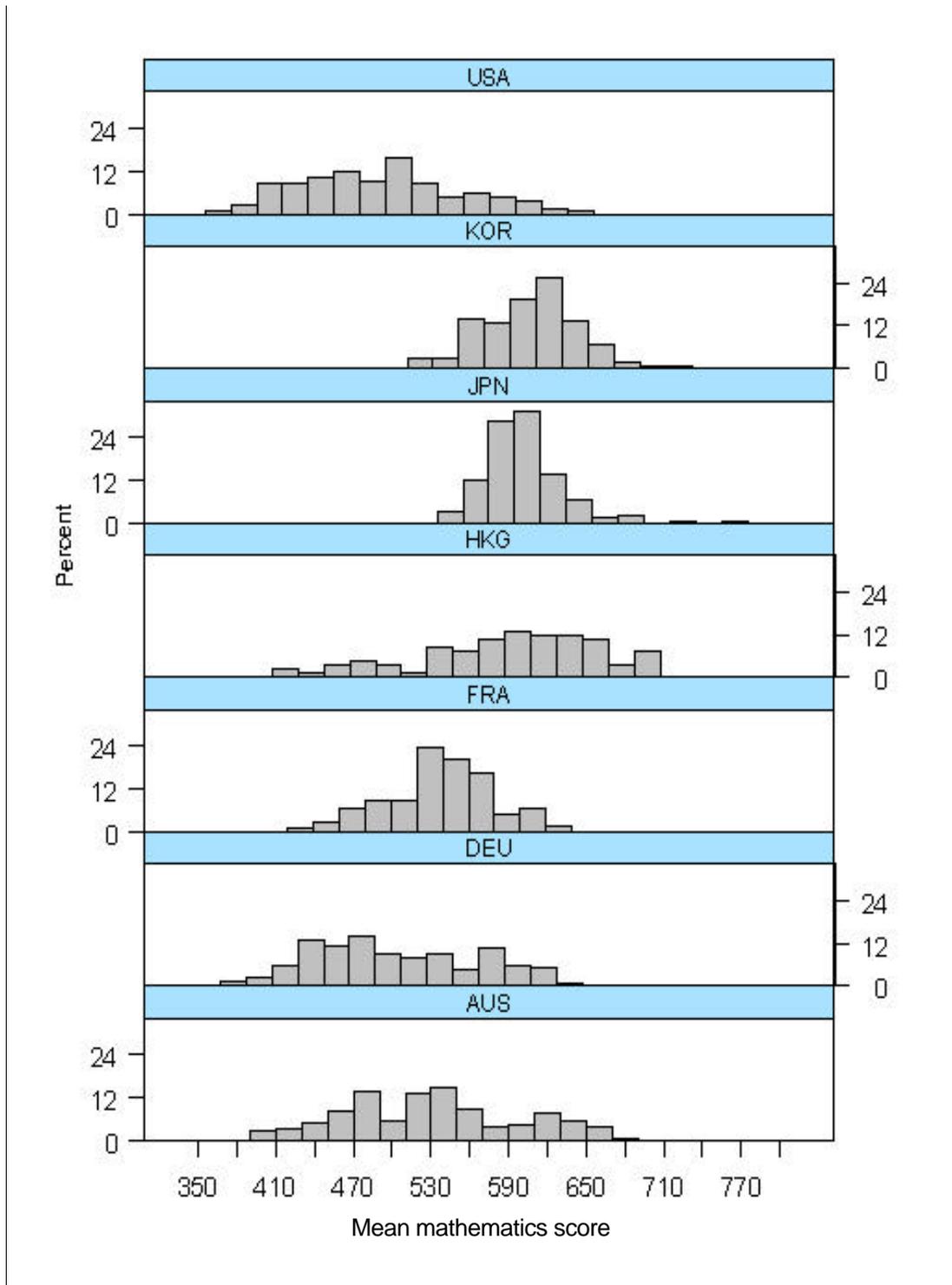
SOURCE: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

27

Summary statistics for these distributions of school means show the differences among them clearly. The standard deviation of classroom means is largest in Hong Kong (69) and Germany (62) (Table 2). Australia and the U.S. have standard deviations of classroom means of approximately 50. In contrast, classrooms vary less in France (SD=42) and especially Korea (SD=32) and Japan (SD=34).

Table 2.—Means and Standard Deviations of Classroom Means, Grade 8 Mathematics (1 classroom per school, weighted): 1995

Country	Mean	SD
Australia	530	52
France	538	42
Germany	509	62
Hong Kong	588	69
Japan	605	34
Korea	607	36
U.S.	500	51

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

These differences among countries are clarified when the total variance in student scores is decomposed into two parts, within-classroom and between-classroom (where the latter includes both variance between schools and variance between classrooms within schools). In the U.S., Hong Kong, Germany, and Australia, a bit over half of the total variance in eighth-grade mathematics scores lies within classrooms, so nearly half lies between classrooms (Table 3). 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. Similarities among some countries in this two-level 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.

Table 3: Percent of Variance Within and Between Classrooms, Grade 8 Mathematics (1 classroom per school, weighted), 1995

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

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA).

Schmidt, Wolfe, and Kifer (1993) partitioned the variance of eighth grade mathematics scores in six countries using data from the Second International Mathematics Study, one of the predecessors of TIMSS. Three of the countries analyzed here – France, Japan, and the U.S. – were analyzed by Schmidt, et al., as well. Their results for France and Japan were very similar to those obtained here; they found 79 and 91 percent of the variance within classrooms, respectively. Their results for the U.S. were less similar. They found only 46 percent of the score variance within classrooms, compared to the 58 percent found here. Schmidt, Wolfe, and Kifer's 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 U.S., only 9 percent of the aggregate variance lay between schools (with the remainder lying between classrooms within schools). The one-classroom-per-school design in most countries precludes similar analysis with TIMSS data.

The seven countries in our sample differ strikingly in terms of the heterogeneity of student performance within the typical classroom, with the U.S. 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, as shown in Table 3. As noted earlier, Japan and Korea have slightly larger national standard deviations than the U.S. in eighth grade mathematics. Because those Japan and Korea also have a much larger share of their total variance lying within classrooms than does the U.S., the typical within-classroom standard deviation in mathematics is considerably larger in Japan (96) and Korea (102) than in the U.S.

(74). (See Table 4.) The average classrooms are slightly less heterogeneous with respect to achievement in Germany and France than in the U.S., while the average classroom in Australia is slightly more heterogeneous. (The greater percentage of variance within classrooms in France compared to the U.S., Germany, and Australia shown in Table 3 is offset by the smaller total national standard deviation in France.)

Table 4: Within-Classroom Standard Deviations:
Grade 8 Mathematics (1 classroom per school, weighted)

| Country	Standard Deviation
Australia	83
France	63
Germany	64
Hong Kong	73
Japan	96
Korea	102
U.S.	74

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA).

Less divergence among countries in the decomposition of variance appears in the fourth grade. In the U.S. and Hong Kong, roughly two-thirds of the total variance in mathematics lies within classrooms; in Australia, 75 percent is within classrooms (Table 5). Japan and Korea again have roughly 90 percent within classrooms. The higher percentage of variance within classrooms in grade 4 compared to grade 8 in Australia, Hong-Kong, and the U.S. may reflect less differentiation among classes on the basis of performance—i.e., less tracking.

Table 5: Percent of Variance Within and Between Classrooms,
Grade 4Mathematics (1 classroom per school, weighted)

|	Percent Between	Percent Within
Australia	25	75
Hong Kong	33	68
Japan	5	95
Korea	12	88
U.S.	35	65

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA).

MULTILEVEL MODELS OF PERFORMANCE VARIATION

Multi-level models were estimated for four countries: the U.S., France, Hong Kong, and Korea. The first part of this section describes the process by which decisions were made about the pool of variables that would be considered in all four countries. The second part describes the type of multi-level models estimated. Subsequent parts present bivariate statistics and multi-level models for each of the four countries. A final section compares the results from the four countries.

Choosing Among Variables for Modeling Performance Variation

Our goal was to compare across countries the contribution of background factors to the variability of performance, distinguishing between variability within and between classrooms. Based on research in the U.S. about predictors of student performance, we chose to examine parental education, other measures of family composition, socio-economic status, 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, particularly in the younger grades. 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. If the propensity to hold back students because of expected poor performance or policies about age and entry differ markedly among countries, the relationships between age and performance could vary substantially as a result.

We did not examine curricular variables. As measured, these will not predict variation within classrooms, and research in the U.S. 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. Ideally, this work and the extant studies of curricular differences would be followed by more ambitious efforts to model the relationships of

performance with background factors and schooling variables jointly. However, even such studies would be limited by the more comprehensive measurement of curriculum in TIMSS and by probable differences among countries in the adequacy of measurement of each set of 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 and other factors collinear with them, and other omitted variables correlated with the measured variables. This is one reason that we place relatively little emphasis on the details of the parameter estimates obtained for specific variables. It is likely that these would change if a more comprehensive set of predictors were available, if predictors were measured with less error, or if we selected differently from among the available measures. There is accordingly a risk of treating the specific parameter estimates as more meaningful than they really are. The ability of the models to predict variance within and between levels, however, should be less sensitive to the specific choice of predictors, and we place more emphasis on those findings.

We carried out extensive exploratory data analysis to determine which of the relevant variables we should include in our models. We examined the distributions of responses across categories of the survey variables, the relationships of the survey variables to each other, and the total, within-classroom, and between-classroom associations between background factors and scores. In numerous cases, the results of these analyses required that we limit our formal modeling. A few key findings of this EDA are described here.

Press and Attitude Variables

Although TIMSS includes numerous attitude and press variables, we focused on a set of 15 Likert variables that ask students how strongly they disagree or agree with statements about the importance the student's mother, the student's friends, and the student herself place on doing well in mathematics, doing well in the language of the test, doing well in sports, being placed in the high-achieving class, and having 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. Some relationships that one would expect to be strong were weak or inconsistent; some that one would expect to be weak were strong; and some were negative or

curvilinear when one would expect them to be positive and monotonic. In several instances, the data showed suggestions of response bias.

For example, several problems can be seen in the responses of eighth-grade students to the BSBMMIP2 variable, "My mother thinks it is important for me to do well in mathematics at school" (Figure 9). This and the other press variables discussed here have four response categories: strongly disagree, disagree, agree, and strongly agree. The analysis of this variable is described in some detail here to illustrate the EDA approach taken.

Figure 9. Mathematics Scores and Responses to BSBMMIP2 Press Variable, Grade 8 Mathematics: 1995
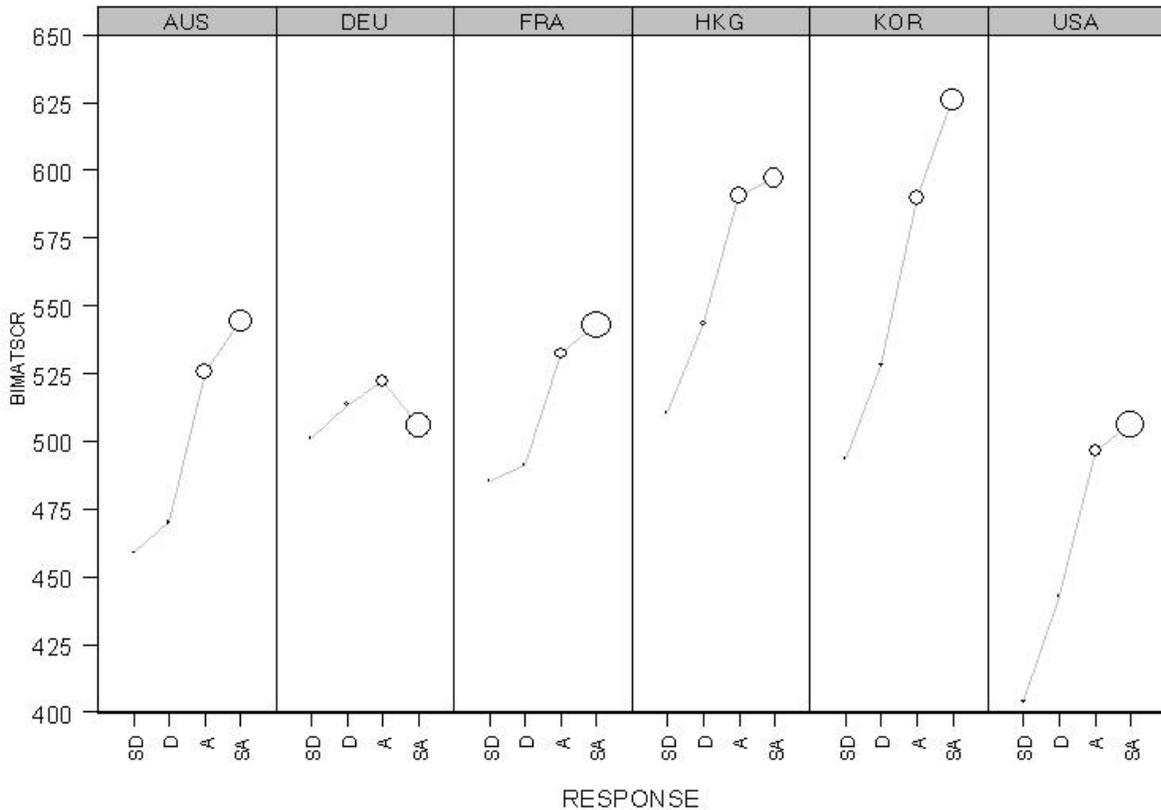
Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

33

Because Figure 9 is an unusual graphic (called a trellis plot [MathSoft, 1998] or a multiway dot plot [Cleveland, 1993]), we explain here how it should be read. Each of the seven panels arrayed across the figure represents one of the seven countries in our sample. The common vertical axis, labeled on the left as 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. Finally, 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, the relative sizes of the circles will understate the actual differences in cell counts. This happened with numerous of the TIMSS press and attitude variables that had extremely small cell counts. For example, there were instances in which the ratio of the largest to the smallest count across the four response categories within a single country was well over 100:1.

In all of the seven 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 many instances, however, this relationship stemmed in large measure from very small groups of students. In the U.S., for example, 97 percent of all students were in the "strongly agree" and "agree" categories. (This is one of the instances in which the graphic understates the severity of differences in cell counts.) The mean mathematics scores of these two groups that included almost all students 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. Somewhat similar patterns appeared in most of the other countries. This pattern does not call the validity of the responses into doubt, but it does indicate that the variable is likely to have relatively little utility in predicting scores. A different measure that more fully captured variation in mothers' press for mathematics achievement would be more useful.

Two other patterns in Figure 9 bear mention: the extremely strong positive relationship between BSBMMIP2 and scores in Korea and the curvilinear relationship with scores in

Germany. Both of these patterns appeared repeatedly across the TIMSS press and attitude variables. Indeed, the unexpected response pattern in Germany was stronger than this in the case of some other variables. For example, one might expect BSBMSIP2, "I think it is important do well in mathematics at school," to show a particularly strong relationship with scores. Yet if one excludes the 3 percent of German students who strongly disagreed with that statement, the relationship between that variable and mean scores in Germany was consistently if modestly negative. Korea poses the opposite problem: relationships with student performance were so consistently and strongly positive that they raise concerns about the validity of responses. The very strong positive relationship appearing in Korea extended to BSBGSIP6 and BSBGMIP6, "I think it is important be placed in the high achieving class" and "My mother thinks it is important for me to be placed in the high achieving class," even though eighth-grade classes are not tracked by achievement in Korea (Hyung Im, 1998). One can only speculate about the causes of these patterns, but one obvious possibility is cross-national differences in response bias.[4]

The relationships between some other press variables and student performance varied markedly, sometimes dramatically, among countries. For example, the relationships between achievement and both "I think it is important be placed in the high achieving class" and "My mother thinks it is important for me to be placed in the high achieving class" are negative in Germany; essentially zero in France and Japan; modestly positive in Australia, Hong Kong, and the U.S., and very strongly positive in Korea. 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 patterns. 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.

The TIMSS press variables include some that one would expect to show weaker 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

---

[4]   The positive relationships between mean scores and both BSBGSIP6 and BSBGMIP6 in Korea might also stem from translation difficulties. Despite the phrase "placed in" in the English version of the question, perhaps students interpreted the question to be whether they consider it important for their class to achieve well.

to have fun. One might expect strength of agreement with these statements to show negative relationships with performance: students who think it particularly important to save time for fun might be less willing to put long hours into study, for example. Yet quite the reverse is true of two of the three variables of this sort: two of the *strongest positive* predictors of mean scores from this set of variables 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 the case of BSBGSIP4, for example, this relationship is monotonic and strongly positive in every country but France. Indeed, these are the only two variables from this set of 15 press and attitude variables in which the strength of agreement is positively and monotonically related to mean scores in all of our seven countries, even Germany.

Thus, EDA of this set of 15 variables reveals several potentially serious limitations. Using the traditional standard of convergent/discriminant evidence—that is, evidence showing whether the size and directions of correlations among the observed variables are consistent with what one would expect in the light of the constructs they are supposed to measure—the set of 15 variables looks questionable. Moreover, some of the country-specific patterns appear suspect as well.

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. While the concentration of responses to these variables weakens them, the pattern of responses to them is otherwise reasonable in most of our seven countries, and their theoretical relationship with achievement is clear. In final analyses, we pooled these two variables for each subject, creating a single "press for mathematics variable" variable from the students' responses pertaining to themselves and to 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 did not extend our multi-level modeling to Germany, where these variables showed the most questionable relationships with scores.

Student and Family Background

Similar exploratory analyses were carried out for 10 of the student and family background variables we considered: 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. These analyses were necessarily restricted to six countries, as Japan did not collect data about these variables.

In general, 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.

The presence of the mother and father in the home was positively associated with scores in every country, although more strongly in some countries than in others. In most of our six countries, however, there was only modest variation in this variable – that is, few students reported that the father did not live at home. For example, roughly 8 percent of respondents in Hong Kong and 11 percent in Korea reported that they did not live with their fathers. In contrast, one-third of U.S. eighth-grade respondents reported that they did not live with their fathers. By way of contrast, the presence of the grandfather in the home showed variable relationships with scores across the seven countries: negative in three and near zero in the other three. It is possible that across countries, the presence of grandparents shows different relationships with other background variables that are more directly related to achievement. We excluded this variable from all models.

Mothers' and fathers' educational attainment—shown by numerous studies in the U.S. to be a particularly powerful predictor of students' educational performance—showed a clear, positive relationship with TIMSS mathematics scores in all six countries, and it showed a reasonable dispersion across the categories of the TIMSS variables in most instances. There were a few patterns in the data, however, that required consideration in estimating models. Mother's educational attainment (BSBGEDUM, Figure 10) shows the categories ordered as in the TIMSS database and shows several instances in which mean scores are not monotonically increasing. Some of these instances, such as the "some university" and "finished university" difference in Korea, may reflect nothing more than sampling error stemming from small cell

counts. (For example, only 23 Korean eighth-grade students reported that their mothers had completed "some university.")

Figure 10. —Mathematics Scores and Responses to Question about Maternal Education, Grade 8 Mathematics: 1995
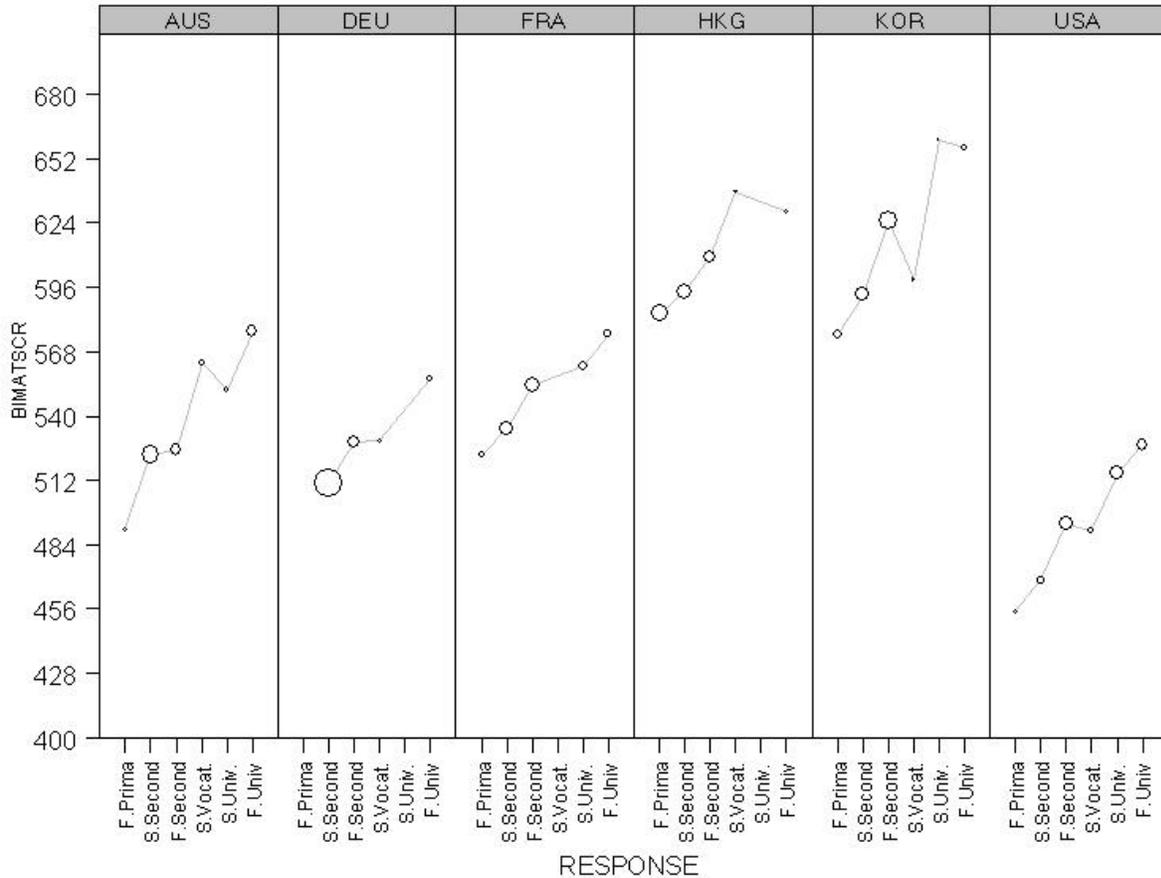
Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

The "some vocational" category, however, poses a more difficult problem. The means of the "some vocational" and the "finished secondary" categories are not always in the same order. Therefore, we combined the "some vocational" and "finished secondary" categories in our models. With this change, the relationship between scores and parental educational attainment was more often monotonic, and in several countries, it was close enough to linear to suggest using the recoded attainment variables as single variables (as opposed to sets of dummy variables) in models. The data patterns in Hong Kong suggested a different approach in that country for mothers' education only, which was to collapse "some vocational" with "finished

university." (These two categories contained a total of only 219 cases; "some university" contained none.)

These generally consistent, positive bivariate relationships between parental education and student achievement are important for interpreting the multivariate models reported below. The multivariate models showed surprisingly weak effects of parental education. In the U.S., we explored these weak effects by carrying out similar analyses using the National Education Longitudinal Study (NELS-88) base year data. The NELS analyses showed stronger effects of maternal education. (The comparison of NELS to TIMSS is presented after the U.S. two-level models.) Unfortunately, we cannot determine the reasons for this discrepancy between TIMSS and NELS. However, we place more confidence in the decomposition of variance between levels than in particular selections of variables or parameter estimates, and this discrepancy should not greatly affect the decomposition of variance.

The three variables about possessions in the home (study desk, computer, and number of books) all showed positive relationships with scores in all six countries, although in several nations, the first two showed such limited variability that their utility in modeling was doubtful. For example, only 4 percent of French eighth-grade students reported that they did not have a study desk. The number of books reported (BSBGBOOK) showed a positive relationship with mean scores in every country, although the strength of the relationship varied considerably across the six. The dispersion of students across the TIMSS *a priori* categories was also generally quite good in our six countries (Figure 11). This relationship was also monotonic in all instances other than one group in France. In France, the top category (more than 200 books) scored on average lower than the next category (101 to 200 books). Therefore, we included the number of books in our models, but in the case of France, we collapsed the top two categories into one.

Figure 11. —Mathematics Scores and Responses to Number of Books in Home, Grade 8 Mathematics: 1995
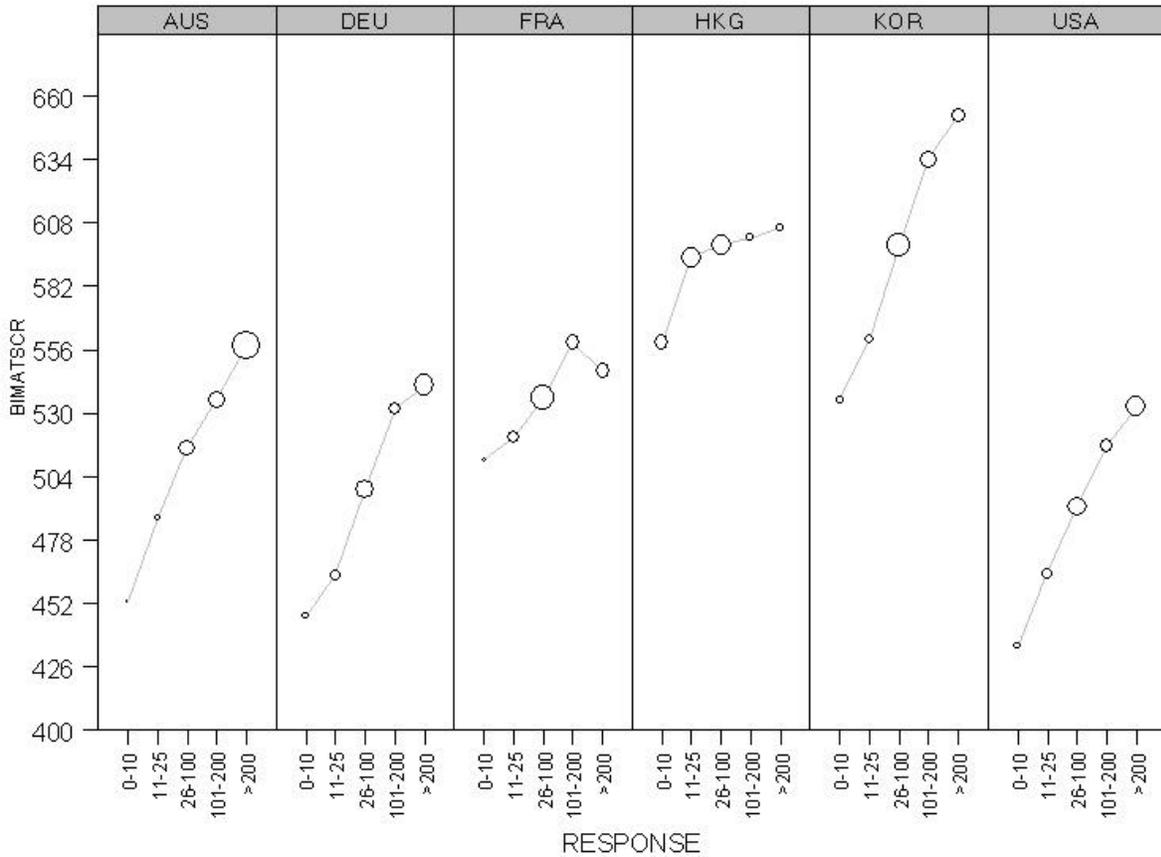
The final variable in this set asks the student whether she was born in the country in which the test was administered. On average, immigrant students scored lower than others in the U.S. and Germany but about the same as others in Australia and higher than others in Hong Kong. The differences in this relationship across countries might reflect real differences in the characteristics of immigration—e.g., differences in countries of origin and in the educational level of immigrant families. Thus we decided to explore the utility of this variable in formal models, with the recognition that the implications of the variable could differ markedly across the countries in our sample.

While response patterns for these variables generally seemed reasonable, missing data and uninformative responses posed serious difficulties in several instances. In some cases, important variables were not collected at all. Japan collected none of these background

variables, and France omitted the variable asking students if they were born outside of the country. (France also did not administer questions about the father's and mother's country of birth.)

In all of our countries, responses to the questions about parents' educational attainment were missing for substantial percentages of students. This problem was greatest in France (where 17 percent were missing for fathers and 16 percent for mothers) and Germany (13 and 12 percent, respectively).[5] More important, of the students who responded to these question, many answered "I don't know." This problem was particularly severe in France, where 16 percent of eighth-grade students did not respond to the question about their mothers' educational attainment, and an additional 34 percent responded "I don't know," so that a total of 50 percent of respondents provided no informative answer (Table 6). Thirty percent of students in Germany and 18 percent in Australia failed to provide an informative response.

Table 6. —Percent of Students with No Response or Response of "I Don't Know" to Question About Mother's Educational Attainment: 1995

	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
U.S.A	3	7	11

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

The frequency of missing data and "I don't know" responses required that we choose among three approaches: (1) omit parental educational attainment from models in France in order to use most of the full sample; (2) include one or more parental education variables without imputation and use a substantially reduced sample; or (3) impute parental educational attainment for half the sample so that we could use the full sample while including one or both parental education variables in the models. Efforts to impute missing parental education data proved

---

[5]   These percentages, which match those in the TIMSS publications, combine cases of student non-response with cases in which the question was not administered. Except in Japan, where these questions were not administered to any students, non-response was more common than instances in which questions were not administered. In Population 2 in France, for example, the question about mother's attainment was not administered to 3.4 percent of students, while 12.7 percent failed to answer the question.

unsatisfactory for several reasons, including the sheer magnitude of the missing data problem, weak prediction of the absence of informative responses, and evidence that imputation biased the results of the model. Comparisons of preliminary models unfortunately showed that the choice between the two options that did not involve imputation affected the results appreciably, although it appeared not to have a major on the overall prediction of score variance. Thus, any choice results in findings for France that have to be interpreted with caution. Analyses of the impact of this choice can be found in Appendix C.

The final selection of variables used in the analysis and their sources are listed 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:

$$y_{ij} = \boldsymbol{a} + \underline{\boldsymbol{b}}(\underline{x}_{ij} - \overline{\underline{x}}_j) + \underline{\boldsymbol{g}}\,\overline{\underline{x}}_j + \boldsymbol{e}_{ij}$$

where the subscript i indicates individuals, j indicates classrooms, an underscore indicates a vector, and a bar over a variable indicates a mean. In other words, 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 estimated for individual characteristics are held constant across classrooms. (That is, there are no cross-level interactions between individual and classroom characteristics.) Equivalently, this can be expressed in terms of two levels as follows:

$$y_{ij} = \boldsymbol{a}_j + \underline{\boldsymbol{b}}(\underline{x}_{ij} - \overline{\underline{x}}_j) + \boldsymbol{e}_{ij}^*$$

$$\boldsymbol{a}_j = \boldsymbol{a} + \underline{\boldsymbol{g}}\,\overline{\underline{x}}_j + \boldsymbol{h}_j$$

$$\boldsymbol{e}_{ij} = \boldsymbol{h}_j + \boldsymbol{e}_{ij}^*$$

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.

These models center observations around classroom means—that is, each student's value on each predictor x is expressed as a deviation from the mean of that variable in that classroom. Without group-mean centering, the predictor variance within and between classrooms would be confounded. For example, if one added to a classroom a student whose value on a given predictor was very low, that addition would contribute to within-classroom variance in the predictor (the student would be low relative to the classroom mean) and between-classroom variance (the student would pull down the mean of the classroom). Centering eliminates confounding of the predictor variance between and within classrooms. Centering also frees us from the assumption that the effects of variations in classroom means equals the effects of deviations within classrooms and makes the model's coefficients straightforward estimates of the within-classroom and between-classroom effects (e.g., Bryk and Raudenbush, 1992). With group-means centering, the classroom-level coefficients ⇕ are not estimates of context effects, that is, the effect of aggregate variables above and beyond the impact of individual-level variables. When slopes are fixed, as in our models, subtraction of a within-classroom coefficient (↕) from the corresponding classroom-level coefficient (⇕) provides an estimate of the context effect.[6]

One could also specify more complex models; for example, one could allow the student-level regression coefficients to vary across classrooms. Preliminary analysis, however, suggested that little would be gained by making the models more complex, either by allowing the slopes to vary randomly or by modeling their variation.

We began with the assumption that the full range of background variables that survived EDA screening 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 selected 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

---

[6] Alternatively, an uncentered model would provide direct estimates of context effects, and the between-classroom effects could then be estimated by adding each ⇕ to the respective ↕. For our purposes, however, it is more straightforward to estimate the within- and between-classroom effects directly.

significance of coefficients.[7] 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 micro-level variables, we included at both levels any variable that was significant at either level. This process is illustrated by Appendix B, which presents all of the models used to select the final U.S. model.

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 variation in 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. The percentage of variance attributable to a given predictor is simply the change in R^2, and we calculated for each factor in our models and at both levels of aggregation the increments in R^2 that would have obtain if that variable had been entered first and last. In analyzing a single sample, that would be sufficient for many purposes, but for comparing across nations, it is not. Countries may differ not only in terms of the impact of a given predictor on performance, but also in terms of the variation (and the clustering of variation) shown by the predictor. Comparisons of R^2 would be affected by both of these considerations. Therefore, we present for each model 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.

---

[7]   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.

Decomposing Performance Variation in the U.S.

We began within-country analysis in the U.S. for two reasons. The EDA showed relatively few problems in the U.S. background data—for example, the U.S. did not show the peculiar associations with scores found in Germany and had much less severe missing data about parental education than did France. Moreover, our familiarity with research on the predictors of achievement in the U.S. gave us more of a basis for formulating and testing models in the U.S. than elsewhere.

Bivariate Relationships

The simple correlations between background variables and scores—pooling students without regard to classrooms—were typically small (Table 7).[8] The largest correlation of any background variable with mathematics scores was .34, the correlation between scores and the number of books in the home. The presence of a computer in the home, father's education, and mother's education all showed roughly the same correlation with math scores, from .24 to .28.[9]

A few of the correlations between background variables were larger. The correlation between mother's and father's education was moderate, .55. The largest of the other correlations among the background variables was .33, and most were smaller or trivial. These modest correlations are not a result of nonlinearities; the only substantial nonlinearities involved cells that included very few cases. This pattern of modest intercorrelation is important for interpretation of the models presented below, because it rules out high collinearity at the student level as an explanation for the modest impact of many variables in the models.

---

[8] The press variables have been recoded so that "strongly disagree" has a value of 1 and "strongly agree" has a value of 4. The TIMSS data have the press variables coded in the opposite direction, which causes substantively positive relationships to appear as negative correlations.

[9] Note, however, that these are different correlations and are not entirely comparable. The correlations with number of books, mother's education, and father's education are point-polyserials, while the correlation with computer present is a point-biserial.

Table 7.—Student-level correlations between background variables and scores, USA, uncentered: 1995

	Father present	Number of books	Computer present	Press	Born in country	Mother's education	Father's education	Age	Math score
Father present	1								
Number of books	0.16	1							
Computer present	0.18	0.33	1						
Press	0.01	0.12	0.06	1					
Born in country	-0.02	0.12	0.06	0.10	1				
Mother's education	0.06	0.31	0.28	0.04	0.01	1			
Father's education	0.12	0.32	0.32	0.06	0.00	0.55	1		
Age	-0.06	-0.11	-0.07	-0.10	-0.04	-0.08	-0.10	1	
Math score	0.14	0.34	0.24	0.12	0.07	0.22	0.28	-0.16	1

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

These already modest student-level correlations dropped sharply when the between-class variance was removed by centering all observations around their classroom means (Table 8). In this case, classroom means were pulled out, and correlations reflect only students' deviations from classroom means. (In effect, these are within-classroom correlations, pooled across classrooms.) The strongest centered correlation between any of the background correlations and mathematics scores was the correlation of .14 between scores and the number of books in the household.

Note that within classrooms, the correlations between scores and mother's and fathers' education were trivial. The relationship between parental education and scores lies almost entirely between classrooms.

The relationships between background variables and scores were generally much stronger at the classroom level—that is, when classroom means were correlated (Table 9). Four of the background variables showed correlations above .60 with mean mathematics scores: mean number of books (.78), percent with computers in the home (.70), mean mother's education (.64), and mean father's education (.66). At the classroom level, mean mother's and father's education were quite collinear with some other background variables, showing correlations of about .75 with mean number of books and .78 with percent with computers. These strong relationships are echoed in the strong classroom-level coefficients in the multilevel models, and the collinearity of these variables at the classroom level is important in interpreting those models. Note that the relatively large classroom-level correlations reflect strong covariance and substantial variations in classroom means, not merely reduced sampling error in means compared to student scores.

Table 8.—Student-level correlations between background variables and scores, USA, centered: 1995

	Father present	Number of books	Computer present	Press	Born in country	Mother's education	Father's education	Age	Math score
Father present	1								
Number of books	0.11	1							
Computer present	0.15	0.22	1						
Press	0.01	0.11	0.04	1					
Born in country	-0.03	0.09	0.05	0.08	1				
Mother's education	0.02	0.19	0.17	0.01	-0.01	1			
Father's education	0.09	0.19	0.19	0.03	-0.01	0.46	1		
Age	-0.04	-0.06	-0.03	-0.07	-0.04	-0.05	-0.05	1	
Math score	0.04	0.15	0.07	0.09	0.04	0.06	0.10	-0.13	1

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

Table 9.—Classroom-level correlations between background variables and scores, USA: 1995

	Father present	Number of books	Computer present	Press	Born in country	Mother's education	Father's education	Age	Math score
Father present	1								
Number of books	0.42	1							
Computer present	0.37	0.78	1						
Press	-0.05	0.25	0.23	1					
Born in country	-0.01	0.32	0.16	0.23	1				
Mother's education	0.23	0.77	0.79	0.28	0.20	1			
Father's education	0.29	0.74	0.80	0.27	0.13	0.85	1		
Age	-0.22	-0.33	-0.26	-0.26	0.02	-0.28	-0.31	1	
Math score	0.55	0.78	0.70	0.30	0.26	0.64	0.66	-0.28	1

Multilevel Model

The final two-level model of mathematics scores in the 1 CPS sample 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, an academic press variable, and student age (with a quadratic term). The press variable was the mean of two variables: the extent to which students agreed with the statements that they and their mothers thought it important to do well in school.[10] 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. As noted, the classroom-level effects were limited to varying intercepts; the within-classroom slopes were held constant across classrooms. All within-classroom variables were centered about the classroom means, in order to eliminate confounding between within- and between-classroom effects.

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 approaches are each described in turn.

Within classrooms, the strongest effects were those of the number of books, the academic press variable, and students' age (Table 10). The effects of having a computer and the father living at home were both smaller and non-significant. It is important to note, however, that most of these estimates are imprecise. The right-most two columns of Table 10 show the upper and lower bounds of a 95% confidence interval around each estimated effect. Most of these bands are wide, as one would expect from the modest t values.

---

[10]   As noted earlier, when one of these variables was missing, the second was used alone in lieu of the mean.

Table 10:.—Two-Level Model of Mathematics Scores, USA, Grade 8: 1995

Variable	Effect	SE	t	p	CILB[a]	CIUB[b]
Intercept	-351.7	265.9	-1.3	0.19	-884.9	181.5
Within class ($\updownarrow$)						
Number books	7.9	1.2	6.7	0.00	5.6	10.3
Computer present	4.4	3.5	1.2	0.22	-2.7	11.4
Father present	1.7	3.3	0.5	0.60	-4.9	8.3
Press	9.6	2.6	3.7	0.00	4.4	14.7
Age	-14.4	3.3	-4.3	0.00	-21.1	-7.7
Age^2	-6.9	3.7	-1.9	0.07	-14.2	0.5
Between-class ($\updownarrow$)						
M Books	45.5	7.4	6.2	0.00	30.7	60.2
M Computer	37.2	16.8	2.2	0.03	3.6	70.9
M Father present	90.3	21.4	4.2	0.00	47.4	133.2
M Press	43.2	17.1	2.5	0.01	9.0	77.4
M Age	33.9	15.3	2.2	0.03	3.2	64.6
M Age^2	-149.4	37.1	-4.0	0.00	-223.8	-75.0
Residual variances						
σ^2 (within)	4570.4					
τ (between)	766.2					

[a] Lower bound of 95% confidence interval around parameter estimate
[b] Upper bound of 95% confidence interval around parameter estimate
NOTE: All estimates of error and significance reflect jackknifed estimates.

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

Most of the estimated within-classroom effects were small to moderate. For example, the estimated effect of the number of books was 7.9. This variable had five categories. (We entered this as a single variable for simplicity because the bivariate plot of scores versus number of books was essentially linear.) The model predicts that holding constant the other variables, the mean difference between students in the lowest and highest categories would be 32 points. The standard deviation of mathematics scores in this subsample was 89.4 points. Thus, the predicted difference in mathematics scores between students in the two most extreme categories of number of books is roughly one-third of a standard deviation. The press coefficient was larger, but since most students are 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 Table 10 and all similar tables following) 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 most of the between-class coefficients are only modestly larger than those of the corresponding within-class estimates.

Nonetheless, there are some striking differences between the within- and between-classroom estimates. The presence of the father in the home had a non-significant and near-zero relationship to scores within classrooms but a substantial relationship between classrooms. The model estimates that on average, the within-classroom effect of having the father present was less than 2 points, roughly 2 percent of a standard deviation. The mean of the father-present variable is the proportion of students with father present, which could in theory vary from 0 to 1. The model predicts that holding other variables constant, an increase from 0 to 1 would be accompanied by an increase of 90 points on the mathematics scale. Classrooms in our grade 8 mathematics model sample ranged from 15 to 100 percent of fathers present, with a mean of 66 and a standard deviation of 16. Thus, the model estimates that holding other variables constant, going from the minimum to the maximum values we observed for the percent of fathers present would be associated with a gain of 77 points, or almost .9 standard deviation. 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 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 was within classrooms, while the remaining 41 percent was between classrooms (Table 11). 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 11. —Total and Predicted Variance in Mathematics Scores at Each Level, USA, Grade 8: 1995

	Between classroom	Within classroom
Total variance at level	3299	4769
Percent of variance at level	41	59
Variance predicted by variables at level	2532	198
Percent of variance at level predicted by variables at level	77	4
Percent of total variance predicted by variables at level	31	2

Comparing TIMSS to NELS

The lack of a substantial relationship between scores and parental education and students scores in the U.S. multilevel models was surprising. Parental educational attainment—in particular, the educational attainment of the mother—is generally considered one of the most powerful predictors of student performance.

The lack of importance of parental education in our models could stem from the particular models we used. Many of the studies that show parental education to be important predictors employed single-level models. Moreover, multilevel models are often very sensitive to the particular specifications employed (Kreft and DeLeeuw, 1998). On the other hand, these results could also stem from idiosyncratic characteristics of the TIMSS database, such as characteristics of the samples or the assessments.

In order to explore these unexpected findings, we carried out parallel analyses of mathematics in TIMSS and the eighth-grade (base year) sample of the National Education Longitudinal Study (NELS-88). Several simplifications were needed to make the models comparable in the two databases, and even with these simplifications, one difference remained between the two. First, the "number of books in the home" variable in TIMSS had five categories: 0-10, 11-25, 26-100, 100-199, and 200+. NELS, however, employed a "50 or more books" dummy variable, so there was no way to collapse the TIMSS variable to be exactly comparable. Based on the frequencies of the responses, we dichotomized the TIMSS variable at 26 books to be most comparable to NELS, even though a variable split at 101 books would likely

have been a more powerful predictor.[11] Second, we collapsed the mother's education variable into four categories: 1=did not finish high school; 2=graduated high school; 3=less than four years of college; and 4=graduated college. This was entered as a single variable. Third, NELS does not have the same press variables as TIMSS, so we deleted that variable from our models. Finally, mathematics scores were standardized to a distribution with mean 0 and variance 1 to allow comparison of parameter estimates across the two databases. This makes parameter estimates from these models different in scale from those in the TIMSS models presented above but would not affect their significance or relative size.

These changes affected the results of the 1 CPS TIMSS model, making mother's education a (barely) significant predictor of mathematics scores at the classroom level, but not at the within-classroom level (Table 12). We can speculate that the stronger partial relationship between scores and mother's education in this model, relative to our final TIMSS model, stemmed from the deletion of the press variable and the dichotomization of the number-of-books variable. Within classrooms, the only significant predictors were age and number of books present. In the TIMSS model reported above, both of these variables and the combined press variable had significant effects within classrooms. At the between-classroom level, the mean of mother's educational attainment had a significant effect, $t = 2.2$. At that level, all of the other variables except for the proportion of students with computers in the home were statistically significant. In the final TIMSS model above, all of these variables, including the proportion with computers, were significant, and most of the t values were similar in the two models.

---

11   When weighted, 89 percent of the NELS sample responded that they have 50 or more books at home. In the TIMSS grade 8 sample, 79 percent responded that they had 26 or more books at home, and 51 percent responded that they had 101 or more. The TIMSS and NELS responses to these variables are quite different. The TIMSS results indicate that somewhere between 21 and 49 percent of students have fewer than 50 books at home, compared to the 11 percent who gave that response in NELS. This could reflect the effects of random sampling error, systematic differences between the samples, or cohort changes.

Table 12.—Two-Level Model of Math, TIMSS Grade 8, 1 CPS, Modified for Comparison to NELS: 1995

Variable	Effect	SE	t	p	CILB[a]	CIUB[b]
Intercept	-7.02	2.20	-3.19	0.00	-11.44	-2.60
Within class ($\updownarrow$)						
Father present	0.04	0.04	1.10	0.28	-0.04	0.13
Number of books	0.26	0.04	5.84	0.00	0.17	0.34
Computer present	0.07	0.04	1.55	0.13	-0.02	0.16
Mother's education	0.02	0.02	0.96	0.34	-0.02	0.06
Age	-0.18	0.04	-4.71	0.00	-0.25	-0.10
Age^2	-0.06	0.04	-1.28	0.21	-0.15	0.03
Between class ($\updownarrow$)						
M Father present	1.10	0.23	4.74	0.00	0.63	1.57
M Number of books	1.49	0.31	4.88	0.00	0.88	2.10
M Computer present	0.38	0.25	1.48	0.15	-0.13	0.89
M Mother's education	0.24	0.11	2.19	0.03	0.02	0.47
M Age	0.32	0.15	2.10	0.04	0.02	0.63
M Age^2	-1.63	0.37	-4.44	0.00	-2.36	-0.89
Residual variances						
$\_^2$ (within)	0.58					
$\_$ (between)	0.10					

[a] Lower bound of 95% confidence interval around parameter estimate
[b] Upper bound of 95% confidence interval around parameter estimate
NOTE: All estimates of error and significance reflect jackknifed estimates.

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

The second step in comparing TIMSS to NELS was to rerun the analysis using schools rather than classrooms as the level-2 unit. NELS, unlike TIMSS, sampled students randomly within schools, rather than sampling intact classrooms. Therefore, to make the TIMSS findings reported in this section more comparable to NELS, the model was re-run with the full Grade 8 TIMSS sample, which comprised 2 classrooms in most schools but a single class in some and more than two in a few.[12] This full TIMSS sample is not entirely comparable to the NELS sample – in schools in which there are more classrooms than were sampled, students in the TIMSS sample will be more clustered than those in NELS – but it is more comparable to NELS

---

[12]   This model used the regular TIMSS sample weights, while the analyses of the 1 CPS sample used our modified weights.

than our standard, one-class-per-school sample. In this full TIMSS sample, as in NELS, the level-1 or within-unit portion of the model is within schools, rather than within classrooms, and the level-2 portion is between schools rather than between classrooms.

Including the full TIMSS sample of classrooms and using schools as the level-2 unit substantially increased the statistical significance of several within-unit effects (Table 13). Most important for present purposes, the estimated effect of each increase of one level in mother's education (on our four-point scale) changed from .02 SD to .07 SD from the 1-CPS to the full-sample model, and the latter estimate was clearly significant. Similarly, in the 1 CPS analysis (Table 12), the estimated within-class effect of having a computer was .07 SD and was nonsignificant. In the full sample (Table 13), the estimated within-school effect was .12 SD and was highly significant (t = 4.4).

It is not surprising that some of the within-unit effects were larger when the full TIMSS sample was used. In the full sample, the within-unit variance includes the variance between classrooms within schools, which is part of the between-unit variance in the 1 CPS models. If there is any differentiation among classes within schools in terms of performance – whether by explicit tracking or by more informal mechanisms – this variance is included in the between-unit level in the 1 CPS analysis but in the within-unit level when the full TIMSS sample is used.

Table 13.—Two-Level Model of Math, TIMSS Grade 8, All Classrooms, Modified for Comparison to NELS: 1995

Variable	Effect	SE	t	p	CILB[a]	CIUB[b]
Intercept	-7.81	3.16	-2.47	0.02	-14.15	-1.46
Within school ($\updownarrow$)						
Father present	0.09	0.02	3.50	0.00	0.04	0.13
Number of books	0.29	0.04	8.02	0.00	0.21	0.36
Computer present	0.12	0.03	4.43	0.00	0.07	0.18
Mother's education	0.07	0.02	3.95	0.00	0.03	0.10
Age	-0.24	0.02	-11.27	0.00	-0.28	-0.20
Age^2	-0.08	0.03	-2.57	0.01	-0.15	-0.02
Between school ($\updownarrow$)						
M Father present	1.27	0.28	4.55	0.00	0.71	1.83
M Number of books	1.42	0.46	3.08	0.00	0.49	2.35
M Computer present	0.46	0.35	1.32	0.19	-0.24	1.16
M Mother's education	0.16	0.15	1.09	0.28	-0.14	0.46
M Age	0.37	0.22	1.68	0.10	-0.07	0.82
M Age^2	-0.94	0.29	-3.21	0.00	-1.53	-0.35
Residual variances						
σ^2 (within)	0.67					
τ (between)	0.08					

[a] Lower bound of 95% confidence interval around parameter estimate
[b] Upper bound of 95% confidence interval around parameter estimate
NOTE: All estimates of error and significance reflect jackknifed estimates.

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

The estimated between-unit effects changed less when the full TIMSS sample was used instead of the 1 CPS sample. The between-unit effects of mean mother's education dropped and became nonsignificant, however, when all classrooms were included. In the 1-CPS analysis, a one-unit change in mean mother's education predicted an increase of .24 SD in scores (Table 12), while in the full TIMSS sample, a one-unit change in mean mother's education predicted a nonsignificant increase of .16 SD (Table 13).

A comparable model of the NELS data produced within-school results that were quite similar to those of the school-level full-sample TIMSS model except for mother's education and age. The estimated within-school effect of a one-step increase in mother's education was more than twice as large in NELS (Table 14) as in the full TIMSS sample (Table 13). The estimate of the impact of mother's education also had a smaller standard error in NELS than in TIMSS. The result of the larger parameter estimate and smaller standard error was a much larger t statistic in NELS than in TIMSS (13 in NELS, vs. 4 in TIMSS), indicating that we can have much more confidence that the parameter in NELS is non-zero. The within-school effect of age was also much larger in NELS.

Table 14.—Two-Level Model of Mathematics, NELS Grade 8: 1988

Variable	Effect	SE	t	p	CILB[a]	CIUB[b]
Intercept	-0.33	1.02	-0.33	0.75	-2.42	1.75
Within school ($\updownarrow$)						
Father present	0.09	0.01	6.65	0.00	0.06	0.11
Number of books	0.22	0.03	8.09	0.00	0.17	0.28
Computer present	0.15	0.02	8.53	0.00	0.12	0.19
Mother's education	0.16	0.01	12.67	0.00	0.13	0.18
Age	-0.34	0.02	-20.49	0.00	-0.37	-0.30
Age^2	-0.06	0.01	-7.01	0.00	-0.08	-0.04
Between school ($\updownarrow$)						
M Father present	0.69	0.08	9.00	0.00	0.53	0.85
M Number of books	0.56	0.12	4.72	0.00	0.32	0.81
M Computer present	0.33	0.08	4.16	0.00	0.17	0.49
M Mother's education	0.43	0.03	16.41	0.00	0.38	0.49
M Age	-0.13	0.07	-1.73	0.09	-0.28	0.02
M Age^2	-0.31	0.08	-3.96	0.00	-0.47	-0.15
Residual variances						
$\_^2$ (within)	0.70					
$\_$ (between)	0.07					

[a] Lower bound of 95% confidence interval around parameter estimate
[b] Upper bound of 95% confidence interval around parameter estimate
NOTE: All estimates of error and significance reflect jackknifed estimates.

Source: National Education Longitudinal Study (NELS-88), 1988, National Center for Education Statistics, U.S. Department of Education.

The difference between NELS and TIMSS in the estimated between-school effect of mother's education was even greater. Recall that in the TIMSS full sample, a one-unit increase in mean mother's education had a nonsignificant effect of .16 SD on scores. In NELS, the estimated effect of a one-unit increase in mean mother's education was nearly three times as large: .43 SD (t = 16). Conversely, the estimated between-school effects of mean age, percent father present, and mean number of books were all markedly smaller in NELS.

These differences between NELS and TIMSS are large enough to be important for some uses and interpretations of the data. They could stem from any number of factors. The TIMSS and NELS performance measures are quite different. They do not have identical content, and the TIMSS measures, unlike the NELS measures, are constructed using an adaptation of the plausible values methodology used in NAEP. Differences in the reliability of the outcome measures would affect the parameter estimates for background factors, but one would expect

these effects would be in the same direction for all background factors, and they were not in this case. Differences in content, however, could conceivably have different effects on different parameter estimates. The differences shown here could also stem from differences in sampling, period effects (the TIMSS Population 2 data were collected roughly seven years later than the NELS baseline data), or response biases (stemming from differences in context or in the operationalization of the background variables).

Decomposing Performance Variation in France

We followed the same procedures in analyzing data from France, but the extent of missing data about parental education complicated both analysis and interpretation. We were forced to chose between a more inclusive model (including parental education) in a reduced sample or a reduced model in a more inclusive sample. We opted for the more inclusive model in the reduced sample that includes only students who provided informative answers to the questions about parental education.

Although the data do not entirely clarify the effects of this choice, analysis of the two samples suggest that our basic findings about the prediction of score variance at each level would not have been fundamentally changed by analyzing a reduced model in the more inclusive sample. Some specific parameter estimates might have been affected, however, but we place less emphasis on specific parameter estimates in our interpretation. The effects of the sample and model differences are discussed in Appendix C.

Bivariate Relationships

As in the U.S., we compared three sets of correlations between scores and background variables: simple student-level correlations, student-level correlations centered on classroom means, and classroom-level correlations.

The uncentered student-level correlations were generally very small (Table 15), and most were smaller than the corresponding correlations in the U.S. (Table 7). Mathematics scores showed weaker correlations in France than in the U.S. with all background variables other than age, and the correlations with both the number of books (r=.14) and computer in the home (r=.10) were much weaker. Correlations between scores and mother's and father's education were somewhat smaller than those in the U.S. Correlations among the background variables themselves were also generally smaller in France than in the U.S. In France as in the U.S.,

centering students' scores around their classroom means reduced most of these already small correlations appreciably (Table 16).

Table 15.—Student-level correlations between background variables and scores, France, uncentered: 1995

	Father present	Number of books	Computer present	Press	Mother's education	Father's education	Age	Math score
Father present	1							
Number of books	0.04	1						
Computer present	-0.01	0.24	1					
Press	-0.01	0.03	0.07	1				
Mother's education	0.00	0.38	0.24	0.09	1			
Father's education	-0.04	0.36	0.26	0.09	0.70	1		
Age	-0.07	-0.19	-0.08	-0.11	-0.24	-0.23	1	
Math score	0.09	0.14	0.10	0.12	0.22	0.18	-0.31	1

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

Table 16.—Student-level correlations between background variables and scores, France, centered: 1995

	Father present	Number of books	Computer present	Press	Mother's education	Father's education	Age	Math score
Father present	1							
Number of books	0.07	1						
Computer present	-0.01	0.21	1					
Press	-0.01	0.03	0.06	1				
Mother's education	0.00	0.26	0.18	0.09	1			
Father's education	-0.04	0.24	0.20	0.09	0.62	1		
Age	-0.08	-0.11	-0.03	-0.08	-0.13	-0.11	1	
Math score	0.06	0.04	0.07	0.08	0.10	0.06	-0.19	1

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

At the between-classroom level as well, most of the correlations between background factors and scores were lower in France (Table 17) than in the U.S. The negative correlation between scores and mean age, however, was much stronger in France, and the correlation between press in mathematics and scores was roughly the same in both countries. In France, mean mother's and father's education showed somewhat lower correlations with scores, somewhat lower correlations with computer present, and much smaller correlations with father present. This suggests that at the aggregate level, the proportion of students with fathers present is less of an SES proxy in France than in the U.S.

Multilevel Model

The final two-level model of mathematics scores in the French sample included only four variables at each level. Three of the variables included in the U.S. model—father present, age (and age^2), and the composite press variable—were included in the French model as well (Table 18). Mother's education, which was not significant in the U.S. models, did have a significant effect in France, despite the smaller bivariate correlation shown by this variable in France. Two of the variables that were significant predictors in the U.S.—computer present and number of books—were not significant predictors in France.

Table 17.—Classroom-level correlations between background variables and scores, France: 1995

	Father present	Number of books	Computer present	Press	Mother's education	Father's education	Age	Math score
Father present	1							
Number of books	-0.13	1						
Computer present	-0.10	0.44	1					
Press	-0.14	0.05	0.02	1				
Mother's education	-0.03	0.71	0.54	0.05	1			
Father's education	-0.11	0.67	0.53	0.10	0.86	1		
Age	0.03	-0.37	-0.30	-0.17	-0.51	-0.49	1	
Math score	0.10	0.48	0.29	0.26	0.50	0.45	-0.55	1

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

Table 18.—Two-Level Model of Mathematics Scores, France Grade 8: 1995

Variable	Effect	SE	t	p	CILB[a]	CIUB[b]
Intercept	592.6	151.7	3.9	0.00	289.8	895.4
Within class ($\updownarrow$)						
Mother's education	4.6	1.7	2.7	0.01	1.2	8.0
Father present	8.9	4.0	2.2	0.03	0.9	17.0
Press	8.6	4.1	2.1	0.04	0.5	16.7
Age	-18.2	3.2	-5.6	0.00	-24.6	-11.7
Age^2	-0.6	2.7	-0.2	0.81	-6.0	4.7
Between-class ($\Updownarrow$)						
M Mother's education	26.4	5.1	5.2	0.00	16.2	36.7
M Father present	59.5	22.8	2.6	0.01	14.0	104.9
M Press	45.0	15.4	2.9	0.00	14.3	75.7
M Age	-23.0	10.1	-2.3	0.03	-43.1	-2.8
M Age^2	-23.3	15.7	-1.5	0.14	-54.7	8.0
Residual variances						
σ^2 (within)	4040.8					
τ (between)	554.7					

[a] Lower bound of 95% confidence interval around parameter estimate
[b] Upper bound of 95% confidence interval around parameter estimate
NOTE: All estimates of error and significance reflect jackknifed estimates.

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

The estimated within-classroom effect of the father being present was much larger in France than in the U.S., where this variable was included in the final model only because it was a

significant predictor at the between-classroom level. In France, the model estimated that within classrooms, students with fathers present will score about 9 points higher than those without, which is a difference of 12 percent of the French standard deviation of 75 points. (Recall that the standard deviation of performance is considerably smaller in France than in the U.S.) The effects of mother's education are somewhat larger. The model estimates that students whose mothers are in the highest category will score on average 18 points, or one-fourth of a standard deviation, above their classmates with mothers in the lowest category. The estimated within-classroom effects of press and age were similar in the two countries, except that the negative effect of age was not curvilinear in France. In France as in the U.S., the effect of age was negative.

At the between-classroom level, all of the predictors included in the French model other than the quadratic term for age were significant. Although the within-classroom effect of father present was much larger in France than in the U.S., the between-classroom effect of the proportion of fathers present was somewhat smaller in France than in the U.S. The effect of press was similar in the two countries. The effect of mean mother's education was sizable.

The predictive power of the model of eighth-grade mathematics in France (Table 19) is in some respects quite similar to the corresponding findings in the U.S. (Table 11). In France as in the U.S., the model predicted much of the between-classroom variance in mathematics, although somewhat less in France: 59 percent in France, compared to 77 percent in the U.S. In both countries, the model predicted very little of the within-classroom variance: 5 percent in France and 4 percent in the U.S.

Table 19.—Total and Predicted Variance in Mathematics Scores at Each Level, France Grade 8: 1995

	Between classroom	Within classroom
Total variance at level	1356	4232
Percent of variance at level	24	76
Variance predicted by variables at level	801	191
Percent of variance at level predicted by variables at level	59	5
Percent of total variance predicted by variables at level	14	3

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

In another sense, however, the between-classroom predictors were much less powerful in France than in the U.S. The predicted between-classroom variance is much smaller in France (801) than in the U.S. (2,532). The percent of between-classroom variance predicted by the model is nonetheless almost as high in France because the total between-classroom variance is much smaller in France (1,356) than in the U.S. (3,299). For this reason, the between-classroom variables predict much less of the total variance in France (14 percent) than in the U.S. (31 percent), even though the total score variance is considerably smaller in France.

Decomposing Performance Variation in Hong Kong

Bivariate Relationships

Student-level (uncentered) correlations showed generally very weak relationships between background variables and scores in Hong Kong (Table 20). The highest correlations between scores and background variables were those with press, and those were below .20. Except for the correlation between mother's and father's education, the correlations among background variables were generally small. An exception was the negative correlation (-.46) between born in country and age. This was the only clue in the correlation matrix about the born in country variable, which had substantial negative effects in the models described below. As in the U.S. and France, centering the student-level correlations reduced many of them (Table 21).

Table 20.—Student-level correlations between background variables and scores, Hong Kong, uncentered: 1995

	Father present	Number of books	Computer present	Press	Born in country	Mother's education	Father's education	Age	Math score
Father present	1								
Number of books	-0.02	1							
Computer present	0.04	0.24	1						
Press	0.02	0.05	0.07	1					
Born in country	0.03	0.05	0.08	0.00	1				
Mother's education	0.02	0.22	0.21	0.04	-0.01	1			
Father's education	-0.01	0.26	0.26	0.08	-0.05	0.60	1		
Age	-0.04	-0.04	-0.07	-0.05	-0.46	-0.06	-0.01	1	
Math score	0.03	0.14	0.13	0.18	-0.05	0.14	0.15	-0.04	1

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

Table 21.—Student-level correlations between background variables and scores, Hong Kong, centered: 1995

	Father present	Number of books	Computer present	Press	Born in country	Mother's education	Father's education	Age	Math score
Father present	1								
Number of books	-0.03	1							
Computer present	0.02	0.18	1						
Press	0.01	0.02	0.02	1					
Born in country	0.02	0.04	0.09	0.00	1				
Mother's education	0.01	0.14	0.13	0.00	0.01	1			
Father's education	-0.01	0.17	0.17	0.04	-0.04	0.54	1		
Age	-0.02	-0.02	-0.05	-0.03	-0.43	-0.06	-0.01	1	
Math score	-0.03	0.00	-0.03	0.07	-0.08	-0.02	0.00	0.03	1

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

Some of the correlations among classroom means, however, were substantial (Table 22), and the patterns among these correlations were somewhat different from those in the U.S. or France. Mean scores showed correlations of more than .50 with six of nine background variables—all but percent father present (.39), age (-.28), and born in country (roughly zero). At the classroom level, the negative correlation between age and born in country increased to -.75. Numerous of the other correlations between background variables were also sizable, including number of books and computers present (.69), computer present and press (.46), press and mother's and father's education (both about .4), and mother's and father's education (.90).

Table 22.—Classroom-level correlations between background variables and scores, Hong Kong: 1995

	Father present	Number of books	Computer present	Press	Born in country	Mother's education	Father's education	Age
Father present	1							
Number of books	0.18	1						
Computer present	0.26	0.69	1					
Press	0.16	0.30	0.46	1				
Born in country	0.23	0.12	0.08	-0.05	1			
Mother's education	0.14	0.70	0.74	0.44	-0.11	1		
Father's education	0.20	0.74	0.78	0.40	-0.08	0.90	1	
Age	-0.36	-0.20	-0.25	-0.19	-0.75	-0.03	-0.04	1
Math score	0.39	0.58	0.66	0.64	-0.02	0.58	0.56	-0.28

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA) Multilevel Model

The final model in Hong Kong was similar to that in the U.S. in terms of selection of variables and predictive power, but it was quite different in some specific details.

The selection of variables in the Hong Kong model differed from that in the U.S. in two respects. Born in country remained in the Hong Kong model, while age did not.

In Hong Kong, only two variables showed significant relationships to scores within classrooms: press, which was positively related to scores, and the 'born in country' dichotomy, which in this case was *negatively* related to scores (Table 23). The within-classroom coefficients for father present and computer present were also negative but were not statistically different from zero.

Table 23.—Two-Level Model of Mathematics Scores, Hong Kong Grade 8: 1995

Variable	Effect	SE	t	p	CILB[a]	CIUB[b]
Intercept	-424.8	140.6	-3.0	0.00	-708.6	-141.0
Within class ($\updownarrow$)						
Number of books	0.3	1.2	0.3	0.78	-2.0	2.7
Computer present	-3.8	3.1	-1.2	0.23	-10.0	2.5
Father present	-7.4	6.1	-1.2	0.24	-19.7	5.0
Press	10.3	2.8	3.7	0.00	4.6	16.0
Born in country	-19.1	5.4	-3.5	0.00	-30.1	-8.2
Between-class ($\updownarrow$)						
M Number of books	44.1	16.1	2.7	0.01	11.7	76.6
M Computer present	89.8	41.8	2.1	0.04	5.4	174.1
M Father present	326.9	86.8	3.8	0.00	151.8	502.1
M Press	174.5	35.1	5.0	0.00	103.5	245.4
M Born in country	-44.7	37.2	-1.2	0.24	-119.9	30.4
Residual variances						
σ^2 (within)	5485.0					
τ (between)	1406.2					

[a] Lower bound of 95% confidence interval around parameter estimate
[b] Upper bound of 95% confidence interval around parameter estimate
NOTE: All estimates of error and significance reflect jackknifed estimates.

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

The between-classroom results in Hong Kong were markedly different from those in the corresponding U.S. model. The coefficient for the proportion of fathers present was huge: 327, vs. 90 in the U.S. The range of percent father present, however, was far more restricted in Hong Kong than in the U.S., ranging in this sample only from 76 to 100 percent. Nonetheless, these results thus suggest that if all other variables in the model were held constant, the mean score in a classroom with 76 percent of fathers present would be 78 points (0.79 SD) lower than that in a classroom with 100 percent of fathers present. The estimated effect of the mean press variable was also very large: 174, about four times as large as in the U.S. The effect of the percent having computers was almost two and a half times as large as in the U.S. model. The effect of the mean number of books was similar in the two countries.

Despite these striking differences in the effects of some aggregate variables, the Hong Kong mathematics model was somewhat similar to the U.S. model in predictive power, showing the by now familiar pattern of strong prediction between classrooms and weak prediction within

classrooms. The Hong Kong model was notably weaker in terms of within-classroom prediction than were the models in the U.S. and France, but the model predicted little of the within-classroom variance in any of the three countries. The percentage of total variance within classrooms was similar in Hong Kong and the U.S. (between 55 and 60 percent in both), but the Hong Kong model predicted only 1 percent of this within-classroom variance (Table 24). The between-classroom variance was somewhat greater in Hong Kong than in the U.S., but the model predicted nearly as large a percentage of it: 69 percent in Hong Kong, compared to 77 percent in the U.S. mathematics model. In Hong Kong, as in the U.S., the between-classroom variables predicted 31 percent of the total score variance.

Table 24.—Total and Predicted Variance in Mathematics Scores at Each Level, Hong Kong Grade 8: 1995

	Between classroom	Within classroom
Total variance at level	4543	5557
Percent of variance at level	45	55
Variance predicted by variables at level	3137	73
Percent of variance at level predicted by variables at level	69	1
Percent of total variance predicted by variables at level	31	1

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

Decomposing Performance Variation in Korea

Bivariate Relationships

      The simple (uncentered) student-level correlations in Korea among background variables and between those variables and scores were not strikingly different from those in the other countries. The strongest correlation between math scores and any of the background factors was with number of books (r = .34; Table 25). Born in country and age showed essentially no relationship with scores, and the other background variables correlated between .18 and .27 with scores. The correlation between mother's and father's education, .72, was similar to that in France and larger than that in the U.S. or Hong Kong. The other background variables showed modest intercorrelations, none greater than .34 and most considerably lower.

Table 25.—Student-level correlations between background variables and scores, Korea, uncentered: 1995

	Father present	Number of books	Computer present	Press	Born in country	Mother's education	Father's education	Age
Father present	1							
Number of books	0.12	1						
Computer present	0.04	0.22	1					
Press	0.08	0.22	0.13	1				
Born in country	0.00	0.07	-0.01	0.01	1			
Mother's education	0.08	0.26	0.23	0.16	0.00	1		
Father's education	0.07	0.34	0.25	0.17	0.02	0.72	1	
Age	-0.02	-0.05	-0.02	-0.04	-0.04	-0.08	-0.07	1
Math score	0.06	0.34	0.18	0.27	0.04	0.22	0.27	-0.03

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

The student-level correlations centered around classroom means, however, were appreciably different in Korea than in the other countries. Because very little of the performance variance in Korea lies between classrooms, removing that variance by centering shrank the correlations less in Korea than in the other countries.

Thus, for example, the centered correlation between press and scores was .25 (Table 26), only trivially less than the uncentered correlation above. The correlations between the two parental education variables and scores shrank noticeably when centered, but they remained considerably larger than the corresponding centered correlations in the U.S., France, or Hong Kong. The correlation between scores and number of books in the home shrank modestly.

Correlations among classroom means in Korea were within the range of those found in the other three countries. The correlations between means of the parental education variables and mean scores were .63 and .69 (Table 27), roughly comparable to those in Hong Kong and the U.S. The correlation between mean scores and the press variable, .50, was stronger than in the U.S. and France but weaker than in Hong Kong. The correlations between mean scores and the means of the possessions-in-home variables were within the range of those in other countries. The correlations between mean scores and the proportion of fathers present was markedly lower than in Hong Kong or especially the U.S. but larger than in France.

Table 26.—Student-level correlations between background variables and scores, Korea, centered: 1995

	Father present	Number of books	Computer present	Press	Born in country	Mother's education	Father's education	Age	Math score
Father present	1								
Number of books	0.10	1							
Computer present	0.02	0.17	1						
Press	0.06	0.19	0.11	1					
Born in country	0.00	0.07	-0.01	0.01	1				
Mother's education	0.07	0.18	0.15	0.13	0.00	1			
Father's education	0.05	0.26	0.16	0.14	0.03	0.65	1		
Age	-0.02	-0.04	-0.02	-0.04	-0.03	-0.07	-0.06	1	
Math score	0.05	0.30	0.12	0.25	0.05	0.15	0.19	-0.03	1

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

Table 27.—Classroom-level correlations between background variables and scores, Korea: 1995

	Father present	Number of books	Computer present	Press	Born in country	Mother's education	Father's education	Age	Math score
Father present	1								
Number of books	0.29	1							
Computer present	0.26	0.56	1						
Press	0.21	0.43	0.34	1					
Born in country	0.09	0.06	-0.04	-0.06	1				
Mother's education	0.17	0.66	0.62	0.41	0.00	1			
Father's education	0.21	0.71	0.66	0.45	-0.05	0.91	1		
Age	-0.03	-0.16	-0.07	-0.04	-0.11	-0.17	-0.19	1	
Math score	0.26	0.66	0.63	0.50	-0.02	0.63	0.69	-0.06	1

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

Multilevel Model

The results of the two-level model of mathematics scores in Korea differed in one fundamental respect from those in the other three countries: the prediction of within-classroom variance was stronger in Korea than elsewhere.

The selection of variables of the final Korean model included number of books, computer present, press, age and age squared, and father's education. It differed from the U.S. model in excluding father present and including father's education, which was included only because of a significant coefficient at the between-classroom level.

With the exception of age, all of the variables included in both the U.S. and Korean models had much larger within-classroom coefficients in the Korean model. For example, the within-classroom coefficient for the number of books in the home was about 2.5 times as large in Korea as in the U.S., and the t was about twice as large (Table 28; compare Table 10). The within-classroom coefficient for computer present was also about 2.5 times as large in Korea as in the U.S., and the coefficient for press was almost 4 times as large. The Korean coefficients were also much larger than those in the French and Hong Kong models, to the extent that the same variables were included.

Table 28:—Two-Level Model of Mathematics Scores, Korea Grade 8: 1995

Variable	Effect	SE	t	p	CILB[a]	CIUB[b]
Intercept	27.9	345.5	0.1	0.94	-660.5	716.3
Within class ($\updownarrow$)						
Number of Books	20.2	1.8	11.5	0.00	16.7	23.7
Computer present	10.9	3.9	2.8	0.01	3.2	18.7
Press	36.2	4.2	8.7	0.00	27.9	44.5
Father's education	9.5	2.0	4.7	0.00	5.5	13.5
Age	-6.0	6.2	-1.0	0.34	-18.4	6.4
Age^2	-14.8	5.6	-2.7	0.01	-26.0	-3.7
Between-class ($\updownarrow$)						
M Books	16.2	7.3	2.2	0.03	1.7	30.7
M Computer	44.5	16.0	2.8	0.01	12.5	76.4
M Press	47.4	17.2	2.8	0.01	13.1	81.7
M Father's education	18.8	5.6	3.4	0.00	7.7	30.0
M Age	20.5	24.1	0.8	0.40	-27.6	68.5
M Age^2	-26.2	11.9	-2.2	0.03	-50.0	-2.4
Residual variances						
σ^2 (within)	9290.6					
τ (between)	48.0					

[a] Lower bound of 95% confidence interval around parameter estimate
[b] Upper bound of 95% confidence interval around parameter estimate
NOTE: All estimates of error and significance reflect jackknifed estimates.

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

In contrast, the between-classroom effects of background variables were not especially large in Korea. For example, the between-classroom coefficient for press was similar in Korea and in the U.S. The estimate for proportion with computers present was slightly larger in Korea, but the estimate for mean number of books was much smaller in Korea (16) than in the U.S. (46).

The relatively strong level of within-classroom prediction in Korea is clearer when one looks at the decomposition of total and predicted variance (Table 29). The within-classroom variables predict 13.4 percent of the within-classroom score variance. This compares to 4 percent in the U.S. (Table 11). The within-classroom variance is far larger in absolute terms in Korea than in the U.S., however, and considerably larger as a percent of total variance (93 percent in Korea, vs. 59 percent in the U.S.). Therefore, the within-classroom variables predict a much larger percentage of the total score variance in Korea (12.4 percent) than in the U.S. (2 percent).

Table 29.—Total and Predicted Variance in Mathematics Scores at Each Level, Korea Grade 8: 1995

	Between classroom	Within classroom
Total variance at level	799	10722
Percent of variance at level	7	93
Variance predicted by variables at level	751	1431
Percent of variance at level predicted by variables at level	94	13
Percent of total variance predicted by variables at level	7	12

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

The between-classroom portion of the model predicted nearly all (94 percent) of the between-classroom variance in scores in Korea (Table 29), far more than in our other three countries. The total between-classroom variance is very small in Korea, however. Therefore, the between-classroom part of the model predicts relatively little of the total mathematics score variance, only 6.5 percent. In contrast, the between-classrooms part of the model predicts 31 percent of the total mathematics score variance in the U.S. and Hong Kong and 14 percent in France.

Comparing the Multilevel Models Across Countries

Perhaps the single most striking consistency in the results reported here is the models' prediction of most of the performance variance between classrooms but little of the variance within classrooms. There are important variations in the predictive power of the models, discussed below, but this general pattern held true in all four countries. In all of our samples, our model predicted at least 59 percent of the between-classroom variance (see Table 30). The prediction of within-classroom variance was modest in Korea and very weak in all other countries.

Table 30.—Percent of Variance at Each Level Predicted by Final Models, Grade 8 Mathematics: 1995

	Between Classroom	Within Classroom
United States	77	4
France	59	5
Hong Kong	69	1
Korea	94	13

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 U.S. Similarly, one of the German TIMSS National Research Coordinators indicated that both state (Land) and inner-city location are correlated with performance in Germany (Baumert, 1998). 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 31; compare Table 30). This finding is an additional reason to suspect that the differences in selection of variables between our four countries and the specific parameter estimates should be interpreted with caution. The ideal models that included the variables that actually determine performance variation could look quite different – that is, they could include different variables and have somewhat different estimates of the parameters already in our models – even though these better models might not predict a great deal more of the variation in performance.

Table 31.—Percent of Variance at Each Level Predicted by Fixed Model, Grade 8 Mathematics: 1995

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

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

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 U.S. 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 32). The models predict much less of the variance in France (18 percent) and Korea (19 percent).

Table 32.—Percent of Total Variance Predicted by Predictors at Each Level, Final Models, Grade 8 Mathematics: 1995

	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.

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

The four countries also differ in terms of the relative predictive power of the models between the student and classroom levels. Again, the U.S. 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 32). 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 U.S. 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 (Table 30). Second, a larger percentage of the total variance lies within classrooms in Korea (93 percent) than in France (76 percent), the U.S. (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.

To explore whether stronger prediction of scores in some countries simply reflects greater variance in the predictors, we took two further steps with our fixed model, which removes the effects of model differences across countries. First, we compared predicted amounts of variance, rather than predicted percentages, from one country to another. We also obtained estimates of the variability of the predictors themselves, decomposed into within- and between-classroom components. We then compared these estimates to see whether differences among countries in the amount of predicted score variance were paralleled by differences in the amount of variance in the predictors themselves.

The predicted variances from the fixed model (Table 33) are largely consistent with the percentages from the final models. The U.S. and Hong Kong are quite similar in terms of predicted variances, except that the predicted within-classroom variance is even smaller in Hong Kong (Table 33). France is similar to the U.S. in having little within-classroom variance predicted, but there is much less predicted between-classroom variance in France.[13] Korea has a small amount of predicted between-classroom variance, but there is not much between-classroom variance to predict; the model predicts most of what variance there is. The more interesting finding in Korea is that the predicted within-classroom variance is far larger there than in the other three countries.

Table 33.—Variance Predicted by Predictors at Each Level,
Fixed Model, Grade 8 Mathematics: 1995

	Between Classroom	Within Classroom
United States	2389	197
France	733	179
Hong Kong	3028	51
Korea	686	1339

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

---

[13] Recall that the total score variance is markedly smaller in France than in the other three countries. That is why the smaller percentage predicted translates into a much smaller amount of predicted variance.

We then decomposed the variance of the predictors in the fixed model into within- and between-classroom components within each country. Because these variables are somewhat collinear, and the collinearity is not constant across countries, looking at one variable at a time is not fully sufficient to ascertain the impact of predictor variance on the country differences in prediction of score variance. However, given that the collinearity among these variables is not very high in these data, this should provide a reasonable if imprecise view.

Differences in predictor variance appear not to account for the much larger predicted score variance within classrooms in Korea. The within-classroom variance of all predictors other than age and age squared was comparable in the U.S. and Korea (Table 34). Age is more variable in Korean classrooms than in U.S. classrooms, but age was a relatively weak predictor of within-classroom score variance in Korea. This is consistent with the larger within-classroom coefficients in the Korean model, compared to the U.S. model.

The contribution of predictor variance to the difference between France and the U.S. 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.

Table 34.—Between- and Within-Classroom Variance of Predictors used in, Fixed Model, Grade 8 Mathematics: 1995

	USA		France		Hong Kong		Korea	
	Between	Within	Between	Within	Between	Within	Between	Within
Number of Books	0.150	0.760	0.080	0.680	0.110	0.970	0.080	0.870
Age	0.030	0.230	0.180	0.420	0.040	0.490	0.000	0.130
Age^2	0.010	0.130	0.020	0.580	0.070	1.700	0.000	0.130
Computer present	0.040	0.210	0.010	0.240	0.020	0.220	0.020	0.220
Press	0.010	0.220	0.010	0.210	0.020	0.280	0.010	0.280

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

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 U.S. – 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 U.S. in terms of the predictive power of the models both within and between classrooms. Table 34 shows that Hong Kong and the U.S. are also similar in terms of the within- and between-classroom variance of the predictors themselves, with the exception of age.

CONCLUSIONS

We begin by noting implications for policy and future research. We then discuss issues for other secondary analysts and raise issues NCES may wish to consider in the design of future international surveys of achievement.

Implications for Policy and Research

This study was prompted in part by a widespread view that performance variance in the U.S. 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, italics in the original). In other instances, this view of variability is implicit, as when the scores for U.S. states or districts are compared to national averages from other countries. In response, we asked whether the distribution of performance in the U.S. is anomalous, how the variance in performance is distributed in the U.S. and other countries, and how well background factors can predict that variation.

TIMSS suggests strongly that the variation in performance in the U.S. is not anomalous. In Population 2, the U.S. 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 U.S., 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 U.S. 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 U.S., 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 U.S., 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 U.S., Germany, Hong Kong, and Australia.

84

While the U.S. 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 U.S. schools are far more variable than are schools elsewhere. Of the countries we considered, only the U.S. and Australia provided samples that allow one to estimate the variability between schools, and one cannot draw inferences about school variability from the TIMSS data on between-classroom variability. 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 U.S. 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 U.S. 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 U.S. 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 U.S. 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 U.S. 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.

Differences in stratification in terms of ability might contribute to the differences in findings between the U.S. and Korea. This hypothesis is consistent with the differences between the U.S. 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 U.S., 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 U.S. (Table 34). However, more of the variance that predicts student performance may lie within classrooms in Korea. In contrast, in countries like the U.S., 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 U.S. 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 U.S. 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 U.S., 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 U.S. and strikingly different from Korea and Japan. Hong Kong is also similar to the U.S. 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 U.S. 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 the univariate decompositions of predictor variance (Table 34) do 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 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 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.

Implications for Secondary Analysis

Because of its prominence and richness, TIMSS is likely to attract the interest of secondary analysts. Although the methods employed by analysts will vary with their purposes, the results presented here point to a number of suggestions that should be pertinent to a wide range of secondary analysis.

This study clearly indicates the importance of using a multi-level approach in exploring the correlates of performance in TIMSS. Even when analysis is restricted to a single country, a multi-level approach is needed when the relationships among variables vary across levels, as they did in the models presented here. Comparative international studies increase the importance of a multi-level approach, however, because the differences among levels also can vary across countries. A good illustration is the finding that the contrast between within- and between-classroom prediction is markedly different in Korea than in the U.S., France, or Hong Kong.

Exploratory data analysis (EDA) should normally be a precursor to formal modeling of survey databases, and international comparisons make this stage of analysis particularly important. Variables may behave very differently in different countries. A variable that shows good dispersion of responses across categories in one country, for example, may be nearly useless in another because of a concentration of responses in one or two categories. Similarly, a variable that shows predicted relationships with other variables in one country may show unexpected relationships in another country. Anomalies in the data, such as exceptions to monotonic relationships, may also arise only in certain countries. All of these possibilities were illustrated in the results of EDA presented above, and there were additional examples in exploratory analyses that we did not present here. When the analyst expects to use multi-level models, this EDA should be carried out within levels as well as in the total sample.

In some contexts, hierarchical modeling adds seriousness to the conventional warning that statistical significance is no substitute for hypotheses or strong theory as a guide for selecting or having confidence in the analyst's choice of variables. One reason that this caution is underscored in multilevel modeling is that the relationships expected at the level of individual may not hold at the aggregate level, and indeed aggregate relationships may depend on the level of aggregation chosen. For example, our hierarchical analysis of the predictors of mathematics performance in NELS found that results were sensitive to the choice of classroom or school as the second level of analysis. In that case, we have prior knowledge that makes the difference sensible: the presence of tracking in many U.S. schools. In some cases, however, the analyst will have little basis in prior research for predicting or interpreting aggregate relationships.

There are a number of steps an analyst can take in response to this uncertainty. One is to examine the collinearity of variables at the aggregate level to explore the possible effects of different choices of variables. A simple approach is to run an ordinary least squares regression at the aggregate level and examine the tolerance of each variable. The tolerance is $(1-R^2)$ from a regression of a given predictor on all other predictors in the model; it indicates the proportion of variance in that predictor not accounted for by other variables already in the model. Table 35 shows the tolerances from two aggregate OLS models run in the U.S. data (one classroom per school): a full model that contains the nine variables from which we selected those for the final model, and our final model that included six variables. In the full model, four of nine variables have tolerances below .30, and the two parental education variables have tolerances of roughly

.2. This may explain the lack of predictive value of parental education in our models; at the aggregate level, most of the variance in these variables is shared with other predictors. With parental education removed, no tolerances remained below .36. These low tolerances may also help explain why modifications to the final model had relatively small effects on the total fit of the model.

Table 35.—Tolerances in Aggregate Models, U.S. Grade 8 Mathematics: 1995

	Full Model	Final Model
M Computer present	.26	.38
M Father present	.71	.76
M Number books	.26	.36
M Age	.69	.73
M Age^2	.71	.74
M Press	.82	
M Born in Country	.77	.86
M Mother's education	.21	
M Father's education	.22	

Source: Third International Mathematics and Science Study, Population 2 data set, International Association for the Evaluation of Educational Achievement (IEA)

The analyst can also test the sensitivity of the final model fit to alternative selections of variables. If the fit of the model is not sensitive to changes in the choice of variables, it may be wise to limit interpretation of specific parameters. Thus, in this report, we responded to this problem by deemphasizing specific parameter estimates and focusing instead on the apparently more robust fit of the models.

The work presented here also suggests the need for caution in using weights in complex analyses of survey data. Although the logic of sampling weights is straightforward, i.e., to offset intentional or unintended differences in the probability of inclusion in the sample, their use in statistical analysis can be difficult, particularly in the case of complex models. Some statistical packages assume that weights are precision weights rather than sample weights; that is, they assume that weights are inversely proportional to the variance of an observation rather than to its probability of selection. For some purposes, sample weights and precision weights are handled very differently, and applying a precision-weighting algorithm to sample weights can result in incorrect results, such as dramatically inflated variances. Moreover, as shown in Appendix D,

there are complex decisions entailed in applying even sample-weighting algorithms to multilevel analysis.

At the very least, analysts applying multi-level models to complex survey data should follow the simple expedient of standardizing the weights so that the sum of the weights is equal to the sum of observations. As shown in Appendix D, this lessened substantially the distortions that otherwise arose in using certain weighting algorithms in our multi-level models of TIMSS. Moreover, some survey data have mean weights much further from 1 than does TIMSS, and in those cases, the importance of standardizing weights could be even greater. Analysts should be cautioned, however, that even standardizing weights will not always eliminate all substantial distortions, and better approaches to weighting are neither fully developed nor well implemented in many software packages. For simple two-level models, the SAS macro provided in Appendix D officers an alternative approach for the case in which weights are uniform within any given level-2 unit.

One specific finding that warrants note is the relatively weak predictive power of parental education in the U.S. TIMSS sample. This was surprising and was inconsistent with NELS. It may be that this is an anomaly from which little can be learned. Nonetheless, this anomaly underscores the general risk of basing conclusions on a single data source. Findings may be sensitive to both intended and unintended idiosyncracies of any particular database. For example, Wolfe (1997) has shown that TIMSS country means in eighth grade mathematics are sensitive to the weighting of particular content areas within mathematics. He illustrated this with the case of Sweden, which scored poorly on algebra compared to the other TIMSS content areas. In many areas of research, the accumulation of other research findings helps researchers recognize potentially anomalous findings, but in the area of international comparisons, the small number of data sources and published findings makes it difficult to check the robustness of findings across sources.

Implications for the Design and Implementation of Future Surveys

The research presented here encountered numerous difficulties stemming from the design and operationalization of TIMSS. These impediments were of two types. Some were aspects of the TIMSS design that were appropriate for the intended primary purposes of TIMSS but were poorly suited to the analyses we conducted. Others were problems of data quality independent of

TIMSS' primary purpose. Both may hold useful lessons for the design of future international surveys of achievement.

Tradeoffs in Survey Design

Designing a complex survey entails many difficult tradeoffs, and many of the important decisions about design depend on the priority assigned to the survey's potential uses. For example, a simple random sample of students within high schools may be less useful than sampling of intact classrooms for obtaining information about instruction but provides an estimate of the variability of performance (and of background variables) within schools. A sample of a single intact classroom per school sacrifices that information about variability but arguably provides a better basis for obtaining information about instruction and may be less burdensome to participating schools. A complex design with multiple stages, clustering, and substantially unequal sampling probabilities lessens the cost and burden entailed in obtaining certain specific statistics, but as we have shown here, it complicates some forms of secondary analysis. Matrix sampling of test items increases efficiency and therefore permits a richer assessment of groups of students, but it can substantially complicate certain types of analysis and often precludes accurate individual-level scores. Many other similar tradeoffs could be listed.

Our expectation is that TIMSS was designed to meet several goals:

- To provide estimates of means and other summary statistics for countries and for large aggregates within them (e.g., males and females) with reasonable statistical efficiency;

- To provide a broad assessment of mathematics and science within limited testing time;

- To provide rich information about curriculum and instruction; and

- To limit administrative and other burdens within schools, survey costs, and total respondent burden.

The TIMSS design, with matrix sampling, sampling of intact classrooms, the use of a complex sampling design with unequal probabilities of selection, and the emphasis on data pertaining to curriculum and instruction, fits well with these goals.

This design, however, is less well suited to many other uses that are likely to arise in secondary analysis. Our analyses, which used TIMSS data in a manner quite different from its

intended primary purposes, illustrate some of the impediments to secondary analysis imposed by the TIMSS design. Some of the difficulties we encountered may be specific to the analytical questions we addressed, but others are likely to arise in a wide variety of secondary analyses.

In our case, the limitation of the sample to a single mathematics classroom per school in most countries imposed particularly large costs because it precluded separating certain important components of performance variance. In mathematics, it precluded separating variations among classrooms within schools from variations among schools. This design precludes answering important questions about variations in performance – for example, how variable schools are in the countries that participated in TIMSS, and how much of the variation among classrooms can be attributed to within-school differentiation rather than stratification among schools. Analysis of data from the Second International Mathematics Study (Schmidt, Wolfe, and Kifer, 1993) illustrated the importance of this shortcoming in that it showed that the partitioning of aggregate score variance into between-classroom and between-school components varied dramatically among countries.

This aspect of the sampling design imposes even more severe costs on analysis of the science data. The sample of students tested in science in a given school do not represent an intact science class and also cannot be considered a random sample of science students within the school. In some schools, the sampled students may happen to be members of the same science class, while in other schools, they are drawn in varying numbers from numerous science classes. For this reason, one cannot discern what mix of variance components constitutes the level-two variance in science. One could analyze the predictors of performance variation in science only by ignoring the inherently hierarchical structure of the data (i.e., by analyzing the data only at the student level). Our two-level results in mathematics, however, suggest that international comparisons based on single-level models are likely to be misleading.

A second aspect of the TIMSS design that complicates secondary analysis is the multi-stage clustered sample design with non-uniform sampling weights. Sample designs of this sort are common, and their advantages are well known. A design without these attributes would be costlier and more burdensome, and for some purposes (e.g., analysis of curriculum), it might yield inferior data. The costs imposed on analysis by complex sampling weights, however, can be high. The application of sampling weights in simple statistical analyses is well understood, and methods for correctly estimating standard errors from clustered samples are increasingly

available and acknowledged. In the case of more complex or more novel analytical methods, however, such as hierarchical modeling, the use of weights is difficult and not well understood, and software is not well developed in this respect.

Our analyses were also handicapped by limitations of the background variables included in TIMSS. These variables are sufficient to predict much of the variation in student performance between classrooms, although little of the variance within classrooms. However, this is a weak test of their adequacy. Even when specific background variables are demonstrably weak, including enough of them in a multivariate model will sometimes predict almost as large a share of the variance as would better variables (see, e.g., Berends and Koretz, 1996). Moreover, we showed that within limits, the specific choice of variables in our models did not greatly affect the percentage of between-classroom variance predicted. Prediction of variance need not imply meaningful explanation.

In other respects, the TIMSS background variables are insufficient for exploring the correlates of performance variation. As noted, TIMSS does not include some of the background variables that are considered to be substantially associated with student performance in some countries. Moreover, the absence of a parent survey presumably limits the quality of some of the background data collected (see Baratz-Snowden, et al., 1988; Kaufman and Rasinski, 1991), particularly for Populations 1 and 2. At this point, we do not know the extent to which deficiencies in the TIMSS background variable set contributed to the very poor prediction of within-classroom performance variance. It seems likely that the weak within-classroom prediction is at least partly structural, not a function of weak operationalization, and the financial and other costs of obtaining information from parents in an international study could be prohibitive. Nonetheless, in the absence of additional data and research, it is not clear how much analyses of background variables will yield biased findings because of sole reliance on student self-reports.

This report thus illustrates the basic and unavoidable tradeoffs that arise in designing large-scale surveys: designs that suit one purpose well often suit another poorly. Some of the difficulties we encountered stemmed from the differences between our goals and the primary purposes for which TIMSS was designed. Nonetheless, the Department of Education has actively encouraged secondary analysis of the large-scale survey databases it has funded. The questions we addressed were not arcane, and the difficulties we encountered would arise in a

wide variety of secondary analysis. Therefore, if the Department of Education anticipates or wishes to encourage diverse uses of the data from these surveys, it may be productive to weigh the design tradeoffs for various of these uses before finalizing the design of surveys. For example, if explorations of performance variation are likely to be important, a two-classroom-per school design, or a hybrid design in which a single intact class and a random sample from the entire grade are both sampled, might be worth the expense. At the minimum, sampling an intact classroom in each tested subject would likely add greatly to the utility of the data. It is important to note, however, that the costs imposed by any decision are likely to be statistical as well as financial. For example, the statistical efficiency of some estimates may be reduced by design changes intended to facilitate other estimates.

Other Limitations of TIMSS

Several of the obstacles we confronted stemmed from problems of operationalization rather than from the fundamental design of the TIMSS survey, and these could be addressed without weakening – indeed, perhaps strengthening – TIMSS for its primary functions. A number of the steps that could be considered might apply to other complex surveys as well. For example, given the problems with school-level non-response, efforts to obtain rudimentary information about non-participants could help estimate the impact of this problem and perhaps correct for it in analysis. A longer timeline might allow additional and more successful recruitment of replacements. The problem of "I don't know" responses, which was extreme in the case of parental education in France but was substantial in other instances as well, might have been lessened by additional pretesting, perhaps followed by in-depth exploration of the reasons for this response in a small number of cases.

The international nature of TIMSS raised additional problems that suggest the need for additional caution and more extensive pretesting of survey questions. We found numerous instances, some of which are shown above, in which variables behaved substantially differently in different countries. For example, in some countries, the limited distribution of cases across the categories of Likert items lessened the usefulness of the data. In addition, some variables showed response patterns suggesting response bias in some countries but not others. The relationships among variables also differed across countries, sometimes in ways that called the meaningfulness of responses into question. It might be feasible to lessen these problems.

REFERENCES

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

Baumert, J. (1998). personal communication, July.

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.

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

Cleveland, W.S. (1993). *Visualizing Data.* Murray Hill, NJ: AT&T Bell Laboratories.

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.

Graubard, B.I., and Korn, E.L. (1996). Modeling the sample design in the analysis of health surveys. *Statistical Methods in Medical Research*, **5**, 263-282.

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.

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

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.

MathSoft (1998). *S-Plus 4 Guide to Statistics*. Seattle: Author.

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.

Snijders, T.A.B., and Bosker, R.J. (1994). Modeled variance in two-level models. *Sociological Methods & Research, 22*,342-363.

Wolfe, R. . (1997). Country-by-item interactions: Problems with content validity in scaling. In *Validity in Cross-National Assessments: Problems and Pitfalls*. Symposium presented at the annual meeting of the American Educational Research Association, Chicago, April 27.

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 warranted by test of linearity
Father's education	BSBGEDUF	Sometimes recoded as noted in text; sometimes entered as a single variable, if warranted by test of linearity
Born in country	BSBGBRN1	

APPENDIX B. SELECTING A MODEL IN THE U.S.

As explained in the Methods section of the text, we used a "judgmental stepwise" procedure to select the final two-level models in each country. This process involved stepping up from a null model to an inclusive one and then stepping down by eliminating unimportant variables. Although statistical significance was the primary consideration in this process, this process differs from empirical subsets procedures (such as conventional stepwise procedures, backward selection, and forward selection) in that the decisions were judgmental rather than algorithmic.

To illustrate this process, this Appendix includes the series of models used for this purpose in the U.S. These models were run after most of the other key decisions had been made—e.g., decisions to eliminate ill-behaved background variables and to restrict the analysis to a fixed-coefficients model with group-means centering. However, they differ from those shown in the body of the report in several respects. The samples in each of the models in this Appendix comprised all cases with all variables present; thus the counts differ and are smaller in the more inclusive models. (A comparable set was run with all samples restricted to that of the most inclusive model as a check against major effects of these sample differences. Except in the case of France, described in Appendix C, the differences between these sets of runs were generally minor.) These models were also run with SAS Proc Mixed without weighting.

The final model in this series (model 8) was refined for reporting.

The reasons for excluding mother's education (BSBGEDUM) and father's education (BSBGEDUF) can be seen in the weak effects of these variables in model 2 and thereafter. Note that in model 4 we replaced the single ordinal mother's and father's education variables with sets of dummy variables to explore whether the surprising unimportance of these constructs in the model was a function of including them as single variables.

```
                          TIMSS Pop2, 1CPS, Upper Grade
                       HLM Results from Proc Mixed, UNWEIGHTED
                       TIMSS.990105.003      source: D_proc_hlm_sas.txt

---------------------- *Country Abbreviation*=USA DEPVAR=BIMATSCR MODEL=1 ------------

              _EFFECT_          _EST_           _SE_     _DF_      _T_        _PT_

              INTERCEPT     491.06431195      4.58838171    180    107.02     0.0001
              Tau          3552.7600207     402.47068810      .      8.83     0.0001
              Sigma^2      4692.2528156     112.75854991      .     41.61     0.0001


---------------------- *Country Abbreviation*=USA DEPVAR=BIMATSCR MODEL=2 ------------

              _EFFECT_          _EST_           _SE_     _DF_      _T_        _PT_

              INTERCEPT     494.52153739      4.59270045    180    107.68     0.0001
              BSBGADU2        3.80578452      2.89069033   2786      1.32     0.1881
              BSBGPS02        4.98121848      2.87288831   2786      1.73     0.0831
              BSBGBOOK        7.87457756      1.18097866   2786      6.67     0.0001
              BSBMSIP2        3.68904364      2.58894274   2786      1.42     0.1543
              BSBMMIP2        5.50508061      2.69114638   2786      2.05     0.0409
              BSDAGE        -12.48738151      2.79570110   2786     -4.47     0.0001
              BSDAGESQ       -6.86384608      3.94653887   2786     -1.74     0.0821
              BSBGBRN1        0.68016961      4.99629003   2786      0.14     0.8917
              BSBGEDUF        1.65866596      1.08692359   2786      1.53     0.1271
              BSBGEDUM       -0.70654341      1.07179498   2786     -0.66     0.5098
              Tau          3368.8508637     388.89107562      .      8.66     0.0001
              Sigma^2      4440.1011211     118.80121066      .     37.37     0.0001


---------------------- *Country Abbreviation*=USA DEPVAR=BIMATSCR MODEL=3 -----------

              _EFFECT_          _EST_           _SE_     _DF_      _T_        _PT_

              INTERCEPT    -536.5463051     204.82510461    170     -2.62     0.0096
              BSBGADU2        3.92628579      2.88528104   2786      1.36     0.1737
              BSBGPS02        5.16078404      2.86904641   2786      1.80     0.0722
              BSBGBOOK        7.78697862      1.17939065   2786      6.60     0.0001
              BSBMSIP2        3.65145387      2.58494438   2786      1.41     0.1579
              BSBMMIP2        5.71261830      2.68696307   2786      2.13     0.0336
              BSDAGE        -13.43344077      2.79474028   2786     -4.81     0.0001
              BSDAGESQ       -3.61518446      3.96686342   2786     -0.91     0.3622
              BSBGBRN1        0.64220260      4.98609971   2786      0.13     0.8975
              BSBGEDUF        1.73329963      1.08608042   2786      1.60     0.1106
              BSBGEDUM       -0.81083289      1.07060771   2786     -0.76     0.4489
              SRBSADU2      105.40839539     17.60837444    170      5.99     0.0001
              SRBSPS02       23.67913408     21.54298365    170      1.10     0.2733
              SBSBGBOO       38.54991692      7.21997134    170      5.34     0.0001
              SSBMSIP2        9.94767218     18.20517696    170      0.55     0.5855
              SSBMMIP2       54.94886834     22.09050064    170      2.49     0.0138
              SBSDAGE        38.51924047     12.90452094    170      2.98     0.0033
              SBDAGESQ     -107.1474027      21.84702186    170     -4.90     0.0001
              SBSBGBRN       14.00960769     27.06697048    170      0.52     0.6054
              SBSBGEDF        5.41114993      6.83280947    170      0.79     0.4295
              SBSBGEDM        4.47839816      7.62216283    170      0.59     0.5576
              Tau           682.43569696    102.92002095      .      6.63     0.0001
              Sigma^2      4439.1342712     118.72175984      .     37.39     0.0001
```

```
----------------------- *Country Abbreviation*=USA DEPVAR=BIMATSCR MODEL=4 ------------

          _EFFECT_          _EST_            _SE_      _DF_        _T_         _PT_

          INTERCEPT    -504.6369199    205.98986535     172       -2.45       0.0153
          BSBGADU2        3.91021938      2.88684671    2779        1.35       0.1757
          BSBGPS02        4.95144103      2.86206319    2779        1.73       0.0837
          BSBGBOOK        7.79133311      1.17581686    2779        6.63       0.0001
          BSBMSIP2        3.89679334      2.58163398    2779        1.51       0.1313
          BSBMMIP2        5.62793498      2.68614721    2779        2.10       0.0362
          BSDAGE        -13.53783888      2.78942751    2779       -4.85       0.0001
          BSDAGESQ       -3.83130089      3.96338219    2779       -0.97       0.3338
          BSBGBRN1        0.69147684      5.00077202    2779        0.14       0.8900
          BSBGEDF2      -10.59517734      4.99866397    2779       -2.12       0.0341
          BSBGEDF3       -3.43554587      3.99554497    2779       -0.86       0.3899
          BSBGEDF4       -2.68330980      4.56735464    2779       -0.59       0.5569
          BSBGEDF5        1.54300170      3.86498312    2779        0.40       0.6898
          BSBGEDM2       -5.80362787      8.17302428    2779       -0.71       0.4777
          BSBGEDM3       -2.60757750      7.92925631    2779       -0.33       0.7423
          BSBGEDM4       -6.20144168      8.64922968    2779       -0.72       0.4734
          BSBGEDM5       -1.30676023      7.96419248    2779       -0.16       0.8697
          BSBGEDM6       -7.33978108      7.97267209    2779       -0.92       0.3573
          SRBSADU2      100.24736314     17.33250131     172        5.78       0.0001
          SRBSPS02       38.85948832     18.18098344     172        2.14       0.0340
          SBSBGBOO       42.74728043      6.66593146     172        6.41       0.0001
          SSBMSIP2       11.73400488     18.28230999     172        0.64       0.5218
          SSBMMIP2       55.25966840     22.27203860     172        2.48       0.0141
          SBSDAGE        37.99607299     13.00196387     172        2.92       0.0039
          SBDAGESQ     -108.4775058     21.90179863     172       -4.95       0.0001
          SBSBGBRN        9.34730900     27.13090723     172        0.34       0.7309
          Tau           699.05292803    104.73283112       .        6.67       0.0001
          Sigma^2      4423.7579010    118.32140351       .       37.39       0.0001


----------------------- *Country Abbreviation*=USA DEPVAR=BIMATSCR MODEL=5 ------------

          _EFFECT_          _EST_            _SE_      _DF_        _T_         _PT_

          INTERCEPT    -523.1109567    203.76721514     171       -2.57       0.0111
          BSBGADU2        3.74799544      2.86716247    2838        1.31       0.1912
          BSBGPS02        4.81841978      2.85252181    2838        1.69       0.0913
          BSBGBOOK        7.96437257      1.17198688    2838        6.80       0.0001
          BOTHSIP         9.55664007      2.76243295    2838        3.46       0.0005
          BSDAGE        -13.85406433      2.77435181    2838       -4.99       0.0001
          BSDAGESQ       -5.20513005      3.89974942    2838       -1.33       0.1821
          BSBGBRN1        0.78088495      4.91171889    2838        0.16       0.8737
          BSBGEDUM       -0.83667048      1.06199587    2838       -0.79       0.4309
          BSBGEDUF        1.74057001      1.07566869    2838        1.62       0.1057
          SRBSADU2      105.73509836     17.72691869     171        5.96       0.0001
          SRBSPS02       30.01342653     21.38104613     171        1.40       0.1622
          SBSBGBOO       37.43149260      7.23451840     171        5.17       0.0001
          SBOTHSIP       51.06286273     17.07811229     171        2.99       0.0032
          SBSDAGE        38.29976032     12.94040589     171        2.96       0.0035
          SBDAGESQ     -102.6060435     21.88833741     171       -4.69       0.0001
          SBSBGBRN       20.93219577     26.83706557     171        0.78       0.4365
          SBSBGEDM        4.72562273      7.66780923     171        0.62       0.5385
          SBSBGEDF        5.08009316      6.87202093     171        0.74       0.4608
          Tau           699.32293927    104.64622222       .        6.68       0.0001
          Sigma^2      4462.7871601    118.29817382       .       37.72       0.0001
```

```
---------------------- *Country Abbreviation*=USA DEPVAR=BIMATSCR MODEL=6 -------------

          _EFFECT_           _EST_            _SE_      _DF_       _T_        _PT_

          INTERCEPT      -543.6083459    202.98594017    172      -2.68      0.0081
          BSBGADU2          3.50886305      2.86157386   2844       1.23      0.2202
          BSBGPS02          5.01527446      2.84872252   2844       1.76      0.0784
          BSBGBOOK          8.03972611      1.16808535   2844       6.88      0.0001
          BOTHSIP           9.65529739      2.75989961   2844       3.50      0.0005
          BSDAGE          -14.00887399      2.77016192   2844      -5.06      0.0001
          BSDAGESQ         -5.00682231      3.89508456   2844      -1.29      0.1987
          BSBGEDUM         -0.76261982      1.06160569   2844      -0.72      0.4726
          BSBGEDUF          1.69573990      1.07501873   2844       1.58      0.1148
          SRBSADU2        104.58088954     17.64747066    172       5.93      0.0001
          SRBSPS02         28.40472989     21.40340361    172       1.33      0.1862
          SBSBGBOO         39.52103341      6.82229515    172       5.79      0.0001
          SBOTHSIP         54.06504714     16.86747924    172       3.21      0.0016
          SBSDAGE          40.10399629     12.73520343    172       3.15      0.0019
          SBDAGESQ       -104.7440078      21.76129332    172      -4.81      0.0001
          SBSBGEDM          4.85188570      7.69312653    172       0.63      0.5291
          SBSBGEDF          4.37320652      6.84124810    172       0.64      0.5235
          Tau             707.51856617    105.50377300      .       6.71      0.0001
          Sigma^2        4465.5197810     118.26773352      .      37.76      0.0001


---------------------- *Country Abbreviation*=USA DEPVAR=BIMATSCR MODEL=7 ------------

          _EFFECT_           _EST_            _SE_      _DF_       _T_        _PT_

          INTERCEPT      -530.4133272    202.78149526    173      -2.62      0.0097
          BSBGADU2          3.72425848      2.82309908   2916       1.32      0.1872
          BSBGPS02          4.43813189      2.79863109   2916       1.59      0.1129
          BSBGBOOK          7.70277912      1.14397980   2916       6.73      0.0001
          BOTHSIP           9.59739085      2.72301595   2916       3.52      0.0004
          BSDAGE          -14.36439444      2.73743060   2916      -5.25      0.0001
          BSDAGESQ         -5.39575414      3.83572024   2916      -1.41      0.1596
          BSBGEDUF          1.59753524      0.94965050   2916       1.68      0.0926
          SRBSADU2        103.04442056     17.30423074    173       5.95      0.0001
          SRBSPS02         31.70841582     20.97773005    173       1.51      0.1325
          SBSBGBOO         41.09147362      6.43441476    173       6.39      0.0001
          SBOTHSIP         54.57300192     16.86773296    173       3.24      0.0015
          SBSDAGE          39.23310942     12.73250816    173       3.08      0.0024
          SBDAGESQ       -101.3513887      21.73045006    173      -4.66      0.0001
          SBSBGEDF          6.89926659      5.51788038    173       1.25      0.2129
          Tau             714.49425472    105.80798968      .       6.75      0.0001
          Sigma^2        4458.2833055     116.65296776      .      38.22      0.0001


---------------------- *Country Abbreviation*=USA DEPVAR=BIMATSCR MODEL=8 -------------

          _EFFECT_           _EST_            _SE_      _DF_       _T_        _PT_

          INTERCEPT      -438.4316276    199.08475500    174      -2.20      0.0290
          BSBGADU2          2.92557313      2.54131366   3458       1.15      0.2497
          BSBGPS02          4.21512920      2.55928358   3458       1.65      0.0996
          BSBGBOOK          7.22712763      1.02033353   3458       7.08      0.0001
          BOTHSIP          10.12991047      2.47030659   3458       4.10      0.0001
          BSDAGE          -12.07757975      2.51939914   3458      -4.79      0.0001
          BSDAGESQ         -7.37765706      3.40666381   3458      -2.17      0.0304
          SRBSADU2         99.04571960     16.88302168    174       5.87      0.0001
          SRBSPS02         44.11311839     17.26661550    174       2.55      0.0115
          SBSBGBOO         46.10028489      6.09051537    174       7.57      0.0001
          SBOTHSIP         54.46070202     16.58109607    174       3.28      0.0012
          SBSDAGE          32.86299898     12.52547716    174       2.62      0.0095
          SBDAGESQ        -84.92542356     21.27544129    174      -3.99      0.0001
          Tau             734.93292084    104.20060147      .       7.05      0.0001
          Sigma^2        4514.0191784     108.53899811      .      41.59      0.0001
```

APPENDIX C. SELECTING A MODEL IN FRANCE

The process of selecting models in France followed the same 'judgmental stepwise' procedure illustrated above for the U.S. but with an additional complication caused by the large number of students in the French sample who replied "I don't know" to questions about parental education.

To address this issue, a set of unweighted hierarchical models were run to isolate the effects of the sample restriction caused by this missing data problem and the inclusion or exclusion of parental education. Two samples were used. In mathematics, Sample D included all students with non-missing values on all variables in the models below, with the exception of maternal education. Sample C was the subset of Sample D that excluded all students with missing values or responses of "I don't know" to the question about maternal education. Models were then run with and without maternal education. (R_DUM in the models below is the recoded version of maternal education used in France; MR_DUM is the classroom mean of R_DUM.)

The first of the three models below, labeled "model=1 group=C", includes R_DUM and is therefore restricted to Sample C. The second model, labeled "model=2a group=C" excludes R_DUM and MR_DUM but is nonetheless restricted to Sample C, the students who had informative values for those two variables. The third model, labeled "model=2b group=D", is the same model as model 2a and also excludes R_DUM and MR_DUM, but it takes advantage of this to use the full Sample D.

Thus a comparison of models 2a and 2b shows how much the difference between Samples C and D affect the results of a simplified model that excludes maternal education, because a direct comparison of samples cannot be carried out with maternal education included. The models are similar at the within-classroom level, but the estimated effect for proportion of fathers present (MSBGADU2) is much larger in the larger Sample D. More important given the emphasis in this report on predicted variance are the residual errors in the two models. The residual variances are fairly similar between the two models, but the between-classroom residual variance is about 13 percent smaller in Sample C (model=2a). This suggests that if we had been able to run our final model in the larger Sample D, we might have predicted slightly less of the between-classroom variance.

Simpler analyses confirmed that Samples C and D were different in important ways. Correlations between background variables and scores were calculated at the student and classroom levels for Sample C and Sample C', defined as all students in Sample D who were not in Sample C. These correlations were quite similar across samples at the level of students, but they differed appreciably at the level of school means. For example, in Sample C, the proportion of fathers present correlated .10 with mean math scores; in Sample C', the corresponding correlation was .24. In Sample C, the correlation between mean number of books and proportion of fathers present was -.13; in Sample C', it was +.15.

The comparison of models 1a and 2a then shows the impact of including or excluding maternal education (R_DUM and R_DUM) in the smaller sample C that includes only students with informative responses to that variable. At both levels, the estimates for most of the variables are quite similar in the two models. The exception is the aggregate age variable, which had a larger coefficient when maternal education was excluded. However, the effect of mother's education was significant at both levels in model 1a. As one would expect, the residual variances are smaller at both levels when maternal education is included, but the difference is appreciable only at the between-classroom level.

These analyses indicate that there is no straightforward solution to the problem of missing and uninformative responses to the parental education questions in France. Maternal education was an important predictor in the subsample that had informative values, and that subsample was clearly somewhat different than the complementary group that had missing or uninformative values. Thus, the simple choices are a better model in a nonrepresentative sample or a weaker model in a representative sample.

Faced with this choice and with the relative paucity of strong background variables in TIMSS, we opted for the more inclusive model, including maternal education, at the expense of some nonrepresentativeness of the sample. Other analysts might chose the other alternative. Either choice limits the appropriate inferences from the results, but the analyses here suggest that our basic conclusions about the prediction of variance would not have been fundamentally altered by the choice.

```
                     TIMSS Pop2, 1CPS, Upper Grade
                  HLM Results FRA from Proc Mixed, UNWEIGHTED
           TIMSS.990625.002        source: compare_c_and_d_groups_in_france.sas.txt

------------------ *Country Abbreviation*=FRA depvar=BIMATSCR model=1a group=C -----

          Effect        Estimate      StdErr       DF       tValue      Probt

          Intercept       695.55      136.66       117        5.09      <.0001
          R_DUM            4.6135       1.6113     1489        2.86      0.0043
          BSBGADU2         8.4723       4.4303     1489        1.91      0.0560
          bothsip          8.6783       3.6190     1489        2.40      0.0166
          BSDAGE         -17.1930       2.9034     1489       -5.92      <.0001
          bsdagesq        -0.1554       2.7575     1489       -0.06      0.9551
          MR_DUM          23.1767       5.4181      117        4.28      <.0001
          MSBGADU2        46.6976      20.7134      117        2.25      0.0260
          MOTHSIP         39.1505      14.9107      117        2.63      0.0098
          MSDAGE         -27.2209       7.5916      117       -3.59      0.0005
          msdagesq       -21.9343      11.4347      117       -1.92      0.0575
          Tau            520.93       111.75        .          4.66      <.0001
          Sigma^2       4093.21       149.85        .         27.32      <.0001


------------------ *Country Abbreviation*=FRA depvar=BIMATSCR model=2a group=C ------

          Effect        Estimate      StdErr       DF       tValue      Probt

          Intercept       981.87      126.96       118        7.73      <.0001
          BSBGADU2         8.2578       4.4421     1490        1.86      0.0632
          bothsip          9.4146       3.6200     1490        2.60      0.0094
          BSDAGE         -18.0965       2.8943     1490       -6.25      <.0001
          bsdagesq        -0.2275       2.7652     1490       -0.08      0.9344
          MSBGADU2        45.6070      22.1136      118        2.06      0.0414
          MOTHSIP         38.3958      15.8951      118        2.42      0.0172
          MSDAGE         -42.5287       7.1359      118       -5.96      <.0001
          msdagesq       -17.5852      12.1144      118       -1.45      0.1493
          Tau            646.43       129.53        .          4.99      <.0001
          Sigma^2       4116.31       150.79        .         27.30      <.0001


------------------ *Country Abbreviation*=FRA depvar=BIMATSCR model=2b group=D ------

          Effect        Estimate      StdErr       DF       tValue      Probt

          Intercept       893.36      144.21       118        6.19      <.0001
          BSBGADU2         9.1873       3.3305     2616        2.76      0.0058
          bothsip          7.2899       2.7026     2616        2.70      0.0070
          BSDAGE         -19.4497       2.0437     2616       -9.52      <.0001
          bsdagesq         0.8787       1.6730     2616        0.53      0.5995
          MSBGADU2        78.2348      28.0223      118        2.79      0.0061
          MOTHSIP         42.4893      19.6811      118        2.16      0.0329
          MSDAGE         -39.0743       7.5211      118       -5.20      <.0001
          msdagesq       -37.3806      13.8659      118       -2.70      0.0080
          Tau            745.49       119.72        .          6.23      <.0001
          Sigma^2       4037.84       111.58        .         36.19      <.0001
```

C-3

APPENDIX D. WEIGHTING MULTILEVEL MODELS

Within each country, the TIMSS sample design is complex, with unequal sampling probabilities for different schools in the sample. When the sampling probabilities are unequal, parameter estimates can be biased unless the model controls for the sample design. The most common way of controlling for the sample design is to weight analyses to compensate for unequal sampling probabilities. One could also control for the sampling design by including design features as covariates in the model. Because the design features are not completely known for all the countries included in our analyses, we used the more common approach of weighting our analyses.

Weighting our TIMSS analyses correctly, however, turned out to be problematic. Most statistical software, including the software used for multi-level modeling, such as HLM and SAS PROC MIXED, has the capability of weighting data. Like most SAS procedures, SAS PROC MIXED treats weights as a specific form of precision weights (i.e., the weights are related to the variance of observations). However, sampling weights reflect differences in inclusion probabilities, not the variance of observations and for most analyses SAS PROC MIXED does not produce the exactly the desired estimates. HLM also appears not to provide the desired estimates. Accordingly, we wrote SAS macros that do provide the weighted estimates we consider most desirable. The code for these macros is included in Figure D.1 at the end of this appendix.

In the remainder of this appendix, we describe the pseudo-maximum likelihood estimator (PMLE) as our preferred estimator for weighted multilevel modeling. We then discuss our exploration of weighting, focusing on PROC MIXED because our analyses were primarily conducted using SAS. Using a series of examples, we compare the results of weighted analyses with PROC MIXED to the PMLE and other estimators with known properties. In all our examples, we assume weights vary among schools, but are constant within schools, as they are in the TIMSS one-classroom-per-school sample. This assumption would apply to the full TIMSS samples in the U.S., Australia, and Cyprus only if the level 2 unit were classrooms rather than schools. We conclude that PROC MIXED is not appropriate for weighted multilevel modeling when the weights are sampling weights. Finally, we note the results of a more limited test of HLM, which also failed to provide PMLE estimates consistently.

The examples given here proceed from simple and restrictive cases in which weighting is relatively unproblematic to more realistic illustrations in which the problems caused by weighting can be severe. The final two examples in our discussion of SAS are the two that are most pertinent to the work presented in this report.

Sampling Weights and Complex Models

Pfeffermann et al. (1998) discuss the use of sampling weights when fitting multilevel models to complex sample data with unequal sampling probabilities. Pfeffermann et al. show that weighted multilevel analyses yield consistent estimators under mild regularity conditions. They call their weighted estimates pseudo-maximum likelihood estimates (PMLE).[14] Pfeffermann (1996) shows that in general the PMLE have desirable properties of minimum error in a class of estimators.[15] Thus, the PMLE is a reasonable estimator to consider for TIMSS.

We will describe the PMLE for the simple random intercepts model without covariates, but the method can be easily expanded to include covariates. For a sample of n classrooms nested in n schools (because we have one classroom per school in our analyses), with m_j observations from the jth classroom, the log likelihood for the outcome y and parameters μ, the mean, σ^2, the level-one variance component, and τ^2, the level-two variance component, is given by:

$$-2l(\boldsymbol{m}, \boldsymbol{s}^2 \boldsymbol{t}; \underline{y}) = \quad log(\boldsymbol{s}^2) \sum_{j=1}^{n}(m_j - 1) + \sum_{j=1}^{n} log(m_j \boldsymbol{t} + \boldsymbol{s}^2) +$$

$$\frac{\sum_{j=1}^{n}\sum_{i=1}^{m_j}(y_{ij} - \overline{y}_{\cdot j})^2}{\boldsymbol{s}^2} + \sum_{j=1}^{n}\frac{m_j(\overline{y}_{\cdot j} - \boldsymbol{m})^2}{(m_j \boldsymbol{t} + \boldsymbol{s}^2)} \quad\quad (1)$$

where $\overline{y}_{\cdot j} = \sum_i y_{ij}/m_j$. The estimates are found by solving the likelihood equations:

---

[14]   Pfefferman et al. (1998) actually define a probability weighted version of an iteratively reweighted generalized least squares estimator (PWIGLS), which is asymptotically equivalent to maximum likelihood. In this appendix we actually consider the true PMLE but the estimates are essentially identical to the PWIGLS estimator for our simple examples. The analyses in the body of the report use the PWIGLS estimator.

[15]   The results of Pfefferman (1996) might not apply to multilevel models in general. However, the TIMSS sample does not subsample within the level-two units (classrooms) of our model. Hence, the Pfefferman results should apply to the TIMSS one-classroom-per-school sample and the simple examples in this appendix.

$$\frac{\partial l}{\partial \boldsymbol{m}} = \sum_j \frac{m_j\left(\bar{y}_{\cdot j} - \boldsymbol{m}\right)}{\left(m_j \boldsymbol{t} + \boldsymbol{s}^2\right)} = 0 \tag{2a}$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{\sum_j (m_j - 1)}{2\sigma^2} + \frac{\sum_j \sum_i \left(y_{ij} - \bar{y}_{\cdot j}\right)^2}{2\sigma^4} + \sum_j \frac{m_j\left(\bar{y}_{\cdot j} - \mu\right)^2}{2\left(m_j \tau + \sigma^2\right)^2} = 0 \tag{2b}$$

$$\frac{\partial l}{\partial \boldsymbol{t}} = -\sum_j \frac{m_j}{2\left(m_j \boldsymbol{t} + \boldsymbol{s}^2\right)} + \sum_j \frac{m_j^2\left(\bar{y}_{\cdot j} - \boldsymbol{m}\right)^2}{2\left(m_j \boldsymbol{t} + \boldsymbol{s}^2\right)^2} = 0 \tag{2c}$$

To estimate the PMLE, we replace sums over the *n* level two units with weighted sums. If $m_j = m$ for all level two units, then the PMLE solve:

$$\sum_j mw_j\left(\bar{y}_{\cdot j} - \boldsymbol{m}\right) = 0 \tag{3a}$$

$$\frac{\sum_j w_j \sum_i \left(y_{ij} - \bar{y}_{\cdot j}\right)^2}{\boldsymbol{s}^4} - \frac{(m-1)\sum_j w_j}{\boldsymbol{s}^2} = 0 \tag{3b}$$

$$\frac{\sum_j mw_j\left(\bar{y}_{\cdot j} - \boldsymbol{m}\right)^2}{\boldsymbol{q}^2} - \frac{\sum_j w_j}{\boldsymbol{q}} = 0 \tag{3c}$$

where $\boldsymbol{q} = m\boldsymbol{t} + \boldsymbol{s}^2$. Let $\bar{y}_w = \sum_{ij} w_j y_{ij} \Big/ \sum_j mw_j$, the weighted mean, $SSW = \sum_j w_j \sum_i \left(y_{ij} - \bar{y}_{\cdot j}\right)^2$ and $SSB = \sum_j mw_j\left(\bar{y}_{\cdot j} - \bar{y}_w\right)^2$, then the PMLEs for μ, σ^2 and τ are $\hat{\boldsymbol{m}} = \bar{y}_w$,

$\hat{\boldsymbol{s}}^2 = SSW \Big/ \sum_j w_j (m-1)$ and $\boldsymbol{t} = \left(SSB \Big/ \sum_j w_j - \hat{\boldsymbol{s}}^2\right)\Big/ m$. When the m_j's are not constant, then the solutions must be found iteratively.

If we were not conducting a weighted analysis, we would consider method-of-moments or REML estimates (which are equal for balanced data) as an unbiased alternative to the MLE. The MLE is biased toward zero but the bias converges to zero at a rate of *1/n*. By analogy we can consider weighted method-of-moment estimates $\tilde{\boldsymbol{m}}_c = \hat{\boldsymbol{m}}$, $\tilde{\boldsymbol{s}}_c^2 = \hat{\boldsymbol{s}}^2$ and

$$\tilde{t}_c = \left(SSB \Big/ \left(\sum_j w_j - 1 \right) - \tilde{s}_c^2 \right) \Big/ m \,.$$ These are the estimators suggested by Graubard and Korn

(1996), and they are consistent. We will refer to these estimators as the consistent method-of-moment estimators, CMMs. However, under the assumption of the random intercepts model and assuming that the sample design does not alter the distribution of values, i.e., the design is ignorable (see Pfeffermann, 1996, for a discussion of ignorable designs), then an alternative unbiased method-of-moments estimators exists. To distinguish between the two method-of-moment estimators, we will refer to these estimators as the unbiased method-of-moments estimators, UMMs, even though the estimators are unbiased only under possibly restrictive assumptions. These unbiased estimates are given by: $\tilde{m}_u = \hat{m}$, $\tilde{s}_u^2 = \hat{s}^2$ and

$$\tilde{t}_u = \frac{1}{m} \left(\frac{SSB}{\sum_j w_j - \sum_j w_j^2 \Big/ \sum_j w_j} - \tilde{s}_u^2 \right).^{16}$$ The UMM estimators are invariant to rescaling the

weights by a constant. The CMMs are not.

PROC MIXED does not directly provide the PMLE or a method-of-moment estimator. PROC MIXED provides at least four estimators depending on whether the model is specified using the "random" or the "repeated" statement and whether MLE or REML estimates are used. The assumptions of the random intercepts model are that $Var(\underline{y}_j) = t\mathbf{J} + s^2\mathbf{I}$, where $\underline{y}_j$ denotes the vector of observations for the jth classroom, $\mathbf{J}$ is an m_j by m_j matrix of 1's and $\mathbf{I}$ is an m_j by m_j identity matrix. Without weights, this model can be specified in SAS using PROC MIXED with either the random statement "random intercept /subject=id;" or the repeated statement "repeated / type=cs subject=id;" where the variable id identifies classrooms. However, the models specified by the random and the repeated statements differ when weights are used. With the random statement, PROC MIXED assumes that the $Var(\underline{y}_i) = t/w_j\mathbf{J} + s^2\mathbf{I}$. With the repeated statement, PROC MIXED assumes that the $Var(\underline{y}_i) = (t\mathbf{J} + s^2\mathbf{I})/w_j$.

When the data are balanced and the weights sum to sample size (nm), then the MLEs from the model specified by the repeated statement equal the PMLE. If the weights do not sum

---

16   The unbiasedness of the UMMs is probably of limited value. The UMMs are unbiased only under the assumption that the sample design is ignorable. If the design is ignorable, then weighting is unnecessary and inefficient. However, the estimates do present a principled method for deriving method-of-moment estimators that are invariant to rescaling the weights.

to the sample size, then the MLEs of the variance components produced by PROC MIXED with the repeated statement equal $q\hat{s}^2$ and $q\hat{t}$, where $q = \sum_j mw_j \big/ nm$. If the data are unbalanced PROC MIXED with the repeated statement does not estimate the PMLE. With weighted data, the REML estimates for PROC MIXED with the repeated statement are $\tilde{s}_r^2 = SSW/n(m-1)$ and $\tilde{t}_r = \left(SSB/(n-1) - \tilde{s}_r^2\right)/m$. When the weights are standardized to equal the sample size, then the REML estimates for PROC MIXED with the repeated statement equal the CMMs.

The estimates produced by PROC MIXED with the random statement are not available in closed form for even this simple model. However, as shown below, the estimates from PROC MIXED with the random statement do not generally equal any of the other estimates.

Simple Test Cases

We used a simple test case to explore weighted analyses with SAS PROC MIXED. We selected a sample of 101 schools from the grade 8 U.S. TIMSS sample. All 101 schools had 20 or more students in the TIMSS sample. From each classroom, we selected 20 students at random to be in our special balanced subsample. We adjusted the standard TIMSS weight (multiplied by two and added one) and rounded it to the nearest whole number. Table D.1 gives the distribution of weights. The weights vary across schools but are constant for students from the same school. The sum of the weights is 265.

Table D.1.—Distribution of Weights for Balanced Sample Schools

Weight	Frequency	Percent
1	6	5.9
2	49	48.5
3	30	29.7
4	13	12.9
5	1	1.0
6	1	1.0
8	1	1.0

We used the students' math scores (BIMATSCR) as the outcome variable and fit the simplest mixed model—a one-way random effects ANOVA or a random intercepts models with the grand mean as the only fixed effect. This model has three parameters: the mean, μ, the level-

one variance component, σ^2, and the level-two variance component, τ.[17] We used seven weighted estimators: a) the unbiased method-of-moments (UMM); b) the consistent method-of-moments (CMM); c) the pseudo-maximum likelihood estimator (PSME); d) SAS PROC MIXED with a random statement and MLE; e) SAS PROC MIXED with a random statement and REML; f) SAS PROC MIXED with a repeated statement and MLE; and g) SAS PROC MIXED with a repeated statement and REML. For these PROC MIXED estimates we standardized the weights to sum to 101. (If the weights are not standardized then the variance component estimates will be scaled by the sum of the weights.) We also expanded the data set by repeating all the observations from each school w times, where w is the weight for the school. We reran the analyses using this expanded data set.

Table D.2 gives the results of our comparison study. The first seven rows contain the results of weighted analyses performed on the sample of 20 students from 101 schools. The last five rows contain the results of analyses on the expanded data where schools were replicated w times. The expanded data set contains 5300 observations (20 students from 265 "schools"), although many of the observations are duplicates as a result of replication.

---

[17]   No single notation for level-one and level-two variance components is consistently used by writers in this field. We use σ^2 for the level-one variance and τ for the level-two variance to be consistent with Bryk and Raudenbush (1992), the text that we expect is most familiar to our readers.

Table D.2.—Comparison of Estimators on Sample with 20 Observations from Every School.

Estimation Method	μ	σ^2	τ
Weighted Data			
Unbiased Method-of-Moments*	25.1311	12.3575	7.2406
Consistent Method-of-Moments**	25.1311	12.3575	7.1799
PMLE*	25.1311	12.3575	7.1505
PROC MIXED, RANDOM, MLE***	24.8605	12.3559	7.1316
PROC MIXED, RANDOM, REML***	24.8603	12.3560	7.2098
PROC MIXED, REPEATED, MLE***	25.1311	12.3575	7.1505
PROC MIXED, REPEATED, REML***	25.1311	12.3575	7.2282
Expanded Data			
Method-of-Moments	25.1311	12.3575	7.1799
PROC MIXED, RANDOM, MLE	25.1311	12.3575	7.1505
PROC MIXED, RANDOM, REML	25.1311	12.3575	7.1799
PROC MIXED, REPEATED, MLE	25.1311	12.3575	7.1505
PROC MIXED, REPEATED, REML	25.1311	12.3575	7.1799

*Invariant to sum of the weights
**Sum of the weights equals 5300, not the sample size of 2020.
***Sum of the weights equals 2020.

Among the various methods, the estimates of μ and σ^2 are similar but not identical. In particular, PROC MIXED with the random statement does not use the weighted mean to estimate μ. This is true for both the MLE and REML methods. Also, both the MLE and the REML estimates of σ^2 from PROC MIXED with the random statement differ slightly from all the other estimates.

The estimates of τ vary more among the methods than the estimates of the other parameters. As shown in the table and discussed above, for the special case of balanced data with a simple random intercepts model, PROC MIXED with the repeated statement provides the PMLE for all three parameters including τ, provided the weights are standardized to equal the sample size. Also the PMLE equals the MLE from the expanded data. As expected, the PMLE estimate of τ is smaller than the CMM estimator. The CMM estimate equals the method-of-moment estimator and the REML estimators from the expanded data. Neither the repeated nor the random statement produces a weighted REML estimator that equals the CMM or the REML estimator from the expanded data.

For the expanded data, the random and the repeated statements in PROC MIXED fit the same model and produce the same results. Because the data are balanced, the REML estimates equal the method-of-moments estimates for the expanded data.

The weighted UMM estimate of τ is larger than any of the other estimates are. We expect this estimate to be greater than the CMM estimate. Heuristically, the consistent estimator assumes that the population would look like the expanded data—data from unobserved schools would be a direct replicate of the data from schools in the sample. The unbiased estimator adjusts for the fact that data from the unobserved schools would be similar but not identical to the data from the observed schools. A small simulation study demonstrated that when the sampling design does not affect the distribution of the data, then the unbiased estimator is indeed unbiased and the consistent estimator is slightly biased for a sample of 101 schools with 20 students per school.

We then created an unbalanced data set by deleting 8 observations from half of the sampled schools (51 schools). We then fit our model using this data. We derived the PMLE and the PROC MIXED repeated estimates using this unbalanced and weighted data. We also expanded this data set by repeating schools. We derived the REML and MLE estimates using the expanded data.

Table D.3 presents the results of these comparisons. The PMLE again is essentially equal to the MLE estimates from the expanded data. The small differences in the estimate of τ probably reflect difference in the algorithms for calculating the estimates. For unbalanced data we used the WIGLS algorithm of Pfeffermann et al. (1998) to estimate the PMLE. WIGLS is a weighted version of the iteratively reweighted generalized least squares estimator. Even without weights, the IGLS estimator differs slightly from the MLE, although they are asymptotically equivalent.

Table D.3.—Comparison of Estimators on Sample with 12 or 20 Observations Per School.

Estimation Method	μ	σ^2	τ^2
Weighted Data			
PMLE*	25.0647	12.3469	7.1390
PROC MIXED, REPEATED, MLE**	25.0646	12.3411	7.1958
PROC MIXED, REPEATED, REML**	25.0645	12.3411	7.2755
Expanded Data			
PROC MIXED, REPEATED, MLE	25.0647	12.3471	7.1378
PROC MIXED, REPEATED, REML	25.0646	12.3471	7.1677

*Invariant to sum of the weights.
**Sum of the weights equals 1612.

Weighted estimates from PROC MIXED in Table D.3 do not equal the PMLE or the estimates from the expanded data. The MLEs from PROC MIXED with a repeated statement minimize the likelihood given by:

$$-2l(\mu, \sigma^2\tau^2; \underline{y}) = log(\sigma^2)\sum_{j=1}^{n}(m_j - 1) + \sum_{j=1}^{n}log(m_j\tau + \sigma^2) + \sum_{j=1}^{n}w_j(\underline{y}_j - \underline{\mu}_j)'\mathbf{V}_j^{-1}(\underline{y}_j - \underline{\mu}_j) \tag{4}$$

where $\underline{y}_j$ denotes the vector of observations from the *j*th classroom and $\underline{m} = \mathbf{X}_j\beta$ where β are the regression coefficients and $\mathbf{X}_j$ is the matrix of predictors for the *j*th classroom and $\mathbf{V}_j = t\mathbf{J} + s^2\mathbf{I}$. The PMLEs, on the other hand, maximize the likelihood given by:

$$-2l(\mathbf{m}, s^2t^2; \underline{y}) = log(s^2)\sum_{j=1}^{n}w_j(m_j - 1) + \sum_{j=1}^{n}w_j log(m_jt + s^2) + \sum_{j=1}^{n}w_j(\underline{y}_j - \underline{\mathbf{m}}_j)'\mathbf{V}_j^{-1}(\underline{y}_j - \underline{\mathbf{m}}_j) \tag{5}$$

The two likelihoods are not the same, and this results in the difference in the estimated variance components.

TIMSS Example

　　Our main analysis data set for our grade 8 U.S. math, contained 3647 observations from 181 classrooms in 181 schools, one classroom per school. As described in the body of the report and in Appendix B, we used this data set to fit a model to predict math scores as a function of the student's age and age squared, the number of books in the student's home, whether or not a computer is present in the home, whether or not the father lives in the household and our press variable. The model includes the group mean (between-class) and the group mean centered (within-classroom) values for each predictor. For this comparison, we ran weighted analyses using the TIMSS sampling weight standardized to sum to 3647, the sample size. We calculated the PMLE and the PROC MIXED random and repeated statement MLEs.

　　The results are given in Table D.4. The results for the PMLE are the same as those given in Table 12 of the report. The three estimation procedures provide very similar estimates of the coefficients for the within-class predictors. The PMLE and the PROC MIXED repeated MLE estimates of the coefficients for the between-class predictors are again very similar. However, the PROC MIXED random MLE estimates of coefficients for the between-class predictors diverge from the other estimates. Differences between PROC MIXED with the random statement and the other methods are not surprising. The model fit by random statement assumes a different covariance structure than the model used by the PMLE or the repeated statement. In particular, the random statement model assumes a different value for the ratio of the within classroom residual variance to the between classroom residual variance than do the other models. This ratio controls the weighting of large classrooms relative to smaller classroom. The weighting of classrooms will have the greatest effect on the between classroom predictors.

Table D.4.—Comparison of Estimates for U.S. Grade 8 Example

| | | PROC MIXED | |
	PMLE	Repeated, MLE	Random, MLE
Intercept	-351.67	-350.60	-431.90
Within Class Predictors (Group Mean Centered)			
Number of books	7.93	7.93	7.93
Computer present	4.35	4.35	4.35
Father present	1.73	1.73	1.73
Press	9.55	9.55	9.55
Age	-14.43	-14.43	-14.43
Age^2	-6.86	-6.86	-6.86
Between Classroom Predictors (Group Means)			
Number of books	45.48	45.43	46.54
Computer present	37.23	37.22	42.26
Father present	90.29	90.33	96.85
Press	43.20	43.14	58.58
Age	33.90	33.86	34.21
Age^2	-149.38	-149.78	-106.05
Variance Components			
τ (between)	766.16	823.65	747.84
σ^2 (within)	4570.39	4557.02	4541.79

The estimates of the variance components in Table D.4 also differ across the three methods. These differences are consistent with our findings from the simple test cases described above. PROC MIXED estimates do not maximize the weighted likelihood that the PMLEs maximize. The differences are not large in this example, although the largest estimate of τ (from PROC MIXED with the repeated statement) is about 10% larger than the smallest estimate (from PROC MIXED with the random statement).

Our discussion has so far focused only on the estimated parameters. As shown in Table D.4, the estimates produced by PROC MIXED with the repeated statement and standardized weights are very similar to the PMLEs. However, the standard errors produced by PROC MIXED will tend to be too small. PROC MIXED is not treating the weights as design weights and therefore does not properly adjust the standard error estimates to account for the affect of weighting. PROC MIXED assumes that the variance-covariance matrix for the estimated coefficients is $\mathbf{V}(\hat{\underline{b}}) = (\mathbf{X'W}^{1/2}\hat{\mathbf{V}}^{-1}\mathbf{W}^{1/2}\mathbf{X})^{-1}$ when the true variance (assuming the model is

correct and the design is ignorable) is given by

$$\mathbf{V}\left(\underline{\hat{b}}\right) = \left(\mathbf{X}'\mathbf{W}^{1/2}\hat{\mathbf{V}}^{-1}\mathbf{W}^{1/2}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{1/2}\hat{\mathbf{V}}^{-1}\mathbf{W}^{1/2}\mathbf{V}\mathbf{W}^{1/2}\hat{\mathbf{V}}^{-1}\mathbf{W}^{1/2}\mathbf{X}\left(\mathbf{X}'\mathbf{W}^{1/2}\hat{\mathbf{V}}^{-1}\mathbf{W}^{1/2}\mathbf{X}\right)^{-1}.$$

Table D.5 provides the standard error estimates that correspond to the coefficients in Table D.4. The PROC MIXED standard errors are the default estimates. The PMLE standard errors were estimated using a jackknife procedure and equal those from Table 12 of the main report. Because of the use of the jackknife, our standard errors adjust for the use of sampling weights. As shown in the table the standard errors from PROC MIXED tend to be smaller than the jackknife-standard errors for the PMLEs. We expect that jackknife standard error for PROC MIXED would be similar to those for the PMLE, and therefore the table demonstrates the likely bias in PROC MIXED standard errors.

Table D.5.—Comparison of Standard Error Estimates for U.S. Grade 8 Example

	PMLE	PROC MIXED Repeated, MLE	PROC MIXED Random, MLE
Intercept	265.9	207.51	205.55
Within Class Predictors (Group Mean Centered)			
Number of books	1.2	1.03	1.03
Computer present	3.5	2.61	2.60
Father present	3.3	2.59	2.58
Press	2.6	2.50	2.49
Age	3.3	2.56	2.55
Age^2	3.7	3.44	3.44
Between Classroom Predictors (Group Means)			
Number of books	7.4	6.23	6.27
Computer present	16.8	16.71	17.45
Father present	21.4	18.71	17.64
Press	17.1	16.63	16.89
Age	15.3	13.04	12.95
Age^2	37.1	24.49	22.62

Table D.6 shows the results of the fitting the same model with unstandardized weights (i.e., the sum of the weights equals 5414.74, not 3647). The estimates of the regression coefficients are invariant to the sum of the weights for all three methods. However, as discussed

above, the PROC MIXED estimates of τ and σ^2 are scaled with the ratio of the sum of the weights to the sample size. Hence, the repeated estimates of τ and σ^2 are 1.48 (=5414.76/3647) times larger then the estimate given in Table D.4. Also, the random estimate of σ^2 is 1.48 times larger than the value given in Table D.4. The sensitivity of PROC MIXED to the sum of the weights could result in misleading conclusions. For example, the estimates from the random statement yield biased estimates of the intraclass correlation and the estimates from the repeated statement could bias conclusions about the proportion of variability explained by the predictors. We do not explore standard errors with these estimates because they would obviously be distorted by the estimates of the variance components.

Table D.6.—Comparison of Estimates for U.S. Grade 8 Example

| | | PROC MIXED | |
	PMLE	Repeated, MLE	Random, MLE
Intercept	-351.67	-350.60	-431.90
Within Class Predictors (Group Mean Centered)			
Number of books	7.93	7.93	7.93
Computer present	4.35	4.35	4.35
Father present	1.73	1.73	1.73
Press	9.55	9.55	9.55
Age	-14.43	-14.43	-14.43
Age^2	-6.86	-6.86	-6.86
Between Classroom Predictors (Group Means)			
Number of books	45.48	45.43	46.54
Computer present	37.23	37.22	42.26
Father present	90.29	90.33	96.85
Press	43.20	43.14	58.58
Age	33.90	33.86	34.21
Age^2	-149.38	-149.78	-106.05
Variance Components			
τ (between)	766.16	1222.89	747.84
σ^2 (within)	4570.39	6765.85	6743.24

Weights and the HLM Software

The HLM software (Bryk, Raudenbush and Congdon, 1996) also fits multilevel or hierarchical linear models. This software also allows users to conduct weighted analysis using sampling or design weights. The HLM software provides users with numerous options for conducting weighted analyses. See Bryk, Raudenbush and Congdon (1996) for details. We explored various combinations of the weighting options and found that none of the methods returned the PMLEs for our TIMSS example. The estimates were sensitive to standardizing the weights, and some combinations of estimates do agree with estimates produced by SAS PROC MIXED with the random statement. The HLM documentation does not provide explicit details on the likelihood being maximized under the various weighting methods, so we cannot determine exactly how all estimates will compare to the PMLE. However, we can conclude that in general HLM does not return the PMLE and that we cannot assume that the HLM estimates share the properties of the PMLE.[18]

Summary

Multilevel modeling is a natural approach to analyze of data gathered through a complex multistage sampling design, but using weights to address differences in sampling probabilities can be problematic. Unweighted estimates may be biased, but the weighting options in SAS and HLM do not necessarily provide the desired pseudo-maximum likelihood (PMLE) estimates under many conditions. In the case of SAS, the differences in estimates result from differences between the likelihood maximized by the PROC MIXED estimates and the likelihood maximized by the PMLEs. The properties of the PROC MIXED estimators are unknown but we cannot assume that they share the properties of the PMLE.

The simplest and probably most common use of weights--applying weights without standardization – can yield estimates that are substantially different from the PMLE. This is shown clearly by the between-classroom coefficients for press and age^2 and the variance components in Table D.5 above. For example, the estimate of τ using the repeated statement differed from the PMLE estimate in that case by 60 percent, and the estimate of σ^2 differed by 48

---

[18]   The default weighting scheme in the MLWIN software is equivalent to PROC MIXED with standardized weights and the repeated statement. See the paper at http://www.ioe.ac.uk/mlwin/weights.pdf.

percent. The use of the random statement resulted in an estimate of τ that was much closer, but the estimate of σ^2 still differed from the PMLE by 48 percent.

The simple expedient of standardizing the weights such that the sum of the standardized weights equals the number of observations greatly reduced these problems in our example, but it did not eliminate them. This can be seen from Table D.4 above, and in particular by comparing Table D.4 (which reflects standardized weights) to Table D.5 (which reflects unstandardized weights). PROC MIXED with the random statement and MLE estimation provided two estimates of level-two parameters that differed substantially from PMLE. For example, the estimate for mean press differed by 36 percent. In this particular case, PROC MIXED with the repeated statement and MLE estimation produced parameter estimates very close to PMLE, but the estimate of τ differed by about 8 percent. We expect that in general when weights vary among classrooms, for random intercept models weighted estimates from PROC MIXED with the repeated statement will be preferable to those from PROC MIXED with the random statement. However, the default standard errors should not be used. Jackknifing or an alternative (e.g., linearization as discussed in Pfeffermann et al., 1998) should be used. Finally, we did not explore situations where weights varied among students in the same classroom. We do not expect that our findings will necessarily generalize to such situations, and readers should not assume that PROC MIXED with the repeated statement will provide acceptable estimates for such data.

For many cases with random intercepts models with constant weights within level 2 units, estimate of both parameters and variance components that are consistent with PMLE can be obtained by using SAS macros written for this project. These macros are time-consuming when jackknifing is required, as with TIMSS, but are fast otherwise. The macro code is given in Figure 1.

For this paper we relied on the PMLE because it is a principled approach which is asymptotically optimal. However, the PMLE might not be optimal for all analyses. The small sample properties of the various estimation methods are unknown and need to be explored. In particular the PROC MIXED estimates might be less variable or have smaller small sample bias then the PMLE. In addition, for random intercept models with no covariates, the consistent and unbiased method-of-moment estimators serve as alternatives to the PMLE that might also have better small sample properties. These alternatives account for estimation of the mean and its

effect of the expected value of the sum-of-squares (SSB). The consistent method-of-moment estimator is sensitive to rescaling the weights, but the unbiased method-of-moment estimator is invariant to the scale of the weights.

Figure D.1. SAS Macro Used to Implement WIGLS.

```
%macro gls(y, xvars, dsname, class, intercpt, xlist=%str());
"


"
data _tmp03;

"
set _tmp;
"

   if _n_ = 1 then set sigma;
   array vars &y &xvars &intercpt;
   array avars _y &xlist &intercpt;
   lambda = sigma2 / (_nj * tau2 + sigma2);
   do over vars;
      vars = vars - avars * (1 - sqrt(lambda) );
      end;
run;

proc reg data=_tmp03 noprint outest=beta;
   model &y = &intercpt &xvars / noint;
   weight _wgt;
   output out=_tmp03 residual=_resid;
run;

proc summary data=_tmp03 nway;
   class &class;
   var _resid;
   output out=_tmp04(keep=&class _rbar) mean=_rbar;
run;

data _tmp03;
   merge _tmp03 _tmp04;
   by &class;
   _resid = _resid - _rbar * (1 - sqrt(lambda)) ;
run;

%mend;



%macro igls(y, xvars, dsname, class, weight=1,
            maxiter=500, intercpt=_intcpt);

%let nx = 1;
%let xvar = tmp;
%let xlist = %str();
%do %until(&nx = 0);
    %if %scan(&xvars, &nx) ^= %str() %then %do;
        %let xlist = &xlist _x&nx ;
        %let nx = %eval(&nx + 1);
        %end;
    %else %let nx = 0;
```

```
%end;

%put &xlist;
%if &intercpt ^= _intcpt %then %let intercpt = %str();

data _tmp;
   set &dsname;
   _wgt = &weight;
   _intcpt = 1;
run;

proc means data=_tmp;
   var _wgt houwgt;
run;

proc summary data=_tmp nway;
   class &class;
   var &y &xvars;
   output out=_tmp02(keep=&class _y &xlist _freq_) mean=_y &xlist ;
run;

data _tmp;
   merge _tmp _tmp02(rename=(_freq_=_nj));
   by &class;
run;

data sigma;
   sigma2 = 1;
   tau2 = 0;
   output;
run;

**Iteration 0 betas **;

%gls(&y, &xvars, &dsname, &class, &intercpt, xlist=&xlist)

data newest;
   merge sigma(keep=sigma2 tau2)
         beta(keep=&intercpt &xvars);
run;

proc transpose data=newest out=newest;
   var sigma2 tau2 &intercpt &xvars;
run;

%let conv = 0;
%let i = 0;

**Loop**;

%do %while(&conv = 0);

data oldest;
   set newest;
run;

proc summary data=_tmp03 nway;
```

```
    class &class;
    var _resid;
    id _wgt lambda _nj;
    output out=_s(keep=_rsum _rpr _wgt lambda _nj)
           sum=_rsum uss=_rpr;
run;

data _r;
    set _s;
    _r11 = _wgt * lambda**2 * _nj**2;
    _r12 = _wgt * lambda**2 * _nj;
    _r22 = _wgt * (lambda**2 + _nj - 1);
run;

proc summary data=_r;
    var _r11 _r12 _r22;
    output out=_r(keep= _r11 _r12 _r22) sum=;
run;

data _s;
    set _s;
    _rsum2 = (_rsum)**2;
    keep _rsum2 _rpr _wgt;
run;

proc summary data=_s;
    var _rsum2 _rpr;
    weight _wgt;
    output out=_s(keep=_rsum2 _rpr) sum=;
run;

data sigma;
    merge _r _s;
    detr = _r11*_r22 - _r12**2;
    tau2 = (_r22 * _rsum2 - _r12 * _rpr) / detr;
    sigma2 = (_r11 * _rpr - _r12 * _rsum2) / detr;
    if tau2 < 0 then put "*** NEGATIVE TAU2 ***";
    keep tau2 sigma2;
run;

*proc print data=sigma;

%gls(&y, &xvars, &dsname, &class, &intercpt, xlist=&xlist)

data newest;
    merge sigma(keep=sigma2 tau2)
          beta(keep=&intercpt &xvars);
    iter = &i;
run;

proc print data=newest;
    var iter &intercpt &xvars sigma2 tau2;
run;

proc transpose data=newest out=newest;
    var &intercpt &xvars sigma2 tau2;
run;
```

```
data conv;
   merge newest oldest(rename=(col1=old));
   diff = abs(col1 - old);
run;

proc summary data=conv;
   var diff;
   output out=conv max=;
run;

data conv;
   set conv;
   if 1000000*diff < 1 then
      call symput("conv", 1);
   if &i = &maxiter then
      call symput("conv", 2);
run;

%put &conv;

%let i = %eval(&i + 1);

%end;

proc transpose data=newest out=finalest;
   var col1;
   id _name_;
run;

proc print data=finalest;
   var &intercpt &xvars sigma2 tau2;
   title "Final Estimates from IGSL Algorithm";
   title2 "Covergence Criteria = &conv";
run;

%mend;
```

APPENDIX E. DECOMPOSING THE VARIABILITY IN MULTILEVEL DATA

In this report we explore the predictive power of a small set of predictor variables for explaining both the classroom level variability and the student within classroom level variability of test scores. For traditional linear regression models, which estimate the predictive power of variables at a single level, the R^2 statistic is used to describe the predictive power of covariates. No single method exists, however, for extending the R^2 statistic to multilevel analysis.

In this appendix we describe the approach used in our analyses for calculating the multilevel R^2 statistic. As shown below, different methods for calculating this statistic produce slightly different results. However, these differences are small enough that our main findings are insensitive to the R^2 statistic we chose. We nonetheless present this description of the method we chose for two reasons. First, the careful reader will notice that our unmodeled decompositions of variance (that is, the decomposition of score variance into within- and between-classroom components, taking no predictors into account) do not precisely match the decompositions we present along with our multilevel models. This Appendix explains the reasons for those discrepancies. Second, we believe that our method has merit because it provides a means of decomposing the variability into between and within classroom variability and then decomposing each of these sources of variability into variability modeled by the predictors and residual error. Other methods do not provide a complete decomposition of the variability.

The R^2 statistic in single-level linear models

In traditional, single-level linear models, in which error is assumed to be homoscedastic and independent, the R^2 statistic can be viewed from many perspectives. The R^2 statistic is the square of the sample correlation between the predicted and the observed values. The R^2 statistic is also the ratio of the sum-of-squares due to regression or model sum-of-squares (MSS) to the total sum-of-squares (TSS), R^2 = MSS/TSS. If the sample size equals *n*, then the sum-of-squares due to regression is *n*-1 times the sample variance of the predicted values (the $\hat{y}$'s), and total sum-of-squares is *n*-1 times the sample variance in the observed *y* values. Equivalently the R^2 statistic equals 1-RSS/TSS, where the RSS is the sum of the squared residuals, i.e., the residual sum-of-squares. See Weisberg (1985) for an introduction to this development of the R^2 statistic. The R^2 statistic can also be seen as the proportional reduction in prediction error that results from

using covariates to predict the outcome. The mean square prediction error for predicting y with no covariates is the variance of y. The mean square prediction error for predicting y with the best linear predictor (best implies minimization of the squared prediction error) based on the covariates $\underline{x}$ is the $\mathrm{Var}(y - \underline{x}'\underline{\beta})$. Thus the proportional reduction in mean squared prediction error is

$$\theta = \frac{\mathrm{Var}(y) - \mathrm{Var}(y - \underline{x}'\underline{\beta})}{\mathrm{Var}(y)} = 1 - \frac{\mathrm{Var}(y - \underline{x}'\underline{\beta})}{\mathrm{Var}(y)} \tag{1}$$

If we estimate $\mathrm{Var}(y - \underline{x}'\underline{\beta})$ with *RSS/(n-1)* and $\mathrm{Var}(y)$ with *TSS/(n-1)*, then R^2 statistic is an estimate of θ. See Snijders and Bosker (1994) for details on this explication of R^2.

R^2 statistics in multilevel models

Defining variability in a multilevel model is much more complex. We first must define the variability between classrooms and the variability within classrooms. Let y_{ij} denote the test score for the ith ($i = 1, ..., m_j$) student from the jth classroom ($j = 1,...,n$). Let $\underline{z}_j$ be a vector of classroom level predictors for students from the jth classroom and let $\underline{x}_{ij}$ be a vector of student level predictors, $\sum_i x_{ij} = 0$. Assuming a random intercepts model,

$$y_{ij} = \mu_j + \underline{x}'_{ij}\gamma + \varepsilon_{ij}, \tag{2a}$$

$$\mu_j = \underline{z}'_j\underline{\beta} + \eta_j, \tag{2b}$$

where $\mathrm{Var}(\varepsilon_{ij}) = \boldsymbol{s}^2$, the $\mathrm{Var}(\mu_j) = \tau$ and ε_{ij} and η_j are independent and independent of the predictors. $\mathrm{Var}(y_{ij}) = \mathrm{Var}(y_{ij} - \mu_j) + \mathrm{Var}(\mu_j)$ which is the natural decomposition of the variance of y into variance between classrooms $\mathrm{Var}(\mu_j)$ and variance within $\mathrm{Var}(y_{ij} - \mu_j)$. Thus, to create statistics analogous to R^2, ideally we would like to decompose the sample variability, i.e.,

$TSS/(M-1)$, into $v_w = \sum_{ij}\left(y_{ij}^* - \bar{y}^*\right)^2 \big/(M-1)$ and $v_b = \sum_j n_j\left(\mu_j - \bar{\mu}\right)^2\big/(M-1)$, where $y_{ij}^* = y_{ij} - \boldsymbol{m}_j$,

$M = \sum_j m_j$ and $\bar{y}^*$ and $\bar{\mu}$ denote the respective sample means. We would then decompose v_w and v_b into modeled and residual variability using methods analogous to those for linear

regression. However, we do not directly observe the μ_j's, and so we cannot decompose the variability directly.

Instead, we decompose the conditional expected value of TSS, conditioning on the observed values of $\underline{x}$ and $\underline{z}$.

$$E(TSS / \underline{x}, \underline{z}) = \hat{\sigma}\left(M - \frac{\sum\limits_j m_j^2}{M}\right) + \sum_j m_j \underline{\beta}'\left(\underline{z}_j - \overline{\underline{z}}_.\right)\left(\underline{z}_j - \overline{\underline{z}}_.\right)'\underline{\beta} + \tag{3}$$

$$\sigma^2(M-1) + \sum_{i,j} \underline{\gamma}' \underline{x}_{ij} \underline{x}'_{ij} \underline{\gamma}$$

Because $(M-1)E(v_b / \underline{x}, \underline{z}) = \hat{\sigma}\left(M - \sum_j m_j^2 / M\right) + \sum_j m_j \underline{\beta}'\left(\underline{z}_j - \overline{\underline{z}}_.\right)\left(\underline{z}_j - \overline{\underline{z}}_.\right)'\underline{\beta}$ and

$(M-1)E(v_w / \underline{x}, \underline{z}) = \sigma^2(M-1) + \sum_{i,j} \underline{\gamma}' \underline{x}_{ij} \underline{x}'_{ij} \underline{\gamma}$, we can decompose $E(TSS \mid \underline{x}, \underline{z})$ into a

"between" classroom component, $\hat{\sigma}\left(M - \sum_j m_j^2 / M\right) + \sum_j m_j \underline{\beta}'\left(\underline{z}_j - \overline{\underline{z}}_.\right)\left(\underline{z}_j - \overline{\underline{z}}_.\right)'\underline{\beta}$, and a "within"

classroom component, $\sigma^2(M-1) + \sum_{i,j} \underline{\gamma}' \underline{x}_{ij} \underline{x}'_{ij} \underline{\gamma}$. We estimate $E(TSS \mid \underline{x}, \underline{z})$ and the between and within components using the estimates of τ, σ^2, $\underline{\beta}$ and γ found by fitting the model given in (2). We denote the resulting estimate of $E(TSS \mid \underline{x}, \underline{z})$ by *ETSS*, and the estimates of the between and within components by *BSS* and *WSS*, respectively.

The percent of total variability in scores that is between classrooms is defined as 100 x *BSS/ETSS* and the percent of variability that is within classrooms is defined as 100 x *WSS/ETSS*. We also define the percent of between classroom variability modeled by the classroom level predictors (z_j's) as $100 \times \sum \hat{\underline{\beta}}'\left(\underline{z}_j - \overline{\underline{z}}_.\right)\left(\underline{z}_j - \overline{\underline{z}}_.\right)'\hat{\underline{\beta}} / BSS$. We estimate the percent of within classroom variability explained by the within classroom predictors ($\underline{x}_{ij}$'s) using an analogous formula.

The percent of variability explained by a single covariate at either level also can be estimated by this approach. For example, to determine the variability explained by a classroom level predictor, z, we use standard linear model results to determine the contribution of single predictor to the model sum-of-squares $\sum_i \hat{\underline{\beta}}'\left(\underline{z}_j - \overline{\underline{z}}_.\right)\left(\underline{z}_j - \overline{\underline{z}}_.\right)'\hat{\underline{\beta}}$. We can use this method to find

the full contribution of the predictor, i.e., the proportion of variability explained when this predictor is "entered into the model first," or the additional contribution of the predictor after controlling for other covariates i.e., the proportion of variability explained when this predictor is "entered into the model last."

Although we developed our estimates of the percent of variability explained as an approximation to the decomposition of the sample variability in the y_{ij}'s, *ETSS, BSS* and *WSS* can also be viewed estimates of the $\mathrm{Var}(y_{ij})$, $\mathrm{Var}(\mu_j)$ and $\mathrm{Var}(y_{ij} - \mu_j)$ respectively. Therefore our method can be viewed as an approximate decomposition of the variance in the y_{ij}'s or as a measure of the relative prediction error in predicting either the μ_j's or the $y_{ij} - \mu_j$'s.

Within and between classroom variability are not invariant to the predictors included in the model. Our method for decomposing the variability in scores is conditional on a given set of predictors. The percent of variability within or between classrooms is defined conditionally on the within and between predictors in the full model. In addition the proportion of variability explained at either level of the decomposition is conditional on the full set of predictors we are considering.

Such conditioning is necessary because between and within variance and variability depend on the within classroom level covariates in the model. For example, let's suppose that instead of fitting model (2) we consider the reduced model:

$$y_{ij} = \mu_j{}^* + \varepsilon_{ij}{}^*, \tag{4a}$$

$$\mu_j{}^* = \mu_0 + \eta_i{}^*. \tag{4b}$$

where the within classroom variance is $\mathrm{Var}(\varepsilon_{ij}{}^*) = \sigma^{*2}$ and the between classroom variance is $\mathrm{Var}(\eta_j{}^*) = \tau^*$. Note that we use the asterisk (*) notation to distinguish the parameters of model (4) from their counterparts in model (2). The relationship between the parameters in the two models is described in detail below. Model (4) is perfectly reasonable. This is the model we consider if we want to obtain the decomposition of variance into between and within classrooms irrespective of any covariates, for example, the decomposition given Table 3 of the main report.

If models (2) and (4) are both valid models for the same joint distribution of the y_{ij}'s and the predictors, then the $\varepsilon_{ij}{}^*$'s and the $\eta_j{}^*$'s must incorporate the $\gamma' \underline{z}_j$'s and the $\underline{\beta}' \underline{x}_{ij}$'s. Furthermore,

the variance components for model (4)--σ^{*2} and τ^*--are formally related to σ^2, τ, the variance components from model (2). (To simplify the following presentation, we assume $m_j = m$ and that the z_j's and x_{ij}'s are scalars. We will also assume that the z_j's equal the classroom mean of variables u_{ij}'s and that $x_{ij} = u_{ij} - z_j$. The results are not sensitive to these assumptions, but these simplifications make the presentation much clearer.) Let v_x and v_z denote the variance of the x_{ij}'s and z_j's respectively. Because the x_{ij}'s sum to zero within each classroom, the $\text{Cov}(x_{ij}, x_{ij'}) = -v_z/(m-1)$. Given models (2) and (4), the marginal variance and covariance of the y_{ij}'s satisfy

$$\text{Var}(y_{ij}) = \beta^2 v_x + \tau + \gamma^2 v_z + \sigma^2 = \tau^* + \sigma^{*2} \qquad (5a)$$

$$\text{Cov}(y_{ij}, y_{i'j}) = \beta^2 v_x + \tau - \gamma^2 v_z/(m-1) = \tau^*, \; i \neq i', \qquad (5b)$$

and, therefore, $\sigma^{*2} = m\gamma^2 v_z/(m-1) + \sigma^2$. However, based on model (2) we define the between variance as $\beta^2 v_x + \tau$ and the within variance as $\gamma^2 v_z + \sigma^2$. Thus, the variance components of marginal model (4) do not equal the total between and within variance that we would like to explore. In particular, τ^* is less than the between variance and σ^{*2} is greater than the within variance. The ratio of $\beta^2 v_x$ to τ^* is greater than the ratio of $\beta^2 v_x$ to $\beta^2 v_x + \tau$ or alternatively $1 - \tau/\tau^*$ is greater than $1 - \tau/(\beta^2 v_x + \tau)$. See Snijders and Bosker (1994) for details on the relationship of the variance components (and corresponding estimates) among models with different sets of predictors.

The difference between τ^* and $\beta^2 v_x + \tau$ is demonstrated by the result in Tables 3 and 13 of the main report. For Table D.3 we fit model (4) and found that the between classroom variability (τ^*) accounted for 42 percent of the total variability in U.S. Population 2 math scores. For Table 13 we fit model (2) using the covariates given in Table 12 and found that, on the basis of our decomposition of *ETSS*, the between classroom variability accounted for 41 percent of the total variability.

Although these differences are small, they resulted in counterintuitive estimates when we attempted to decompose the variability using the methods suggested by Bryk and Raudenbush (1992). Let $\hat{\tau}^*$ denote the estimate of τ^* from model (4) and $\hat{\tau}$ denote the estimate of τ from fitting model (2). Bryk and Raudenbush suggest using $100 \times (1 - \hat{\tau}/\hat{\tau}^*)$ to estimate the percent of between class variability explained by the between class predictors. When we used this method

for the TIMSS data, we found that, as demonstrated above, adding within class predictors to our model decreased the percent of between class variability modeled by the between class predictors. Therefore, the Bryk and Raudenbush approach can not be used to separately decompose the between and within variability into modeled and residual components.   Changes to the within-classrooms predictors change both the proportion of the within classroom *and* the proportion of between classroom variability that is modeled by the predictors.

Results are not qualitatively sensitive to the definition of R^2

We compared our method for defining the percent of variability modeled by predictors to two alternative methods. The first alternative is the method of Bryk and Raudenbush (1992) that was discussed above. We call this the HLM method. As discussed, changes to the within class predictors affect the proportion of between variance that is modeled by the predictors. For Table E.1, our HLM estimate of between class variance modeled by between class predictors is always conditional on the full set of within class predictors. Our HLM estimate of the proportion of within class variance explained by a single predictor (the variable father present in our example) ignores the effects on the between classroom variance. We ignore the effect that changes to within class predictors have on the between classroom variance component only to allow for comparison. We are not suggesting that this approach be used in general as a means of forcing the HLM method to produce a decomposition of variability.

The second alternative method we use for our comparison is the ANOVA decomposition. The ANOVA method decomposes $TSS = \sum_{ij}\left(y_{ij} - \overline{y}\right)^2$ into $SSB = \sum_j n_j\left(\overline{y}_{.j} - \overline{y}\right)^2$ and

$SSR = \sum_{ij}\left(y_{ij} - \overline{y}_{.j}\right)^2$ and then decomposes each of these into modeled and residual sum-of-squares. The results of the ANOVA method are equivalent to fitting a model on classroom level means and separate model to the within classroom deviations from the mean. Although the ANOVA method is straightforward, the method will tend to overestimate the between classroom variance because $\overline{y}_{.j} = \mu_j + \overline{\varepsilon}_{.j}$ and $SSB = \sum_j n_j\left(\overline{y}_{.j} - \overline{y}\right)^2 > \sum_j n_j\left(\mu_j - \overline{\mu}\right)^2$. However, these differences will tend to be small provided the sample size within each classroom is at least moderate.

Table E.1 compares the three methods for estimating R^2 statistics for multilevel data. The data are the U.S. Population 2 math scores. The model includes as covariates: the number of

books; computer present; father present and press. The model includes each predictor as both a classroom level (classroom mean) and student within classroom level (group mean centered) variable. The table gives the percent of variance between classrooms, the percent of the between classroom variance explained by the classroom level predictors (the z's) and percent of the within classroom variance explained by the student level predictors (the x's). The table also gives the percent of between and within variance modeled by the father present variable.

The three methods give very similar results. Each method estimates that roughly 43 percent of the total variability is between classrooms and about 57 percent is within classroom. All three methods estimate that somewhat more than 70 percent of the between classroom variability is explained by classroom level predictors (the estimates range from 71 to 76 percent), while a little less than 3 percent of the within classroom variability is explained by student level variables. Also, all methods provide similar estimates of the percent of variability modeled by the father present variable.

The slight differences in the estimates are in the expected directions. As noted above, the ANOVA method includes the average within classroom residual error in the between class variability. Hence the ANOVA method estimates the share of the variance that is between classrooms is larger than the estimate produced by the other methods. Because the additional residual error in the ANOVA method estimate of between class variability cannot be explained by between class predictors, the ANOVA method estimates the smallest percent of between class variability modeled by the z's. Similarly, compared to the ETSS method, the ANOVA method estimates a smaller share of variability is within classrooms and a larger share of that variability is modeled by the x's. The HLM method is something of a compromise between our ETSS method and the ANOVA method. The HLM methods gives estimates that are very similar to the ETSS method for the between classroom variability and nearly identical to the ANOVA method for within classroom variability. However, as discussed above the HLM method does not provide estimates that truly decompose the total variability.

Table E.1.—Three Methods of Estimating Percent of Variability Modeled for U.S. Population 2 Mathematics Scores: 1995

Source	ETSS	HLM	ANOVA
Between Classrooms			
Percent of Total	41.97	42.93	44.71
Percent of Between for:			
All Four Predictors	74.96	75.87	71.03
Father Present			
Total	33.45	32.61	31.65
Additional	7.56	7.79	7.15
Within Classrooms			
Percent of Total	58.03	57.07	55.29
Percent of Within for:	2.62	2.76	2.76
All Four Predictors	0.05	0.05	0.05
Father Present			
Total	0.18	0.19	0.19
Additional	0.05	0.05	0.05

Summary

Determining the percent of variability explained at the different levels of a multilevel model is complicated by the fact that we do not directly observe the variability at the various levels of the data. As an alternative we decompose the expected total sum-of-squares. Using this decomposition we can decompose the total variability (as measured by the expected total sum-of-squares) into variability within and between classrooms and we can then decompose the variability within and between into modeled and residual variability.

Because the total variability between classrooms depends on the within classroom predictors in the model, our decomposition is conditional on the full set of predictors of interest. For decomposing the variance of a fitted model, this method provides a meaningful description of the size of the various sources of variability. The decomposition of expected total sum-of-squares would not be appropriate for other uses of an R^2 statistic such as model comparison. In

addition, because the total between classroom variability depends on the within classroom predictors, a model fit with no predictors does not provide the correct total between and within variability for determining percent of between or within variance modeled by a set of predictors. The result is that the percent of variability between classrooms differs when estimated using our decomposition of expected total sum-of-squares and when estimated using a reduced model with no covariates. The differences are extremely small, as illustrated by Tables 3 and 13 of the body of this report.

Because we cannot directly observe the variability at each level of the model, different methods for approximating a decomposition of the variability exist. We have demonstrated a method that has clear advantages compared to other, more common approaches but have also shown that for the TIMSS data, these differences are too small to alter our findings qualitatively.

APPENDIX F: PARTITIONING WEIGHTS FOR INTERMEDIATE STATISTICS

When an analyst calculates statistics that subsume all levels of a of a multi-stage sample such as TIMSS, the application of sample weights is often straightforward. The database includes a weight (in some databases, numerous weights) that take into account all levels of sampling and non-response down to the level of the individual student, and applying that weight will provide appropriate estimates for statistics such as national means.

In many cases, however, analysts will need to calculate statistics at lower levels of aggregation, which we call here intermediate statistics, and then aggregate those statistics further. An example would be calculating the national distribution of classroom means. In such cases, the use of weights can be more complex, with different weights applied to different stages of the calculations. The process of using weights to create weighted estimates of weighted intermediate statistics is illustrated here with the process of obtaining a weighted distribution of weighted classroom or school means from TIMSS data.

TIMSS is a three-stage survey that samples schools, then classrooms within schools, and finally students within classrooms. In some instances, units are sampled with certainty or with equal probabilities, but that does not affect the generality of the points made here. The TIMSS student weight reflects sampling probabilities at all three levels, but non-response factors only at the level of schools and students:

$$w_{ijk} = \prod_{ijk} (p_i^{-1} \cdot r_i^{-1}) \cdot p_{ij}^{-1} \cdot (p_{ijk}^{-1} \cdot r_{ijk}^{-1}). \qquad (1)$$

where

p is the probability of selection

r is the probability of response or participation, given selection

i is schools

j is classrooms

k is students

In equation (1), factors are grouped into levels of sampling with parentheses. All of these factors are included separately in the TIMSS database (see Gonzalez, Smith, et al., 1997, pp. 3-15 to 3-19).

Typically (as in the TIMSS documentation), weights are labeled in keeping with the stages of sampling, that is, starting with the highest level of sampling and working down only as far as needed. For example, in TIMSS, the school weight (SCHWGT) is the first pair of factors in (1), corresponding to the inverses of the probabilities of selection and response at the first stage of sampling:

$$w_{i\bullet\bullet} = \prod_i (p_i^{-1} \cdot r_i^{-1}). \qquad (2)$$

This labeling is ambiguous. For analytical purposes, it would be clearer to call this fractional weight the *between-school* weight because it is used to weight school-level statistics to obtain correct between-school statistics. The *between-classroom* weight (not separately noted in the TIMSS documentation) would simply multiply the between-school weight by the inverse of the probability of sampling a classroom within a school:

$$w_{ij\bullet} = \prod_{ij} (p_i^{-1} \cdot r_i^{-1}) \cdot p_{ij}^{-1}. \qquad (3)$$

Logically, the weight needed to calculate intermediate statistics at the level of the classroom, the *within-classroom* weight, comprises only the final two factors in (1), that is, the inverses of the probability of being selected from within the classroom and the probability of responding if selected:

$$w_{(c)ijk} = \prod_{ijk} (p_{ijk}^{-1} \cdot r_{ijk}^{-1}). \qquad (4)$$

The *c* in the subscript denotes the classroom level. The product of the within-classroom weight (4) and the between-classroom weight (3) is the total student weight (1). Similarly, the within-school weight is obtained by multiplying the within-classroom weight by the inverse of the probability of sampling a classroom within a school:

$$w_{(s)ijk} = \prod_{ijk} p_{ij}^{-1} \cdot (p_{ijk}^{-1} \cdot r_{ijk}^{-1}). \qquad (5)$$

Again, the product of the within- and between-school weights (5 and 2) is the total student weight (1).

Although it is useful to conceptualize weights in these terms, it is typically unnecessary in practice to partition weights in this fashion to calculate the intermediate statistics. For example, the between school weight is a constant for any school and thus does not influence school-level statistics such as means. As noted above, the student

weight can be expressed as the product of the between-school and within-school components:

$$w_{ijk} = w_{i\bullet\bullet} \prod_{ijk} p_{ij}^{-1} \cdot (p_{ijk}^{-1} \cdot r_{ijk}^{-1}) = w_{i\bullet\bullet} \cdot w_{(s)ijk}.$$

The weighted mean of variable X in school i would therefore be:

$$\overline{X}_i = \frac{\sum_{jk}(w_{ijk} X_{ijk})}{\sum_{jk} w_{ijk}}$$

$$= \frac{w_{i\bullet\bullet} \sum_{jk}(w_{(s)ijk} \cdot X_{ijk})}{w_{i\bullet\bullet} \sum_{jk} w_{(s)ijk}}$$

$$= \frac{\sum_{jk}(w_{(s)ijk} \cdot X_{ijk})}{\sum_{jk} w_{(s)ijk}}.$$

Thus, the weighted school mean is calculated equivalently with either the total student weights or the within-school weights.

To calculate distributional statistics from weighted intermediate statistics (e.g., distributional statistics where schools rather than students are the unit of analysis), however, does require partitioning the total weight into within- and between-unit components. When calculating, for example, the weighted mean of school means, the appropriate weight to apply to the school means is the between-school weight (2) above, i.e., SCHWGT in the TIMSS database. The total student weight (1) is not equivalent to the weighted sum of school mean weights because the within-school weighting components (5) need not be constant within a school. To calculate the weighted mean of classroom means, one would need to compute a between-classroom weight (4) by multiplying the between-school weight SCHWGT by p_{ij}^{-1}, labeled WGTFAC2 in the TIMSS database (Gonzalez, Smith, et al., 1997, p. 3-15).

Depending on the inferences the statistic is intended to support, using the between-unit weight may not be sufficient. For example, suppose that one wanted the national weighted mean and standard deviations of weighted school means (as in Table 2 above). If one wants the weighted mean of means to be equal to the weighted grand

mean calculated with student level weights (i.e., equal to the national mean in the published TIMSS reports), weighting by the between-school weights would not be sufficient, because taking the mean at the school level removes the information about the relative size of the schools. Accordingly, it is necessary to adjust the between-school weights by multiplying them by the sum of the within-school weights. This yields the following adjusted between-school weight:

$$\widetilde{w}_{i\bullet\bullet} = \prod_i (p_i^{-1} \cdot r_i^{-1}) \cdot \sum_{jk} w_{(s)ijk} \,. \qquad (6)$$

This procedure was used for calculating simple statistics in this report. For example, the means and standard deviations of classroom means were calculated first by using the within-school weights (5) to calculate the means for each school and then the adjusted between-school weights (6) to calculate the means and standard deviations of these weighted means.

Reference

Gonzalez, E.J., Smith, T.A., et al. (1997). *User Guide for the TIMSS International Database: Primary and Middle School Years.* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

National Center for Education Statistics

left		blank	Description of NCES	Visit the U.S. Department of Education
	Go to Electronic Catalog			
	Go to Students' Classroom			
	Go to Survey and Program Areas			
	Go to Encyclopedia of ED Stats			
	Go to Quick Tables and Figures			
	Go to Global Locator			
	Go to FastFacts			
	bottom			

[Web Version of Education Statistics Quarterly (Spring 2001) Now Available](#)

Inside the Stats!

Percentage of 1992 high school graduates who enrolled in a 4-year institution within two years of graduation, by parent's level of education.

[Percentage of 1992 high school graduates who enrolled in a 4-year institution within two years of graduation, by parent's level of education.](#)

Did You Know?

In 1998-99, on average, faculty in 4-year postsecondary schools earned over $9,000 more per year than those in 2-year institutions. ([NCES Reference](#))

[Text Only](#)   [Privacy & Security Policy](#)

National Center for Education Statistics

[map](#)   *1990 K Street, NW,   Washington, DC 20006,   USA,   Phone: (202) 502-7300*

Search Tools and Related Information

Conduct customized searches to locate NCES publications and data products. Once located, browse the content of publications or download files.

● **Search for Products by NCES #:**

Enter NCES #: (e.g. 2001017 or 98293)

● **Search on one or more of the following fields:**
Enter Searchword(s): *(time default = products released within last 5 years)*

1. by

2. **Help on Area**

3. **Help on Type of Product**

4. Release Date: Month Year

● **Popular NCES Reports** NEW!   ● **Data Access Tools**

● **How to Order NCES Products**
  1. Education Publications Center (ED Pubs),
  2. Government Printing Office (GPO), and
  3. Federal Depository Libraries

● **Restricted Data Licenses** - **Access restricted data from NCES surveys.**

● **News Flash Subscription Service** - **Receive e-mail alerts about new NCES Publications and Data Products.**

● **Educational Resources Information Center (ERIC) and Other Clearinghouses** - **ERIC, funded by the U.S. Department of Education, is a nationwide information network that acquires, catalogs, summarizes, and provides access to education information from all sources. All ED publications are included in its inventory.**

Text Only

Electronic Catalog of NCES Products (National Center for Education Statistics). Publications and data products.