

What Can TIMSS Surveys Tell Us About 1990's Mathematics Reforms in the United States?

Laura S. Hamilton
RAND Corporation

José Felipe Martinez
RAND Corporation and University of California, Los Angeles

DRAFT

Not for citation or attribution. The opinions expressed are solely those of the authors and do not represent those of RAND or its sponsors.

Introduction

Throughout the 1990's, a number of mathematics reforms were introduced in K-12 schools throughout the United States. Many of these reforms were influenced by the National Council of Teachers of Mathematics (NCTM), which published national standards for mathematics curriculum and instruction (NCTM, 1989; 2000), and by the National Science Foundation, which funded the development of curriculum materials aligned to the NCTM Standards (see Porter et al., 1994; Linn et al., 2000). These reforms and many of the new curricula emphasized the importance of problem solving and inquiry, and were designed to promote increased use of instructional activities that reformers believed would promote students' thinking skills. These activities included cooperative groups, writing about mathematics, and use of open-ended assessment strategies. Reform advocates also pointed to the value of studying topics in depth rather than covering a large number of topics in a shallow way, and they encouraged certain uses of technology, including calculators, when the technology served to promote deeper knowledge and understanding. Although NCTM did not endorse instructional approaches that ignored the development of basic skills and factual knowledge, and has published several subsequent documents that clarify this position (e.g., NCTM, 2006), critics of the reforms expressed concerns certain types of knowledge and skills were often de-emphasized. Looking back on this decade of reform, it is important to ask whether and how these reforms might have influenced teachers' practices and student achievement in mathematics.

In this paper, we review work that has examined relationships between students' mathematics achievement and teachers' use of so-called "reform-oriented" instructional practices. We then present new analyses using data from the Trends in International Mathematics and Science Study (TIMSS) to examine teachers' use of these practices in the

United States and in a small number of other countries, and their relationships with student achievement in mathematics. We also discuss ways in which existing research falls short of answering the most important questions about reform-oriented instruction.

Distinguishing Reform from Traditional Practices: Understanding the Math Wars

For the past several decades, discussions about mathematics curricula have often been characterized by conflict between two camps—advocates of what is often called “traditional” instruction on the one side, and reformers who espouse student-centered or inquiry-based approaches on the other. Proponents of more traditional approaches often argue that student learn mathematics most effectively when given extensive opportunities to develop and practice skills, and when provided direct instruction from a teacher. They point to low student achievement, particularly on tests that measure computational skills (e.g., Loveless, 2003), and note the risks this poses for these students’ future mathematical performance and for the economy as a whole. The reform-oriented approach, by contrast, places greater emphasis on having students construct their own knowledge through investigations and other student-led activities (Le et al., 2006). Reform-minded educators and scholars often cite low test scores as evidence that the traditional approach that has been prevalent in U.S. schools has not worked (Hiebert, 1999). Some members of each camp have assigned labels to the other side, with traditionalists being labeled as in favor of “drill and kill” approaches, and reform advocates being accused of promoting “fuzzy math.” To illustrate, Chester Finn was quoted in a New York Times article, referring to “the constructivist approach some educators prefer, in which children learn what they want to learn when they're ready to learn it" (Lewin, 2006). This view represents a distortion of the kind of constructivist approach that has been thoughtfully described by

scholars such as Battista (2001), but it is not an uncommon one (see the volume edited by Loveless, 2001, for additional discussions of the debate and examples of how each side is often depicted).

There are signs of a partial cease-fire in the math wars. NCTM has published a number of documents that clarify its own positions, most recently the *Curriculum Focal Points for Prekindergarten through Grade 8 Mathematics* (NCTM, 2006), which includes references to the importance of computation and other basic skills. Although NCTM officials argue that this has been their position all along, many observers of the math wars perceive this and related documents as indicators of a growing acceptance of more traditional approaches on the part of NCTM (see Cavanaugh, 2006; Lewin, 2006). The National Mathematics Advisory Panel, created in April 2006 by President Bush to provide advice on the effectiveness of various approaches to teaching mathematics, includes individuals who have been associated with both sides of the debate. Early indications of the panel's work suggest a greater degree of consensus than has been obtained in the past.

Despite the often rancorous environment that characterizes the debate, the two approaches do not necessarily operate in opposition to one another. Moreover, the notion that "reform" and "traditional" approaches can be easily distinguished from one another is simplistic. Most of those who espouse the traditional viewpoint do not believe that students should engage in mindless drills or that higher-order reasoning is unimportant, and the idea that reform-oriented instruction shuns computation in favor of purely "fuzzy" activities that put the students in charge does not consider the ways in which many NCTM-aligned curricula incorporate a variety of activities including computation practice and teacher-led instruction. Interest in reform-oriented instruction is often grounded in a belief that while students are likely to benefit from traditional

approaches, these approaches alone are insufficient: Students also need to be exposed to instruction that promotes conceptual understanding and the ability to solve complex, ill-structured problems, that encourages them to communicate mathematically and to ask questions, and that helps them make connections between mathematics and other disciplines (National Research Council, 1999). The relevance of traditional approaches and goals to effective reform instruction is described by Battista (2001):

“...it is clearly insufficient to involve students only in sense making, reasoning, and the construction of mathematical knowledge. Sound curricula must have clear long-range goals for assuring that students become *fluent* in utilizing those mathematical concepts, ways of reasoning, and procedures that our culture has found most useful...They should possess knowledge that supports mathematical reasoning. For example, students should know the “basic number facts” because such knowledge is essential for mental computation, estimation, performance of computational procedures, and problem solving.” (p.46).

Investigations about the effectiveness of reform-oriented instruction, therefore, need to recognize that the use of more traditional practices is not antithetical to reform and is likely to occur concurrently with reform-oriented practices in many classrooms. In fact, efforts to measure teachers’ use of both approaches typically reveal either null or positive correlations between reform and traditional approaches (Desimone et al., 2005; Klein et al., 2000; Hamilton et al., 2003). Moreover, the field has not reached consensus on how to define “reform-oriented” or “traditional” instruction, so while there is extensive overlap across studies in how these constructs are conceptualized and measured, they do not always mean exactly the same thing.

This discussion makes it clear that simple comparisons of reform versus traditional practices are unlikely to shed much light on questions about the effectiveness of various teaching strategies, given the complex ways in which the two approaches interact and the lack of a universally accepted definition of either approach. Nonetheless, it is worth examining existing data and literature to understand how frequently various instructional approaches are used, and how these approaches are associated with student achievement. In the next section we briefly review a number of studies that have addressed this topic in recent years before turning to a new set of analyses using TIMSS data.

Existing Research on Reform Practices

The research literature suggests that only a small minority of U.S. teachers embody the principles of reform-oriented curriculum and instruction in their lessons. In one recent study, Jacobs and colleagues (Jacobs et al., 2006) used the 1995 and 1999 TIMSS video studies to examine various dimensions of reform practice among middle school teachers and found that the typical teacher used few reform-oriented practices. Moreover, that study's comparison of lessons in 1995 and 1999 suggested that changes in fidelity to reform practices were relatively few. The changes that were observed went in both directions, with 1999 teachers slightly more likely to adopt some aspects of the reforms and less likely to adopt others. Other work lends support to the conclusion that reform-oriented instruction is relatively rare compared with more traditional approaches (e.g., Hieber, 1999; Ravitz, Becker, & Wong, 2000) and indicates that even when teachers are given reform-oriented curriculum materials, they often fail to implement the curriculum in ways consistent with developers' intentions (Battista, 1994; 2001; Spillane & Zeuli, 1999). The relative infrequency of reform-oriented instruction is not unique to the U.S.:

one recent international comparison using 1999 8th grade TIMSS survey data showed that the rate of use among U.S. teachers is similar to the international average (Desimone et al., 2005).

One reason for the relatively infrequent use of reform practices might be the lack of a strong research base for concluding that these practices promote improved achievement.

Research in cognitive and developmental psychology provides some evidence that student learning in mathematics is associated with exposure to aspects of reform-oriented instruction, such as the opportunity for students to construct their own ideas about mathematics (Bransford, Brown, & Cocking, 1999; Cobb et al., 1991; Greeno, Collins, & Resnick, 1996; Hiebert & Carpenter, 1992; Wood & Sellers, 1997). Most of this work has been done in a relatively small number of classrooms, which made it possible to collect rich data on instruction and achievement but limited the generalizability of results.

Several studies examining how teachers' use of reform practices is related to student achievement in large numbers of schools and classrooms have also been conducted (Hamilton et al., 2003; Mayer, 1998; Shouse, 2001; Smith, Lee, & Newmann, 2001; Wenglinsky, 2002). Together, the findings can be described as suggesting at best a small, positive relationship between some aspects of reform-oriented instruction and achievement, but the body of work is far from conclusive, and many of the studies also point to positive relationships with traditional practices as well. Most of this work relies on surveys that ask teachers to report the frequency of various instructional activities such as use of cooperative groups and administration of open-ended assessments. Undoubtedly such surveys fail to capture some of what distinguishes true reform-oriented instruction from other approaches, a problem to which we return in the final section of this paper. In particular, content is often ignored in these studies in favor of a focus on pedagogy. Nonetheless, these survey-based studies are the source of much of what we currently

know about how instructional practices are related to student achievement. Moreover, several studies suggest that questionnaires can provide reasonably accurate information about practices and that acceptable levels of agreement between questionnaire responses and observations are often obtained (Burstein, Chen, & Kim, 1989; Burstein et al., 1995; Porter, 1995).

Moreover, researchers have called attention to the possibility that different outcomes or constructs may be affected differently by instructional practices, even within the same content area. Even with a test that functions in a unidimensional manner, conclusions about the relationship between mathematics achievement and instruction have been shown to depend on the specific choice of items included or the weights assigned to each subtopic in the test (Kupermintz et al., 1995; Hamilton, 1998; Lockwood et al., in press). Estimates of relationships between reform-oriented instruction and achievement are likely to depend in part on the degree to which the outcome measure is sensitive to the specific instruction offered (Le et al., 2006).

Using TIMSS Surveys to Examine Reform-Oriented Instruction

As discussed above, TIMSS survey data have previously been used to investigate teachers' use of reform-oriented instruction in mathematics (e.g., Desimone et al., 2005). In addition, the TIMSS surveys have been adapted for use in other studies of instructional practices (e.g., Spillane & Zeuli). In this paper we present the results of several analyses using the 2003, 1999, and 1995 8th-grade TIMSS data, with a focus on understanding the extent to which teachers in the U.S. and elsewhere have adopted reform-oriented instructional methods and whether these methods are associated with differences in student achievement. We address four questions using data from the U.S. and the other participating countries:

- How has teachers' reported use of reform-oriented instruction changed over time, including their use of reform-oriented practices and the number of topics covered in their courses?
- To what extent are teachers' opinions about how mathematics should be taught related to their reported practices?
- How closely do student and teacher reports of practices match?
- What is the relationship between reform-oriented practices and student achievement, considering both total mathematics scores and scores on each of five content strands?

We focus on the 8th grade for a few reasons. First, a somewhat richer and more consistent set of practice-related variables is available for 8th grade teachers and students across the three waves of data collection. In particular, the 1999 TIMSS study did not include 4th grade classrooms, making analysis of trends across time more difficult. Second, preliminary analyses of the 2003 4th grade practice items yielded less consistent results than the 8th grade data, both between teachers and students and among countries. This lack of consistency complicates the construction of meaningful scales to measure practices. Nevertheless, many of our analyses could easily be adapted to the 4th-grade data.

This study is intended to shed light on instruction in the U.S. during a period of extensive mathematics reform. But our understanding of what is happening in the U.S. can be enhanced by an examination of instruction and achievement elsewhere, so we use the TIMSS data to explore relationships in three other countries (Japan, Singapore, and the Netherlands) that differ in important ways from the U.S. They are not intended to be representative of any larger set of countries, but are merely included to provide a comparative context for understanding U.S.

mathematics instruction and achievement. Japan is an especially high-achieving country, whereas achievement scores for students in the Netherlands are approximately midway between those of the U.S. and Japan. Both of these countries participated in the 1999 TIMSS video study¹, which provided rich information about classroom environments. The video study suggests some potentially important differences among the three countries in how mathematics is taught. For example, classrooms in the Netherlands were characterized by a greater emphasis on real-life mathematics problems, more individual and small-group work, and greater use of calculators than in the U.S. Japanese classrooms stood out from all of the other participating countries in several respects, including a larger amount of time devoted to the introduction of new content, a larger percentage of problems that were judged to be of high complexity, a larger percentage of problems that involved making connections, and a lower percentage of problems that were judged to be repetitions (Hiebert et al., 2003; Stigler et al., 1999). Singapore did not participate in the 1999 video study, but it is consistently one of the highest-scoring nations and its mathematics curriculum has been the subject of great attention among U.S. educators and policymakers². Together, the four countries represent a wide variety of classroom contexts in which to examine the use of reform-oriented instructional practices.

Analytic Approach

As discussed earlier, the focus of this paper is on exploring the utility of the TIMSS survey data for understanding teachers' use of reform-oriented instructional practices and their

¹ Analyses of the 1999 video study relied on videos collected in 1995 for Japan, and in 1999 for the other countries (Hiebert et al., 2003).

² As with many of the curricula used in the U.S., the curricula in these countries have been subject to varying descriptions that reflect different opinions on their "reformedness". Singapore, for example, is often praised for its basic skills-oriented curriculum (see New York Times Editorial, *Teaching Math: Singapore Style*, 9/18/06), but the website for the U.S. Singapore Math curriculum materials indicates that the curriculum "encourages active thinking process, communication of mathematical ideas, and problem solving" (http://www.singaporemath.com/Primary_Math_Textbook_1A_U_S_EDITION_p/pmust1a.htm).

relationships with achievement in mathematics. In this section we discuss the data, instruments, and statistical methods used to analyze the data.

Data

We relied on three primary sources of data: the 8th grade mathematics teacher survey files from the 1995, 1999, and 2003 administrations of TIMSS in the United States, Japan, the Netherlands, and Singapore; the 8th grade students' mathematics test-scores files from the 2003 administration in each of these countries; and the student background survey datasets. We also refer briefly to other data sources, such as the video studies, where relevant. However, the focus of this paper is on the surveys rather than the videos, in part because we are interested in examining the utility of survey-based measures for understanding instructional practice.

For the analyses of instructional practices over time we used the teacher survey datasets for each of the 4 countries across the three waves of TIMSS. The surveys include questions about personal background, education and experience, and also about teacher use of or perceptions about different kinds of instructional practices in the classroom. On the other hand, our analyses of the relationship between instructional practices and student achievement refer to 2003 data only. These analyses combined the student achievement datasets with the files containing the information from the student and teacher background surveys.

TIMSS data have several advantages over many other data sources, including the availability of information for a large number of countries, representative sampling of classrooms, and high-quality measures of student mathematics achievement. At the same time, these data have a number of limitations. Perhaps most significantly, they do not allow researchers to follow individual students over time. It is widely acknowledged that the most

valid approach for evaluating the effects of any educational intervention on student achievement (short of conducting a randomized experiment or quasi-experiment) is to measure each individual's achievement at several points in time to control for unmeasured student characteristics (Charter School Achievement Consensus Panel, 2006). The absence of multiple measures of individual students' achievement over time severely limits the range of analyses that can be conducted and the strength of the conclusions that can be drawn from these analyses. Additional limitations include a somewhat sparse set of items that directly address reform practices, changes in the wording of items from one wave to the next, and the strong possibility of incomparability of survey responses across countries. We address these and other limitations later in this paper.

Instructional Practices and Perceptions Items and Scales (2003 Student and Teacher Surveys)

Figure A in the Appendix presents the complete list of items in the 2003 student and teacher surveys that address instructional practices in the classroom. To measure exposure to reform-oriented practices, we identified items on the teacher and student surveys that asked about practices consistent with a reform-oriented approach to mathematics instruction. We examined correlations among the items separately for each country, and estimated the internal consistency reliability of scales created from these items³.

Table 1 lists the five items included in the reform-oriented instruction scale created from the teacher survey results. These items are similar to those used in other studies of reform-oriented practices (e.g., Cohen & Hill, 2000; Le et al., 2006), though there are fewer such items on the TIMSS surveys than in many of the other studies. For each item teachers were asked to

³ We also conducted exploratory factor analyses on the full set of practice items for students and teachers in each country. The results for the student surveys were inconsistent across countries. The teacher survey results were more consistent, with the five reform-oriented items tending to cluster together in each country.

indicate whether they asked students to engage in each activity every or almost every lesson, about half the lessons, some lessons, or never. The internal consistency reliability for a composite created from these five items (using the average of the 1-4 response scale) is moderate ($\alpha=0.64$ for U.S. teachers and between 0.45 and 0.54 for the other three countries).

Unlike some of the other studies examining instructional practices, we did not create a separate traditional practices scale because the TIMSS surveys did not lend themselves to such a scale. Instead, we created a content-focused composite to measure the frequency with which teachers asked students to engage in activities focused on the following activities: practice adding, subtracting, multiplying, and dividing without using a calculator; work on fractions and decimals; interpret data in tables, charts, or graphs; write equations and functions to represent relationships. The first of these would generally be considered to be associated with a more traditional approach to instruction (though, as noted earlier, this type of skill-building is not precluded by a reform-oriented approach), whereas the other three could be associated with either pedagogical style. All of these items were positively correlated. This composite ($\alpha=0.54$ in the U.S; 0.42 to 0.62 in the other countries) was intended to control for exposure to mathematics content in our achievement models.

The student survey items addressing reform-oriented practices did not cluster as clearly or consistently as the teacher survey items. In particular, the items showed stronger relationships with one another in the U.S. than in the other three countries, and their correlations with other, non-reform-oriented items varied substantially across countries. These differences suggest that students in different countries might be interpreting the items differently. Therefore we did not create student-level composite scores using these items. The models of student achievement presented later in this study include some of these individual items as predictors. However, in

interpreting any relationships observed there it should be remembered that the precise *meaning* of attending a classroom where students report more or less of a certain activity may not be the same across countries.

Finally, we examined a set of items addressing teachers' opinions and perceptions about reform principles in mathematics instruction (See Figure A.1 in the Appendix). Although these items do not directly measure teachers' actions in the classroom, their content gets at the heart of many of the differences between the views of reform advocates and detractors. Table 2 shows the results of an exploratory factor analysis of these items, separating a factor capturing *reform-oriented perceptions* (including, for example, items related to multiple representations or solutions) and a second factor capturing what we consider more *traditional* perceptions (e.g., memorization, algorithms). Table A.1 in the appendix suggests there are some differences in the perceptions of teachers in other countries: thus for example Japanese teachers may perceive algorithms/rules (item b) as closer to *reform* principles, while Dutch teachers perceive hypothesis testing (item c) as a more *traditional*. As with instructional practices, however, here there also seem to be more similarities than differences across countries. Consequently, we constructed similar *reform* and *traditional* composites from the original sets of perception items for teachers in all four countries.⁴

Instructional Practices and Topics Covered Across Time (1995, 1999, and 2003 Teacher Surveys)

One of the goals of this study is to investigate whether use of certain instructional practices (specifically *reform-oriented* practices) saw any changes from 1995 to 2003 in the

⁴ It should be noted that the alpha reliability estimates for *traditional perceptions* were generally low across countries, indicating that these composites may be less consistently measured (and thus less meaningful) than it would be desirable.

United States. As in 2003, in the two previous administrations of the TIMSS (1995, 1999) teachers were asked to report on their use of a number of instructional practices in the classroom. However, the list of instructional practice items in the 2003 teacher survey differs from that used in the 1995 and 1999 surveys. In all, only four instructional practice items remained consistent across all three waves of the TIMSS study. Three of these items are included in our reform-oriented instruction scale, and the fourth is one of the content-focused items discussed earlier. The precise wording of the item each year is presented in Figure 1.

For each year and country we estimated the frequency with which teachers reported using each of these four instructional practices in the classroom. For any given country the samples of teachers in each year of the TIMSS study were not representative of the population of teachers in that country. Instead, the teacher estimates are weighted proportional to the students they teach (Martin, 2005); therefore the weighted estimates presented should be interpreted as referring to the reports of teachers of nationally representative samples of students.

We also conducted an additional analysis that used the information from the teacher surveys about the topics teachers covered during the school year. The goal of this analysis was to explore whether the number of topics covered in 8th grade mathematics changed over past decade. Comparisons across time are complicated here by differences in the number of mathematics topics listed in the survey in each study (37 in 1995, 34 in 1999, and 45 in 2003). To facilitate the relative comparisons across time we obtained the average proportion of topics that teachers in a country reported covering each year.⁵ Furthermore, because we expected

⁵ The scales used to report coverage of topics also suffered some modifications. Thus, in the 1999 and 2003 surveys we cannot tell whether a topic is not covered at all during the year, and in 2003 we cannot separate between *not taught* and *just introduced*; etc.

curricula to differ in algebra classes and general mathematics classes in the United States, we obtained separate estimates in 2003 for these two types of classes.⁶

Multilevel Modeling of Student Reports of Instructional Practice and Student Achievement

The TIMSS sample in the United States includes only two teachers per school, and in the other three countries only one teacher per school. This structure prevented estimation of three-level models where the student, teacher, and school levels could be modeled adequately. Given our focus on classroom practice we employ a series of two-level hierarchical linear models (students nested within teachers) to investigate the distribution of student reports of instructional practice across classrooms, and the relationship between student achievement and use of *reform* instructional practices in the classroom (Raudenbush & Bryk, 2002). Unlike a two-level structure nesting students within schools, this approach avoids aggregating information about practice available by teacher to the school level, and provides accurate estimates of the variance at the student level (Opdenakker & Van Damme, 2000; Martinez-Fernandez, 2005).

The HLM software package (Raudenbush, Bryk, Cheong, and Congdon, 2004) was used to estimate these multilevel models. In all cases student weights were applied to the estimation in order to take into account the sampling framework used in TIMSS. As mentioned previously, TIMSS does not survey representative samples of teachers and thus weighted model estimates involving teacher- or classroom-level effects are representative at the student level (that is, they refer not to teachers generally, but to the teachers of nationally representative samples of students; see Martin, 2005).

⁶ Algebra classes were defined as those for which teachers reported spending 75% or more of their instructional time on algebra topics. The survey item that enabled this distinction between Algebra and general math classes existed only in the 2003 survey, and that year the U.S. was the only one of the 4 countries where there were any 8th grade classes that met this criterion.

We first estimate unconditional multilevel models to partition the total variance in student reports of instructional practice into within- and between-classroom components (i.e., to estimate the intra class correlation or ICC), separately for each of four countries and 13 types of instructional practices (52 models in total). Similarly, 24 unconditional models partitioned the variance in student achievement within and between classrooms, one for each of four countries, and six dependent variables (overall mathematics scores, and scores on each of the five content strands in TIMSS). At the student level the equation treats the achievement of student i in classroom j as a function of the classroom mean level of achievement (β_{0j}) and each student's deviation from this mean (ε_{ij}); the classroom mean in turn is modeled as a function of the grand mean score (γ_{00}) and a classroom-specific deviation (u_{0j}).

$$MTOT_{ij} = \beta_{0j} + \varepsilon_{ij} \quad (1)$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (2)$$

where ε , and u are assumed to be normally distributed, with mean zero and variances σ^2 , τ_β respectively. Thus, τ_π and σ^2 reflect respectively the distribution of student achievement within and between classrooms.

In addition to incorporating sampling weights, the models of student achievement further took into account the variation introduced by the use of a multiple *plausible values* methodology. TIMSS tests each student with a different set of items to provide better achievement aggregates by country. However, this also introduces a degree of uncertainty in estimating the achievement of individual students. To account for this uncertainty five scores are generated for each student

from a posterior distribution of plausible scores (as opposed to single maximum likelihood estimates; see Martin, 2005). HLM estimates the models five times separately, one for each plausible value or score, and the five plausible parameters are then pooled together to estimate the final model parameters (see Raudenbush, Bryk, Cheong, & Congdon, 2004). An adjusted standard error for each pooled parameter is estimated from the formula:

$$SE = \sqrt{U^* + (1 + M^{-1})B_m} \quad (3)$$

where B_m is the variance of the five plausible parameters, U^* is the average estimated variance of these parameters, and $M=5$. The resulting pooled parameter estimates and tests of significance take into account the measurement error introduced by the use of plausible values instead of single scores, and are the equivalent of the jackknifed estimates used to generate descriptive statistics in TIMSS (Willms & Smith, 2005).

We then estimated *conditional* models to gauge the degree of relationship between instructional practices and student achievement, controlling for student and teacher background. At the student level the covariates included gender, age, number of books and appliances at home (as proxies for socioeconomic status), and home use of a language different to that of the test. In addition, three items were included in the student-level model to capture students' perceptions of the frequency of instructional practices: lecture style presentations, working problems independently, and use of calculators. For teachers we included years of experience, highest level of education attained, and classroom size as controls. Finally, the teacher-level model included the composites created previously to reflect teacher reform-oriented perceptions and use of reform-oriented instructional practices in the classroom. The level-1 equation

incorporated the student covariates ($X_{1...p}$) as predictors of student achievement, so that β_{0j} was adjusted for student background (see equation 4). At the classroom level (level-2), these adjusted classroom means were modeled as a function of classroom and teacher predictors ($M_{1...q}$), while u_{0j} represented residual classroom variance. Finally, the parameters representing the *effects* of student-level covariates ($\beta_{1j}... \beta_{pj}$) were held constant across schools (as depicted in equation 6).

$$MTOT_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \dots + \beta_{pj}X_{pij} + \varepsilon_{ij} \quad (4)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}M_{1j} + \dots + \gamma_{0q}M_{qj} + u_{0j} \quad (5)$$

$$\beta_{1j} = \gamma_{10} \quad (6)$$

Importantly, while these models controlled for a few important student and teacher covariates included in the background questionnaires, the nature of the data collected by TIMSS (i.e., cross-sectional, non-experimental design with no prior-year scores for students) does not support causal interpretations. For example, if a model shows a strong relationship between instruction and achievement, it is not possible to determine with certainty whether instruction affects student achievement, or whether high-achieving students receive certain kinds of instruction, or both. Our analyses can thus only be interpreted as indicative of relationships between variables and perhaps point at interesting avenues for more rigorous causal investigations.

Findings from Analyses of TIMSS Survey and Achievement Data

In this section we present the results of several sets of analyses of the survey and achievement data. We begin with descriptive information on changes in teachers' reported

practices in the U.S, and elsewhere, and then investigate how teacher reports of instructional practice relate to their perceptions about reform instruction and to student reports of practice in the classroom. Finally, we examine the extent to which these reported practices are associated with achievement.

How has Teachers' Use of Reform Practices Changed Over the Last Decade?

One way to understand the extent to which mathematics reforms took hold U.S. classrooms during the 1990's is to examine changes in teachers' use of reform-oriented instructional practices over the course of the last decade. Of course, such changes cannot be assumed to represent direct effects of the reforms, but they can serve a useful illustrative purpose.

Figure 2 presents the average responses for 8th-grade teachers in each country and year for four items that were included in identical or very similar forms in all three waves.⁷ The first three correspond more strongly to a reform-oriented approach, whereas the last, emphasis on computational skills, is generally considered an indicator of a traditional approach to instruction (though it is important to note that it also indicates a particular content focus that might or might not be related to pedagogical style). Figure 2 shows sizable increases in the average frequency with which U.S. teachers reported using each kind of instructional practice between 1995 and 1999, whereas in most cases the use in other countries did not change substantially.⁸ Between

⁷ As noted previously, only a small proportion of the total number of instructional practice survey items remained consistent through the 1995, 1999, and 2003 TIMSS waves.

⁸ Standard Errors for the yearly averages for each country are presented in Table A.2 in the Appendix and range from 0.02 to 0.09. While providing a sense of the precision of the estimates, confidence intervals can produce overly conservative test of significance (Schenker & Gentleman 2001). Thus, the statistical significance of differences across countries and years can be assessed through a simple pooled standard error of differences as given by the formula: $(Q_1 - Q_2) \pm 1.96 \sqrt{SE_1^2 + SE_2^2}$. While the specific critical value naturally differs for each pair of estimates, differences in the 0.15 to 0.20 range and above are generally significant.

1999 and 2003, reported frequency of *explaining* and *problem solving* continued to increase among U.S. teachers, while the reported emphasis on equations and computational skills remained relatively stable. As a result of the differences in trends among the four countries, the U.S. went from being the country in which the use of these practices was least frequent in 1995 to being among those where reported use was most frequent in 2003. While a causal interpretation of these summaries is not possible, they suggest that something about the way mathematics was taught in the U.S. changed significantly during the 1990's. One possibility is that the changes reflect at least in part the influence of mathematics reforms under way in the country during the same period. At the same time, these trends suggest that increases in the reported use of instructional practices in the U.S. were not limited to reform-oriented items: practice of *computational skills* also increased during the same period.

Another way to examine changes in reform orientation is to consider the number of topics taught over the course of a year. Schmidt and colleagues (e.g., Schmidt et al., 1997) have reported that mathematics curricula in the U.S. tend to include a relatively large number of topics, each of which is taught at a fairly shallow level. Many of the reform-oriented mathematics curricula were designed to promote increased depth by focusing on fewer topics within each course than are typically covered in a more traditional curriculum, and by integrating key mathematical ideas into the curriculum throughout a student's mathematical education.

Although the TIMSS surveys gathered detailed information on topics taught, the extent to which it is possible to use the survey data to understand differences in depth and breadth is limited. Here we present one analysis that is intended to shed some light on how breadth of coverage might differ across years and countries, but much more information would be needed to understand the nature of the mathematics content to which students are exposed.

Table 3 presents the average proportion of topics (out of the total number listed in the survey) that teachers in each country reported spending any amount of time teaching during that year for each of the three TIMSS administrations. It points to a slight decrease in breadth for U.S. teachers from 1995 to 1999, a result that would be consistent with the changes in practice between 1995 and 1999 in that it suggests greater fidelity to reform principles. By 2003, however, the numbers across countries increased back to 1995 levels.⁹

Irrespective of trends over time, the overall results indicate that each year U.S. teachers reported teaching a higher average number of topics than teachers in Japan, the Netherlands and Singapore. On the other hand, the table also indicates that this breadth of coverage does not apply to algebra classes, where U.S. teachers reported teaching about half as many topics as in general mathematics classrooms (a smaller proportion than in the Netherlands and Singapore and roughly the same number as in Japan). While it is impossible to draw any strong conclusions from these results, they do suggest that the relatively high-achieving group of U.S. 8th graders who take algebra are exposed to instruction that is more closely aligned with reform principles in its focus on a smaller number of topics than is typical in U.S. mathematics classrooms.

Are Teachers' Perceptions Consistent with Their Practices?

Table 4 presents the correlations between the teacher-reported practice and perception composites created after the factor analyses results discussed previously (separately for each country). In the United States *traditional* perceptions are negatively correlated to both *reform* perceptions (-0.23) and *reform-oriented* practice (-0.26), suggesting the two may indeed be

⁹ As mentioned before, the specific wording of the questions and scales used to report coverage of topics changed slightly each year. Although the numbers should be comparable after recoding the scales, it is possible that any trends might reflect changes to the teacher survey in addition to (or instead of) real changes in *breadth of coverage*.

perceived by teachers as being somewhat in opposition one another. Teachers' use of reform-oriented practices is positively associated with their emphasis on mathematics content.

Interestingly, the positive correlation between reform practices and content is observed in the other three countries, but not the negative correlation between traditional and reform perceptions. While we have no information that would shed light on the source of these differences, it is possible that one result of the math wars in the U.S. may have been to make teachers perceive reform and traditional approaches as inconsistent with one another.

Do Student and Teacher Descriptions Match?

Because TIMSS 2003 gathered information about practices from both teachers and students, it is possible to compare the responses of these groups to evaluate the degree to which students' reports of instructional practices vary within and across classrooms, and whether they match the reports of provided by their teachers.

In comparing instructional practices across countries it is informative to examine variability in student reports across classrooms (or similarly, the extent to which students in the same classroom provide consistent information about instructional practices). Table 5 shows the intra-class correlation (ICC) in student reports of instructional practices for each of the 4 countries—i.e. the proportion of variance between classrooms estimated from an unconditional multilevel model. Across countries variation in the use of *traditional* or skill-focused practices across classrooms is generally greater than in practices we group under the *reform-oriented* label. Moreover, the table points to a greater degree of between-classroom variability in student reports of instructional practices in the United States (and to a lesser extent in the Netherlands) compared to countries like Japan and Singapore. While this is an admittedly coarse measure, it suggests that the academic experiences of 8th grade students may vary more considerably across

mathematics classrooms in the United States, compared to students in other countries, who report more consistent instructional practices across classrooms. This result is consistent with the more heavily tracked nature of 8th grade mathematics education in the U.S. (as an example, the results in the previous section suggest that algebra classrooms are very different from regular mathematics classrooms).

Table A.3 in the appendix presents the average frequency of instructional practices reported by teachers and students. Compared to teachers, students across countries generally reported significantly more frequent use of *traditional* instructional practices or skills in the classroom (and of practices where they work without teacher supervision). By contrast, teachers tended to report more frequent use of *reform*-oriented instructional practices than their students. This could reflect teachers' better understanding of the nature and goals of different instructional practices, socially desirable answers, or a combination of both. Finally, Table 6 presents the correlation between student reports of instructional practices and those of their teachers.¹⁰ The results in the United States point to a small to moderate correlation between student and teacher reports of instructional practices, a weaker relationship than that reported in other studies (see, e.g., Herman & Abedi, 2004, and Herman, Klein, & Abedi, 2000. Muthén et. al., 1995 found enough overlap between student and teacher reports to suggest that collecting information from both could be unnecessary). In other countries the correlations are substantially lower, perhaps reflecting the limited variability of these classroom aggregates reported in Table 5. In the absence of additional information with which to validate students' and teachers' reports, it is impossible to determine the source of these low correlations or to make conclusions about which

¹⁰ For each instructional practice item student reports of frequency were aggregated by classroom, and these classroom averages were correlated to those of the teacher.

source has greater validity. We include both in the multilevel models discussed in the next section.

Is Reform-Oriented Instruction Related to Student Achievement?

Table 7 presents the average mathematics scores of 8th grade students in each country (total mathematics score and five subscores), along with the proportion of variance across classrooms estimated from the unconditional multilevel model of student achievement. In addition to widely reported differences in the average achievement of students across countries, the table indicates that variation in mathematics achievement across classrooms is greatest in Singapore and the United States. Consistent with results reported in other studies (see, e.g., Koretz, McCaffrey, & Sullivan, 2001) between half and three quarters of the total variance in mathematics scores in these countries is between classrooms. By comparison, Japanese classrooms are much more homogeneous in terms of average mathematics achievement.

The next modeling step is to investigate what characteristics of classrooms could help explain the greater variance in student mathematics scores in the U.S. (and Singapore)¹¹. These characteristics could include various kinds of resources, teacher training and background, and aggregate characteristics of the students, but could also be related to variation in curricula and instructional practices. Tables 8 and 9 present the results of multilevel models of student achievement that investigated this question. As discussed before, the results cannot be interpreted in causal terms, but merely as indicative of relationships that exist between student achievement and features of the classroom context in each country.

¹¹ We conducted several tests of sensitivity of these models. In particular, we tested the effects of including the content emphasis items and found that these did not affect conclusions about other predictors of student achievement; these predictors were dropped from the final models. In the U.S. we also examined the effects of using individual teacher practice items rather than the composite scales, and found a consistent lack of relationship with student achievement.

Table 8 shows the results for the total mathematics score. Across countries the models consistently showed a significant advantage for boys and for students of higher socioeconomic status (proxied through the number of books and appliances in the household), and significant disadvantages for non-native speakers. In the U.S. and the Netherlands there was also a negative relationship with student age (likely a proxy for grade retention), while the age coefficient was positive in Japan. On the other hand, the relationship between student achievement and activities in the classroom (as reported by the students themselves) was inconsistent across countries. In the U.S. achievement was higher among students who reported working on problems on their own more often, and lower for students who reported listening to lecture style presentations more often.¹² The positive association with working on problems independently was also observed in other countries (and was strongest in Japan); however, lecture style presentations were not significantly related to student achievement in other countries. The effect sizes of significant raw coefficients were often small, however.¹³ For example, the model-based estimate of the standard deviation for total mathematics scores in the United States was 79.8; thus, an effect of 6.58 for working problems independently represents only about 8% of a standard deviation. Even for students at the extremes of the 3-point instructional practice scale the effect size would be only about a quarter of a standard deviation.

Unlike the results at the student level, Table 8 shows no significant association in the United States between student achievement and teacher background and experience, classroom size, and reform perceptions. For reform-oriented practice the coefficient was positive but non-

¹² The original instructional practice scales were inverted in these models so that higher values represent more frequent use of instructional practice.

¹³ Overall *d* effect sizes were estimated (Cohen, 1988), by standardizing each parameter with respect to the pooled within- and between-classroom variance estimated in each multilevel model. This gives a general sense of the importance of the effects with respect to the scale of the dependent variable. Alternatively level-specific effect sizes can be estimated through the *proportion of variance* explained by a variable at each level in the model.

significant except in Singapore, where this indicator was positively and significantly related to achievement.

Table 9 presents the results of a model examining achievement in algebra. Given the differentiation of curriculum in some U.S. classrooms this model included two additional predictor variables: one was the proportion of time the teacher spent teaching algebra topics, and the second was the interaction between this indicator and the frequency with which the teacher reported using reform-oriented instructional practices in the classroom.

Except for the absence of a difference in the performance of boys and girls, the results at the student level in Table 9 for algebra scores generally resembled those for the overall mathematics scores in Table 8. In the U.S. student achievement in algebra was positively related to working on problems independently (a relationship most strongly observed among Japanese students), and negatively related to lecture-style presentations (a relationship not observed in other countries). At the classroom level, teacher reported frequency of reform-oriented practices was positively related to student achievement in algebra; importantly, in this case the associated effect size is more considerable at over a quarter of a standard deviation. The same strong relationship between reform practices and student achievement was observed in Japan and Singapore. Finally, no significant relationships were observed between student achievement and time spent teaching algebra in the classroom, or the interaction of time in algebra and frequency of use of reform practices.

Tables A.4 to A.7 in the Appendix present the results of multilevel models identical to that in Table 8 for the relationships between student and classroom predictors and the other four mathematics sub-scores available in TIMSS (data, analysis and probability; number sense; geography; and measurement). While the relevant coefficients were always positive, a constant

in all these models was the absence of a significant relationship in the U.S. between student achievement and the frequency of so-called reform-oriented instructional practices reported by the teacher. In fact, the reform practice coefficient is only significant in the case of Singapore; in that country reform practice is significantly related to student achievement for all mathematics sub-scores.

Summary

Together, the analyses presented in this paper provide some evidence of changes in instructional practices that corresponded with the implementation of new mathematics curricula during the 1990's in the U.S. Relevant findings include an increase in teachers' reports of the frequency with which they engaged in some instructional practices that are consistent with the math reforms, and a decrease in the number of mathematics topics presented over the course of a year. While the findings point to more stability in the practices reported by teachers in three other countries we investigated (Japan, the Netherlands and Singapore), there is also the possibility that something about how teachers in the U.S. describe their practices changed over time while the practices themselves did not. The video studies conducted in 1995 and 1999 do not show an increase in reform-oriented instruction during that period, which suggests this last possibility should be given consideration.

Overall, the evidence of changes in U.S. teachers' use of reform practices is very limited and can only be interpreted as suggestive, among other reasons because of the limited number of practices compared. Even assuming the results reflect real changes in classroom practices, though, it is not clear what the implications of these changes might be for student achievement. The bulk of our cross-sectional analyses of 2003 student achievement data do not suggest a

strong relationship in the U.S. between more frequent use of reform-oriented practices and higher student achievement.

Although it is hindered by a number of weaknesses in the design and analysis (including a design that does not support strong causal inference and a lack of consistency between student and teacher reports of practice) this study adds to a growing body of literature examining relationships between instruction and achievement. Understanding these relationships can help shape the mathematics education research agenda: A relationship that is observed in a variety of contexts would suggest the need for richer and better-designed research to understand the source of that relationship, whereas a consistent finding of no relationship might lead to the conclusion that research resources would be more wisely spent on other topics of research. Combined with the other studies reviewed earlier, the results presented here suggest that investing resources in promoting the kinds of instructional practices examined in this set of studies is unlikely to lead to changes in student achievement. At the same time, the somewhat ambiguous results observed in this study, combined with the limitations discussed in the next section, suggests that better methods for measuring instructional practice are needed in order to understand more fully how the 1990's mathematics reforms influenced instruction and student achievement.

Limitations

The results of the analyses presented here must be interpreted very cautiously. There are a number of limitations of the data and analytic approaches used. Some of these are unique to TIMSS but others characterize virtually all studies that examine instructional practices using survey data. We discuss three sets of limitations: (1) validity problems with survey-based

measures of instructional practice; (2) the need to go beyond instructional practice to understand reform implementation; and (3) other challenges associated with large-scale projects like TIMSS.

Survey-based Measures May Not Accurately Capture Information About Practices

Perhaps the most significant problem in gauging the impact of reform-oriented instructional practice has to do with the specific aspects of reform that can be measured using instruments such as the TIMSS surveys. While the kinds of instructional activities examined in this paper are in many ways consistent with reform-oriented mathematics instruction, they by no means provide a complete picture, and in some instances might provide misleading information. In discussing the implementation of standards-based curriculum and instruction, Burrill (2001) acknowledges “In some cases, there has been a focus on the ‘trappings of the reform’—cooperative groups, manipulatives, hands-on activities—with little attention of mathematics as the focus of instruction” (p.37). Even when supplemented with information about curriculum or topics taught, most surveys fail to capture information about the extent to which teachers are implementing the core principles of reform, such as a focus on students’ thinking.

We also know from other research that teachers do not always interpret survey items in the ways the developers intend. Spillane and Zeuli (1999) used TIMSS survey items to identify high-reform teachers, but found that of the 25 teachers whose survey responses suggested high levels of reform practice, only four were judged by classroom observers to be teaching in a way consistent with reform ideals. Hill’s (2005) validation of instructional logs for elementary mathematics teachers showed that the accuracy of responses was affected by teachers’ knowledge of terms and conventions, and interviews conducted by Le et al. (2006) revealed that teachers sometimes mentally rephrased survey questions in ways that changed their meaning.

Low degrees of consistency between similar items on surveys and logs or between responses to the same survey administered at two time points also raise concerns about the validity of survey items (Mayer, 1999; Smithson & Porter, 1994). However, Meyer (1999) found that consistency over time, as well as correlations between survey responses and observations, were relatively high when using a composite of several items rather than a single item.

The problem of differences in interpretation is likely to be especially acute for between-country comparisons. As shown earlier, factor analyses of student and teacher reports suggest that the instructional practice items do not always function consistently across countries, perhaps because of contextual differences involving language, curriculum, and other factors. Other research indicates that teachers in different countries often interpret survey questions differently (Schulz, 2006), and differences have also been reported for student surveys (Walker, 2006).

Another limitation of relying on teacher surveys is that they typically fail to capture information about differences in students' experiences within a classroom (see Gamoran, 1991). Teachers spend much of their time interacting with small groups or individuals, and even when engaged in whole-class activities might vary the instruction provided to different students through differences in the types of questions they ask or the feedback they provide. Individual students' experiences will also vary as a function of the characteristics and experiences they bring to the classroom, such as their level of engagement with the material (Floden, 2002). It is impossible to fully understand cross-country differences in achievement without considering the many contextual and cultural factors that influence the academic experiences of students in classrooms (Stevenson, Lee, & Stigler, 1986). Our analyses of student reports of instructional practice suggest that these experiences may vary considerably for students in different classrooms in the United States, while classrooms in Japan or Singapore tend to be more

homogeneous. Also, exposure to content and instruction in earlier grades is likely to influence test scores and is an important aspect of opportunity to learn (Floden, 2002). Thus a truly accurate measure of a student's exposure to instructional practices is very difficult to obtain with traditional survey methods and with the resource constraints typical of large-scale cross-sectional data collection efforts.

Finally, however closely we think teachers' or students' survey responses resemble their actual activities in the classroom, most surveys fail to provide any information on the *quality* of the practices in which teachers (and students) engage. There are a variety of ways to present complex problems or use cooperative groups, only some of which represent desirable instructional practice. This is an area where observational techniques such as those used in the TIMSS video studies can be especially valuable (Hiebert & Stigler, 2000), but even these types of studies are likely to provide a less-than-complete picture of the quality of instruction to which students are exposed over the course of an academic year. Promising approaches for measuring instructional quality are becoming available for researchers (e.g., Matsumura et al., 2002; Borko et al., 2005) and are worth considering as ways to collect better information in large-scale studies.

Data on Practices Provides Incomplete Information about Reform Implementation

Even if the validity of survey items were not a concern, it is clear that survey data about instructional practices provide only a limited understanding of what is actually happening in classrooms. Herman and Klein (1997) discuss three aspects of opportunity to learn, all of which can affect student performance: curriculum content, instructional strategies, and instructional resources (see also Haertel, 2003). Others argue that the cognitive demands of the curriculum are

especially important and constitute a form of opportunity that should be examined (see, for example, Shepard, 2001).

In particular, curriculum and instruction interact in complex ways to influence achievement, and an examination of one without any consideration for the other is likely to be incomplete. Evaluations of specific curricula often fail to take into account the ways in which reform-based curricula can become distorted in practice, so that the instruction students receive fails to match the goals of the curriculum developers. Similarly, evaluations of instructional practices should be informed by knowledge of what curriculum is in place. The relationship between practices and achievement has been shown in some cases to depend on the specific curriculum being implemented. McCaffrey et al. (2001) found a relationship between reform practices and achievement in high school classrooms that were implementing reform-oriented mathematics curricula, but not in classrooms that were using more traditional textbooks. Other studies have demonstrated the importance of considering pedagogy in the context of rigorous content when examining effects of instruction on achievement (Gamoran et al., 1997; Porter, 1998). To be of maximum utility, large-scale surveys should include detailed information both on practices and on curriculum, including the specific materials being used. Measuring the content of instruction is important for understanding how mathematics reforms have been implemented. Information on pedagogical strategies does not tell us, for example, whether teachers have integrated algebra into their instruction of early-grade students. But measuring content related to reform-oriented goals through surveys is not easy, and might be subject to even greater problems stemming from lack of shared understanding of terms than measures of practice (Burstein et al., 1995; Hill, 2005). Although TIMSS questionnaires include some information

about topics taught, it does not provide a sufficient level of detail to evaluate the implementation of core reform principles.

Some existing surveys provide more detail on instructional content. An example is the Surveys of Enacted Curriculum (SEC),¹⁴ which not only ask about content but also about the depth and rigor of that content. Surveys like SEC are probably more likely to provide an accurate understanding of what is happening in the classroom, but may pose a substantial response burden. Balancing the desire for accurate information about pedagogy, content, and depth with the need for efficient data collection is likely to remain one of the most significant challenges for studies like TIMSS.

Research Design and Measurement of Achievement Create Additional Challenges

Two other issues are worth mentioning, one having to do with limitations of the study design and the other with the measurement of student achievement. One of the primary limitations of most research conducted with large-scale databases is the inability to utilize research designs that support strong causal inference. In the absence of random assignment of teachers to the use of reform practices (or, perhaps more plausibly, to reform-oriented curricula), there is a high likelihood that teachers' use of those practices will be associated with other, unmeasured teacher characteristics and experiences. This confounding hinders our ability to make causal claims about the effects of reform practices on student achievement. This problem is especially severe in studies such as TIMSS that cannot follow individual students over time. Without the ability to examine prior achievement and prior exposure to reform practices it is impossible to conclude with any certainty that a correlation reflects a causal relationship.

¹⁴ See http://www.ccsso.org/projects/Surveys_of_Enacted_Curriculum/7804.cfm (retrieved 9/14/06)

An additional concern stems from the need to consider the appropriateness of the achievement outcome for measuring instructional effects. As Linn (2002) notes, “The more specific the purpose, the more homogeneous the population of students, and the narrower the domain of measurement...the easier is the task of developing measures that will yield results that support valid interpretations and uses” (p.27). The TIMSS assessments, by contrast, epitomize a situation characterized by multiple purposes, heterogeneity of students, and broad domains of measurement. These assessments were never validated for the purposes of detecting effects of instructional practice, and there is evidence of cross-country differences in their psychometric properties (Ercikan & Koh, 2005; Grisay et al., 2006). Moreover, the validity of the TIMSS tests as measures of the effects of instructional approaches depends in part on the degree to which they are aligned with the curriculum to which students are exposed. The importance of this alignment is evidence in the finding that country rankings can change substantially when only certain items are included in the achievement measure (Schmidt, McKnight, & Raizen, 1997).

Finally, exclusive reliance on multiple-choice items can produce an incomplete picture of student attainment in mathematics, with a relatively smaller weight assigned to dimensions of attainment more likely to be influenced through reform-instruction. Researchers have suggested that open-ended assessments or assessments that emphasize problem solving, understanding, and application may be more likely to demonstrate a positive relationship with reform-oriented instructional practices (see e.g. Saxe, Gearhart, and Seltzer, 1999; Cohen and Hill, 2000; and Thompson and Senk; 2001; Hamilton et al., 2003; and Le et al., 2006), which emphasizes the importance of examining differences in relationships across outcome measures.

What's Next?

After reviewing existing literature and presenting results of a additional analyses, the answer to the question posed in the title of this paper could be summed up as “not very much.” Although TIMSS and other international studies have provided valuable information on a range of topics, it is difficult to find much evidence in the data to support or refute the variety of claims that have been made about 1990’s mathematics reform. Nonetheless, the ongoing TIMSS data collection provides an opportunity to enhance our understanding of the role that instructional practices play in influencing student achievement in mathematics, and given the tremendous resources currently devoted to international surveys it is worth thinking about ways to make the most effective use of the data these surveys generate.

Although valuable information is obtained from the video studies, paper-and-pencil (or eventually on-line) surveys will continue to be relied upon for large-scale data gathering. The kinds of items included in the TIMSS surveys continue to dominate large-scale data collection efforts, but there are initiatives under way to improve the quality of these measures and to develop measures of other important constructs such as teacher knowledge (the work of the Study of Instructional Improvement stands out; see, e.g., Hill et al., 2005).

A few specific directions are especially important for researchers and developers of international education surveys to consider. The discussion in the previous sentence suggests that measures of general pedagogical strategies, such as whole-class or small-group instruction, should be supplemented with information about the content of the lessons taught, such as how algebraic concepts are integrated into instruction throughout the elementary grades. Valid measures of content probably need to include data-collection activities other than surveys that ask teachers to report on the topics they taught. Classroom observations, examination of

curriculum materials, and collection of artifacts such as lesson plans or classroom assessments are some examples of strategies that have been used to gather this information. These approaches vary in cost, feasibility, and fidelity to what is actually happening in classrooms.

Another important consideration for data collection involves whether to gather information about individual students' opportunities to learn. As discussed earlier, even within the same classroom, students' experiences may differ dramatically as a result of a number of factors: their own engagement and prior levels of knowledge and understanding, the attributes their peers bring to the class (which might be especially relevant when small-group instruction is used), and teachers' actions. This type of information could be collected from teachers (e.g., by asking teachers to report on the instruction provided to an individual child rather than the whole class) or students, and could also be gathered through observations and examination of student work.

A final issue, which has not yet been discussed in this paper, involves the role of external, contextual factors such as the existence of a national curriculum or high-stakes testing policies. TIMSS gathers extensive data on some of these contextual factors, and this information has contributed to our understanding of cross-nation differences in achievement. For studies focusing on mathematics education within the U.S., contextual information collected at the state and district levels could enhance our understanding of classroom practices and student achievement. For example, states vary in the rigor and clarity of their content standards, in the difficulty and item formats of their state tests, and in the kinds of consequences they attach to test scores. Within each state, districts vary on such dimensions as the degree to which they mandate specific programs and the kinds of supports they provide for implementation of curriculum and standards. By gathering information on the various supports and sources of pressure affecting

teachers' work, it might be possible to attain a more-refined understanding of the interactions among curriculum, instruction, and achievement.

In the end, large-scale survey methods such as those used in TIMSS are inherently limited; they do not lend themselves to research designs that are best equipped to support causal inference, and they can't possibly provide all of the contextual information needed to understand the mechanisms through which instructional practices influence achievement. But they provide an unparalleled opportunity to gather information across a wide range of districts, states, and countries, and therefore can play an important role in efforts to build the scientific research base in education. It is likely that such surveys will be with us for some time, and recent research provides a number of lessons that can be applied to survey development for future rounds of TIMSS and other large-scale studies.

Tables and Figures

Table 1. 2003 Teacher Survey Items Addressing Reform-Oriented Instruction

In teaching mathematics to the students in the TIMSS class, how often do you usually ask them to do the following?
Work on problems w/no immediately obvious method of solution
Work together in small groups
Relate what they are learning in mathematics to their daily lives
Explain their answers
Decide on their own procedures for solving complex problems

Table 2. Factor Analysis of Perceptions Items in the Teacher Survey. (U.S.)

	Factor Loadings	
	1	2
More than one representation should be used in teaching a math topic	0.71	-0.07
Solving math problems involves hypoth., estimating, testing, modifying findings	0.76	-0.16
There are different ways to solve most mathematical problems	0.61	-0.43
Modeling real-world problems is essential to teaching math	0.71	-0.16
Math. should be learned as algorithms/rules that cover all possibilities	-0.06	0.78
Learning mathematics mainly involves memorizing	-0.18	0.69
Few new discoveries in mathematics are being made	-0.23	0.56

Table 3. Proportion of topics in the 8th Grade Teacher survey taught (By Year)

	Proportion of topics taught		
	1995	1999	2003
Japan	28.8	29.1	24.5
Netherlands	42.7	36.2	41.5
Singapore	38.9	24.4	39.2
USA	49.1	42.6	46.1 (Overall) 24.0 (Algebra)

Table 4. Correlations Among Teachers' Practices and Perceptions By Country

	Practice		Perceptions	
	Content	Reform	Traditional	Reform
Japan				
Content Emphasis	1.00			
Reform Practice	0.12	1.00		
Traditional Perceptions	0.10	-0.01	1.00	
Reform Perceptions	0.27*	0.20*	0.15	1.00
Netherlands				
Content Emphasis	1.00			
Reform Practice	0.32*	1.00		
Traditional Perceptions	-0.16	-0.03	1.00	
Reform Perceptions	0.18*	0.09	-0.08	1.00
Singapore				
Content Emphasis	1.00			
Reform Practice	0.41*	1.00		
Traditional Perceptions	0.00	-0.07	1.00	
Reform Perceptions	0.20*	0.27*	0.06	1.00
United States				
Content Emphasis	1.00			
Reform Practice	0.22*	1.00		
Traditional Perceptions	0.08	-0.26*	1.00	
Reform Perceptions	-0.01	0.33*	-0.23*	1.00

Note: * (p<0.05)

Table 5. ICCs for Student Reports of Instructional Practice, by Country
(Percentage of Variance Between Classrooms).

	Japan	Netherlands	Singapore	USA
Traditional				
a) We practice add, subtract, multiply, divide without calculator	1.1	9.4	3.0	10.1
b) We work on fractions and decimals	2.5	3.6	4.7	10.3
i) We review our homework	19.9	38.4	8.2	34.4
j) We listen to the teacher give a lecture-style presentation	5.1	25.5	3.9	7.2
l) We begin our homework in class	9.9	12.7	12.5	38.1
m) We have a quiz or test	19.2	6.5	7.8	20.1
Average	9.7%	16.1%	6.7%	20.1%
Reform				
c) We interpret data in tables, charts, or graphs	1.4	6.4	3.0	14.8
d) We write equations and functions to represent relationships	3.8	9.8	9.5	14.5
e) We work together in small groups	22	38.2	20.9	39.5
f) We relate what we are learning in math to our daily lives	3.7	6.7	5.1	11.2
g) We explain our answers	12.3	15.2	4.8	10.9
h) We decide our own procedures solving complex problems	4.1	3.5	5.7	5.8
k) We work problems on our own	3.1	9.7	10.1	9.7
Average	7.2%	12.8%	8.5%	15.2%

Table 6. Correlations Between Student and Teacher Reports of Instructional Practices.
(By Country)

	Correlations			
	Japan	Netherlands	Singapore	USA
a) Practice add,subtract,multiply, divide without a calculator	0.00	-0.03	0.11	0.33
b) Work on fractions and decimals	-0.05	-0.06	-0.01	0.21
c) We interpret data in tables, charts, or graphs	-0.05	0.09	-0.04	0.22
d) We write equations and functions to represent relationships	0.07	0.18	0.15	0.23
e) We work together in small groups	0.24	0.25	0.14	0.55
f) We relate what we are learning in math to our daily lives	0.08	0.15	-0.01	0.26
g) We explain our answers	0.22	0.16	0.07	0.22
h) We decide our own procedures solving complex problems	0.04	0.12	0.15	0.19
k) We work problems on our own	-0.01	-0.01	0.06	0.02

Table 7. Average Student Mathematics Achievement and ICCs, by Country

	Japan	Netherlands	Singapore	USA
Average Achievement				
Total Mathematics Score	568.0	530.5	602.9	505.1
Data, Analysis, Probability	571.2	555.1	577.2	527.8
Fractions, Number Sense	554.5	533.1	615.1	508.4
Geometry	585.3	507.7	577.3	472.9
Algebra	566.1	508.5	587.2	510.6
Measurement	557.3	543.3	608.1	496.2
Proportion of Variance between classrooms				
Total Mathematics Score	12.0	42.5	76.8	63.4
Data, Analysis, Probability	8.1	35.0	59.9	51.4
Fractions, Number Sense	11.7	41.1	73.6	62.1
Geometry	8.8	34.9	70.4	53.4
Algebra	10.1	39.5	65.9	59.3
Measurement	11.4	39.1	70.8	57.7
Sample Sizes				
N of Students	4147	2397	5214	6841
N of Classrooms	125	103	283	346

Table 8. Multilevel model of Student Achievement (Total Mathematics Score)
Coefficient and *Effect Size* estimates.

	Japan		Netherlands		Singapore		USA	
	Coeff.	E.S.	Coeff.	E.S.	Coeff.	E.S.	Coeff.	E.S.
Classroom Mean, G00	567.74		533.00		602.61		504.52	
Tchr. Yrs of Exp., G01	-0.32	0.00	0.78	0.01	0.15	0.00	0.48	0.01
Tchr. Highest Ed., G02	15.26	0.20	14.03*	0.19	-0.31	0.00	0.26	0.00
Classroom Size, G03	0.86*	0.01	6.83*	0.09	0.99	0.01	-0.19	0.00
Tchr. <i>Reform</i> Practice, G04	7.49	0.10	-26.45	0.36	27.47*	0.14	12.31	0.16
<i>Reform</i> Perceptions, G05	6.62	0.09	22.08	0.30	-11.33	0.34	-9.37	0.12
Boy, B2	8.82*	0.11	13.08*	0.18	6.30*	0.08	7.48*	0.09
Diff. Language at Home, B3	-5.35	0.07	-8.56*	0.12	2.55*	0.03	-5.31*	0.07
Books at Home, B4	9.35*	0.12	2.74*	0.04	0.31	0.00	7.26*	0.09
Lecture-Style Present., B4	-0.23	0.00	1.03	0.01	0.26	0.00	-2.28*	0.03
Work on Problems Own, B5	31.96*	0.42	2.53	0.03	5.35*	0.07	6.58*	0.08
Use of Calculators, B6	-22.3*	0.29	0.19	0.00	-1.81	0.02	-1.77	0.02
SES (Home appliances), B5	11.49*	0.15	1.52	0.02	1.91*	0.02	0.19	0.00
Student Age, B1	7.28*	0.09	-7.92*	0.11	0.60	0.01	-8.13*	0.10

* Significant ($p < .05$)

Table 9. Multilevel model of Student Achievement (Algebra)
Coefficient and *Effect Size* estimates.

	Japan		Netherlands		Singapore		USA	
	Coeff.	E.S.	Coeff.	E.S.	Coeff.	E.S.	Coeff.	E.S.
Classroom Mean, G00	565.76		511.24		586.94		509.75	
Tchr. Yrs of Exp., G01	-0.40	0.01	0.60	0.01	0.30	0.00	0.46	0.01
Tchr. Highest Ed., G02	12.25	0.16	11.61*	0.15	1.61	0.02	-9.49	0.12
Classroom Size, G03	0.50	0.01	5.33*	0.07	0.59	0.01	-0.43	0.01
Time spent in Algebra, G04	-3.02*	0.04	4.28	0.05	-1.39	0.02	0.48	0.01
Tchr. <i>Reform</i> Practice, G05	47.63*	0.61	-25.17	0.32	74.45*	0.87	21.75*	0.29
<i>Reform</i> Perceptions, G06	7.98	0.10	36.55	0.47	-15.32	0.18	-10.83	0.14
Time in Algebra by <i>Reform</i> Practice Interaction, G07	-1.20*	0.02	0.78	0.01	-1.30	0.02	-0.36	0.00
Boy, B2	1.31	0.02	4.86	0.06	2.20	0.03	1.16	0.02
Diff. Language at Home, B3	-6.11	0.08	-9.33	0.12	2.90*	0.03	-2.78*	0.04
Books at Home, B4	7.95*	0.10	3.14*	0.04	-0.03	0.00	6.70*	0.09
Lecture-Style Present., B4	1.38	0.02	3.24	0.04	-0.78	0.01	-2.23*	0.03
Work on Problems Own, B5	27.77*	0.36	4.41	0.06	6.11*	0.07	5.64*	0.07
Use of Calculators, B6	-20.7*	0.27	-1.10	0.01	-2.38	0.03	1.14	0.01
SES (Home appliances), B5	12.73*	0.16	7.90	0.10	2.18*	0.03	0.14	0.00
Student Age, B1	8.68*	0.11	-5.69*	0.07	4.62	0.05	-7.94*	0.10

* Significant ($p < .05$)

Figure 1. Select Instructional Practice Items from the 1995, 1995, and 2003 versions of the 8th grade Mathematics Teacher Survey

1995 and 1999 Surveys	
Question	<i>In your mathematics lessons, how often do you usually ask students to do the following? :</i>
Survey Items	
Computational Skills	<i>Practice computational skills</i>
Problem Solving	<i>Work on problems for which there is no immediately obvious method of solution</i>
Equations Representation	<i>Write equations to represent relationships</i>
Explain Reasoning	<i>Explain the reasoning behind an idea</i>
2003 Survey	
Question	<i>In teaching mathematics to the students in the TIMSS class, how often do you usually ask them to do the following? :</i>
Survey Items	
Computational Skills	<i>Practice adding, subtracting, multiplying, and dividing without using a calculator</i>
Problem Solving	<i>Work on problems for which there is no immediately obvious method of solution</i>
Equations Representation	<i>Write equations and functions to represent relationships</i>
Explain Reasoning	<i>Explain their answers</i>

Figure 2. Frequency of use of various Instructional Practices (by Country and Year)



- Notes:
- 1) The samples of teachers in each country are independent and different each year.
 - 2) The scale ranges from 1=never / almost never, to 4=every lesson
 - 3) Some items in the teacher survey were not administered in the Netherlands and Japan in 1995.

Appendix

Figure A. Instructional Practice and Perceptions Items in the 2003 Student and Teacher Surveys

Student Survey:

Practices:

How often do you do these things in your mathematics lessons? (*Never to Every or almost every lesson*)

- a) We practice adding, subtracting, multiplying, and dividing without using a calculator
- b) We work on fractions and decimals
- c) We interpret data in tables, charts, or graphs
- d) We write equations and functions
- e) We work together in small groups
- f) We relate what we are learning in mathematics to our daily lives
- g) We explain our answers
- h) We decide on our own procedures for solving complex problems
- i) We review our homework
- j) We listen to the teacher give a lecture-style presentation
- k) We work problems on our own
- l) We begin our homework in class
- m) We have a quiz or test
- n) We use calculators

Teacher Survey:

Perceptions:

To what extent do you agree or disagree with each of the following statements? (*Disagree a Lot to Agree a lot*)

- a) More than one representation (picture, concrete material, symbols, etc.) should be used in teaching a mathematics topic
- b) Mathematics should be learned as sets of algorithms or rules that cover all possibilities
- c) Solving mathematics problems often involves hypothesizing, estimating, testing, and modifying findings
- d) Learning mathematics mainly involves memorizing
- e) There are different ways to solve most mathematical problems
- f) Few new discoveries in mathematics are being made
- g) Modeling real-world problems is essential to teaching mathematics

Practices:

In teaching mathematics to the students in the TIMSS class, how often do you usually ask them to do the following? (*Never to Every or almost every lesson*)

- a) Practice adding, subtracting multiplying, and dividing without using a calculator
 - b) Work on fractions and decimals
 - c) Work on problems for which there is no immediately obvious method of solution
 - d) Interpret data in tables, charts, or graphs
 - e) Write equations and functions to represent relationships
 - f) Work together in small groups
 - g) Relate what they are learning in mathematics to their daily lives
 - h) Explain their answers
 - i) Decide on their own procedures for solving complex problems
-

Table A.1. Factor Analysis of Math Teachers' Perceptions About Reform Principles (by Country).

	Japan		Netherlands		Singapore		USA	
	1	2	1	2	1	2	1	2
a) More than one representation should be used in teaching a math. topic	0.72	-0.04	0.68	-0.03	0.68	0.07	0.71	-0.07
b) Math. should be learned as algorithms/rules that cover all possibilities	0.55	0.13	-0.12	0.75	0.40	0.55	-0.06	0.78
c) Solving math problems involves hypoth., estimating, testing, modifying findings	0.67	0.11	0.13	0.66	0.67	0.21	0.76	-0.16
d) Learning mathematics mainly involves memorizing	0.24	0.78	-0.44	0.50	-0.16	0.67	-0.18	0.69
e) There are different ways to solve most mathematical problems	0.65	-0.10	0.51	-0.09	0.64	-0.24	0.61	-0.43
f) Few new discoveries in mathematics are being made	-0.31	0.64	-0.53	0.38	-0.01	0.67	-0.23	0.56
g) Modeling real-world problems is essential to teaching math	0.53	-0.22	0.63	0.31	0.65	-0.07	0.71	-0.16
	Reform	Trad.	Reform	Trad.	Reform	Trad.	Reform	Trad.
Cronbach Alpha :	0.59	0.11	0.32	0.45	0.60	0.31	0.65	0.45

Table A.2. Average use of instructional practices reported by teachers, by country and year
(Standard Errors in parentheses)

	Mean (Standard Error)		
	1995	1999	2003
Computational Skills			
Japan	--	2.83 (0.07)	2.73 (0.10)
Netherlands	--	2.38 (0.08)	2.00 (0.07)
Singapore	2.41 (0.09)	2.57 (0.08)	2.49 (0.04)
USA	1.94 (0.09)	2.92 (0.07)	2.69 (0.05)
Problem Solving			
Japan	2.18 (0.05)	2.48 (0.07)	2.49 (0.05)
Netherlands	--	2.24 (0.09)	2.10 (0.04)
Singapore	1.80 (0.05)	1.96 (0.06)	1.99 (0.02)
USA	1.36 (0.05)	2.06 (0.04)	2.29 (0.04)
Write Equations to Represent Relationships			
Japan	2.83 (0.06)	3.04 (0.05)	2.74 (0.04)
Netherlands	--	2.15 (0.06)	2.23 (0.05)
Singapore	2.14 (0.05)	2.28 (0.05)	2.41 (0.03)
USA	1.71 (0.08)	2.62 (0.04)	2.55 (0.04)
Explain/Reasoning Answers			
Japan	3.05 (0.06)	3.17 (0.06)	2.52 (0.05)
Netherlands	--	2.86 (0.09)	2.90 (0.09)
Singapore	2.45 (0.05)	2.55 (0.06)	2.64 (0.04)
USA	2.07 (0.09)	2.97 (0.05)	3.29 (0.04)

Note: The statistical significance of the difference between two estimates across countries

or years ($Q_1 - Q_2$) can be assessed through the formula: $(Q_1 - Q_2) \pm 1.96\sqrt{SE_1^2 + SE_2^2}$.

The difference is significant if the confidence interval around the difference does not contain zero.

Table A.3. Student and teacher reports of instructional practices. (Means By Country)

	Japan		Netherlands		Singapore		USA	
	Student	Teachers	Student	Teachers	Student	Teachers	Student	Teachers
Practice add, subtract, multiply, divide without using a calculator	2.98	2.73*	2.72*	1.97	2.84*	2.51	2.92*	2.70
Work on fractions and decimals	2.47	2.03*	2.46*	1.99	2.75*	2.36	2.96*	2.57
Interpret data in tables, charts, graphs	2.57	2.42*	2.55*	2.35	2.37*	2.11	2.74*	2.26
Write equations and functions	2.67	2.73	2.61*	2.20	2.76*	2.39	3.08*	2.54
Work together in small groups	1.71	1.86*	1.75	2.23*	1.88	2.01*	2.32	2.60*
Relate learning in math to daily lives	1.97	2.04	2.00	2.34*	2.36	2.39	2.48	2.90*
Explain answers	2.42	2.52*	2.57	2.92*	2.82*	2.64	3.31	3.27
Decide own procedures for solving complex problems	2.32*	2.18	2.29*	2.06	2.55*	2.24	2.63	2.80*
Work problems on own	3.49*	2.50	3.52*	2.08	2.90*	2.00	3.38*	2.28

* Denotes a significant difference ($p < .05$)

Table A.4. Multilevel model of Student Achievement (Data, Analysis, Probability)
Coefficient and *Effect Size* estimates.

	Japan		Netherlands		Singapore		USA	
	Coeff.	E.S.	Coeff.	E.S.	Coeff.	E.S.	Coeff.	E.S.
Classroom Mean, G00	571.00		556.71		576.91		527.41	
Tchr. Yrs of Exp., G01	-0.26	0.00	0.61	0.01	0.15	0.00	0.35	0.00
Tchr. Highest Ed., G02	13.62	0.19	12.74*	0.18	-0.77	0.01	0.61	0.01
Classroom Size, G03	0.60	0.01	5.75*	0.08	1.14	0.01	0.03	0.00
Tchr. <i>Reform</i> Practice, G04	7.28	0.10	-19.98	0.28	21.63*	0.11	7.87	0.10
<i>Reform</i> Perceptions, G05	4.34	0.06	17.31	0.24	-8.51	0.27	-7.15	0.09
Boy, B2	11.38*	0.16	13.82*	0.19	11.60*	0.15	2.97	0.04
Diff. Language at Home, B3	-3.12	0.04	-8.40*	0.12	-1.78	0.02	-12.79*	0.17
Books at Home, B4	7.65*	0.11	4.58*	0.06	1.61	0.02	8.30*	0.11
Lecture-Style Present., B4	4.14	0.06	1.59	0.02	-0.71	0.01	-0.61	0.01
Work on Problems Own, B5	22.72*	0.32	1.72	0.02	2.99*	0.04	4.41*	0.06
Use of Calculators, B6	-22.1*	0.32	1.95	0.03	-1.25	0.02	-2.46	0.03
SES (Home appliances), B5	8.83*	0.13	6.33	0.09	2.40*	0.03	-0.09	0.00
Student Age, B1	5.98	0.09	-6.67*	0.09	-3.32	0.04	-7.84*	0.10

* Significant ($p < .05$)

Table A.5. Multilevel model of Student Achievement (Fractions, Number Sense)
Coefficient and *Effect Size* estimates.

	Japan		Netherlands		Singapore		USA	
	Coeff.	E.S.	Coeff.	E.S.	Coeff.	E.S.	Coeff.	E.S.
Classroom Mean, G00	554.18		535.66		614.80		507.89	
Tchr. Yrs of Exp., G01	-0.44	0.01	0.73	0.01	0.16	0.00	0.48	0.01
Tchr. Highest Ed., G02	17.12	0.20	13.22*	0.19	0.06	0.00	0.60	0.01
Classroom Size, G03	1.04*	0.01	6.73*	0.09	1.04	0.01	-0.14	0.00
Tchr. <i>Reform</i> Practice, G04	9.13	0.10	-23.58	0.33	25.16*	0.13	11.25	0.14
<i>Reform</i> Perceptions, G05	7.38	0.08	18.80	0.26	-10.44	0.32	-9.92	0.12
Boy, B2	12.41*	0.14	15.85*	0.22	4.63*	0.06	10.07*	0.13
Diff. Language at Home, B3	-7.30	0.08	-9.49*	0.13	3.29*	0.04	-3.84	0.05
Books at Home, B4	10.79*	0.12	1.96	0.03	-0.42	0.01	7.59*	0.09
Lecture-Style Present., B4	-2.37	0.03	0.89	0.01	1.31	0.02	-2.14*	0.03
Work on Problems Own, B5	35.11*	0.40	2.28	0.03	5.89*	0.08	7.29*	0.09
Use of Calculators, B6	-22.4*	0.26	0.99	0.01	-1.96	0.03	-3.29*	0.04
SES (Home appliances), B5	12.34*	0.14	-2.84	0.04	2.13	0.03	1.47*	0.02
Student Age, B1	8.74*	0.10	-6.20*	0.09	-2.15	0.03	-8.46*	0.11

* Significant ($p < .05$)

Table A.6. Multilevel model of Student Achievement (Geometry)
Coefficient and *Effect Size* estimates.

	Japan		Netherlands		Singapore		USA	
	Coeff.	E.S.	Coeff.	E.S.	Coeff.	E.S.	Coeff.	E.S.
Classroom Mean, G00	584.74		508.34		577.07		472.59	
Tchr. Yrs of Exp., G01	-0.19	0.00	0.78	0.01	0.23	0.00	0.32	0.00
Tchr. Highest Ed., G02	13.08	0.17	13.25*	0.17	0.59	0.01	-0.28	0.00
Classroom Size, G03	0.68	0.01	6.82*	0.09	0.73	0.01	-0.24	0.00
Tchr. <i>Reform</i> Practice, G04	4.61	0.06	-27.76	0.37	27.22*	0.11	10.42	0.14
<i>Reform</i> Perceptions, G05	5.53	0.07	19.41	0.26	-9.16	0.34	-8.49	0.12
Boy, B2	3.21	0.04	8.69*	0.11	6.51*	0.08	7.46*	0.10
Diff. Language at Home, B3	-3.90	0.05	-7.34*	0.10	2.55*	0.03	-4.61*	0.06
Books at Home, B4	8.48*	0.11	2.78*	0.04	1.33	0.02	5.46*	0.07
Lecture-Style Present., B4	0.93	0.01	1.54	0.02	0.11	0.00	-0.88	0.01
Work on Problems Own, B5	33.13*	0.42	-1.52	0.02	5.31*	0.07	5.35*	0.07
Use of Calculators, B6	-22.4*	0.29	-1.03	0.01	-1.84	0.02	-0.32	0.00
SES (Home appliances), B5	11.46*	0.15	15.88*	0.21	0.25	0.00	1.19	0.02
Student Age, B1	7.77*	0.10	-7.80*	0.10	4.11	0.05	-5.82*	0.08

* Significant ($p < .05$)

Table A.7. Multilevel model of Student Achievement (Measurement)
Coefficient and *Effect Size* estimates.

	Japan		Netherlands		Singapore		USA	
	Coeff.	E.S.	Coeff.	E.S.	Coeff.	E.S.	Coeff.	E.S.
Classroom Mean, G00	556.93		545.73		607.79		495.54	
Tchr. Yrs of Exp., G01	-0.32	0.00	0.73	0.01	0.18	0.00	0.46	0.01
Tchr. Highest Ed., G02	14.84	0.21	13.02*	0.18	-0.49	0.01	1.88	0.02
Classroom Size, G03	0.73	0.01	6.51*	0.09	1.09	0.01	-0.19	0.00
Tchr. <i>Reform</i> Practice, G04	7.56	0.11	-21.51	0.29	25.39*	0.11	11.81	0.15
<i>Reform</i> Perceptions, G05	7.50	0.11	18.40	0.25	-8.65	0.32	-9.63	0.13
Boy, B2	5.69	0.08	18.97*	0.26	10.55*	0.13	13.55*	0.18
Diff. Language at Home, B3	-5.85	0.08	-9.08*	0.12	1.19	0.01	-5.20*	0.07
Books at Home, B4	7.73*	0.11	1.77	0.02	0.40	0.00	6.59*	0.09
Lecture-Style Present., B4	-1.87	0.03	-1.10	0.02	-0.77	0.01	-1.03	0.01
Work on Problems Own, B5	29.49*	0.42	6.55*	0.09	6.14*	0.08	4.34*	0.06
Use of Calculators, B6	-18.0*	0.25	-0.79	0.01	-2.94*	0.04	-1.80	0.02
SES (Home appliances), B5	10.10*	0.14	-0.69	0.01	0.84	0.01	-0.08	0.00
Student Age, B1	4.85	0.07	-6.69*	0.09	-1.44	0.02	-6.22*	0.08

* Significant ($p < .05$)

References

- Battista, M. T. (1994). Teacher beliefs and the reform movement in mathematics education. *Phi Delta Kappan*, 75(6), 462-470.
- Battista, M. T. (2001). Research and reform in mathematics education. In T. Loveless (Ed.), *The great curriculum debate: How should we teach reading and math?* (pp.42-84). Washington, DC: Brookings.
- Borko, H., Stecher, B. M., Alonzo, A., Moncure, S., & McClam, S. (2005). Artifact packages for measuring instructional practice: A pilot study. *Educational Assessment*, 10(2), 73-104.
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Research Council.
- Burrill, G. (2001). Mathematics education: The future and the past create a context for today's issues. In T. Loveless (Ed.), *The great curriculum debate: How should we teach reading and math?* (pp.25-41). Washington, DC: Brookings.
- Burstein, L., Chen, Z., & Kim, K. S. (1989). *Analyses of procedures for assessing content coverage & its effects on instruction* (CSE Technical Report 305). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
- Burstein, L., McDonnell, L. M., Van Winkle, J., Ormseth, T., Mirocha, J., & Guitton, G. (1995). *Validating national curriculum indicators*. Santa Monica, CA: RAND.
- Cavanaugh, S. (2006, September 12). NCTM issues new guidelines to help schools home in on the essentials of math. *Education Week* (26), available at www.edweek.org/ew/articles/2006/09/12/03nctm_web.h26.html.
- Charter School Achievement Consensus Panel (2006). *Key issues in studying charter schools and achievement: A review and suggestions for national guidelines*. (NCSR White Paper Series, No. 2).
- Cobb, P., Wood, T., Yackel, E., Nicholls, J., Wheatley, G., Trigatti, B., & Perlwitz, M. (1991). Assessment of a problem-centered second-grade mathematics project. *Journal for Research in Mathematics Education*, 22(1), 3-29.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cohen, D. K., & Hill, H. C. (2000). Instructional policy and classroom performance: The mathematics reform in California. *Teachers College Record*, 102(2), 294-343.
- Desimone, L. M., Smith, T. M., Ueno, K. & Baker, D.P. (2005). The distribution of teaching quality in mathematics: Assessing barriers to the reform of United States mathematics instruction from an international perspective. *American Educational Research Journal*, 42, 501-535.

- Ercikan, K., and Koh, K. (2005). Examining the Construct Comparability of the English and French Versions of TIMSS, *International Journal of Testing*, 5(1), 23-35.
- Floden, R. E. (2002). The measurement of opportunity to learn. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp.229-266). Washington, DC: National Academy Press.
- Gamoran, A. (1991). Schooling and achievement: Additive versus interactive models. In S.W. Raudenbush and J.D. Willms (Eds.), *Schools, Classrooms, and Pupils: International studies of schooling from a multilevel perspective*. San Diego: Academic Press.
- Gamoran, A., Porter, A. C., Smithson, J., & White, P. A., (1997). Upgrading high school math instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis*, 19, 325-338.
- Greeno, J. G., Collins, A. M., & Resnick, L. (1996). Cognition and learning. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp.15-46). New York: MacMillan
- Grisay, A., de Jong, J., Gebhart, E., Berezner, A., & Halleux-Monseur, B. (2006). *Translation equivalence across PISA countries*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.
- Haertel, E. H. (2003). Differential prediction and opportunity to learn. *Paper presented at the 2003 Meeting of the American Educational Research Association*, Chicago, IL.
- Hamilton, L.S. (1998). Gender differences on high school science achievement tests: Do format and content matter? *Educational Evaluation and Policy Analysis*, 20, 179-195.
- Hamilton, L.S. (2003). Assessment as a policy tool. *Review of Research in Education*, 27, 25-68.
- Herman, J. L. & Klein, D. (1997). *Assessing opportunity to learn: A California example* (CSE Technical Report 453). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
- Herman, J. L., Klein, D. C. D., & Abedi, J. (2000). Assessing students' opportunity to learn: Teacher and student perspectives. *Educational Measurement: Issues and Practice*, 19(4), 16-24.
- Herman, J. L. & Abedi, J. (2004). *Issues in assessing English language learners' opportunity to learn mathematics* (CSE Technical Report 663). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
- Hiebert, J. (1999). Relationship between research and the NCTM Standards. *Journal for Research in Mathematics Education*, 30(1), 3-19.

Hiebert, J., & Carpenter, T. P. (1992). Learning and teaching with understanding. In D. A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching* (pp.65-97). Reston, VA: NCTM/Macmillan.

Hiebert, J., & Stigler, J. W. (2000). A proposal for improving classroom teaching: Lessons from the TIMSS video study. *Elementary School Journal*, 101, 3-20.

Hiebert, J., Gallimore, R., Garnier, H., Givvin, K. B., Hollingsworth, H., Jacobs, J. Chiu, A., Wearne, D., Smith, M., Kersting, N., Manaster, A., Tseng, E., Etterbeek, W., Manaster, C., Gonzales, P., & Stigler, J. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 video study* (NCES 2003-013). Washington, DC: U. S. Department of Education, National Center for Education Statistics.

Hill, H. C. (2005). Content across communities: Validating measures of elementary mathematics instruction. *Educational Policy*, 19(3), 447-475.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, 42(2), 371-406.

Jacobs, J. K., Hiebert, J., Givvin, K. B., Hollingsworth, H., Garnier, H., & Wearne, D. (2006). Does eighth-grade mathematics teaching in the United States align with the NCBM Standards? Results from the TIMSS 1995 and 1999 video studies. *Journal for Research in Mathematics Education*, 37, 5-32.

Klein, S.P., Hamilton, L.S., McCaffrey, D.F., Stecher, B.M., Robyn, A., & Burroughs, D. (2000). *Teaching practices and student achievement: Report of first-year results from the Mosaic study of Systemic Initiatives in mathematics and science* (MR-1233-EDU). Santa Monica, CA: RAND.

Koretz, D., McCaffrey, D., & Sullivan, T. (2001, September 14). Predicting variations in mathematics performance in four countries using TIMSS. *Education Policy Analysis Archives*, 9(34). Retrieved 10/10/06 from <http://epaa.asu.edu/epaa/v9n34/>.

Kupermintz, H., Ennis, M.M., Hamilton, L.S., Talbert, J.E., & Snow, R.E. (1995). Enhancing the validity and usefulness of large-scale educational assessments: I. NELS:88 mathematics achievement. *American Educational Research Journal*, 32, 525-554.

Le, V., Stecher, B.M., Lockwood, J.R., Hamilton, L.S., Robyn, A., Williams, V., Ryan, G., Kerr, K., Martinez, F., & Klein, S. (2006). Improving mathematics and science education: A longitudinal investigation of the relationship between reform-oriented instruction and student achievement. Santa Monica, CA: RAND.

Lewin, T. (September 13, 2006). Report urges changes in the teaching of math in U.S. schools. *New York Times*. Retrieved 9/14/06 from <http://www.nytimes.com/2006/09/13/education/13math.html?ref=education>.

Linn, M. C., Kessel, C., Lee, K., Levenson, J., Spitulnik, M., and Slotta, J. D. (2000). *Teaching and learning K-8 mathematics and science through inquiry: Program reviews and recommendations*. Unpublished report commissioned by the North Central Regional Educational Laboratory. Available at <http://www.ncrel.org/engage/resource/techno/k8.htm>.

Linn, R. L. (2002). The measurement of student achievement in international studies. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp.27-57). Washington, DC: National Academy Press.

Lockwood, J.R., McCaffrey, D.F., Hamilton, L.S., Stecher, B., Le, V., and Martinez, F. (in press) The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures. In print in *Journal of Educational Measurement*.

Loveless, T. (2001, Ed.). *The great curriculum debate: How should we teach reading and math?* Washington, DC: Brookings.

Loveless, T. (2003). Trends in math: The importance of basic skills. *The Brookings Review*, 21(4), 41-43.

Martin, M O. (2005) TIMSS 2003 User Guide for the International Database. Chestnut Hill, MA: Boston College

Martinez-Fernandez, J.F. (2005). A multilevel study of the effects of opportunity to learn (OTL) on student reading achievement: Issues of measurement, equity, and validity (Doctoral Dissertation, University of California, Los Angeles). *Dissertation Abstracts International*, DAI-A 65/11, 3155006.

Matsumura, L.C., Garnier, H., Pascal, J., & Valdés, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and student achievement. *Educational Assessment*, 8(3), 207-229.

Mayer, D. P. (1998). Do new teaching standards undermine performance on old tests? *Educational Evaluation and Policy Analysis*, 20, 53-73.

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis*, 21, 29-45.

McCaffrey, D. F., Hamilton, L. S., Stecher, B. M., Klein, S. P., Bugliari, D., & Robyn, A. (2001). Interactions among instructional practices, curriculum, and student achievement: The case of standards-based high school mathematics. *Journal for Research in Mathematics Education*, 32(5), 493-517.

Muthén, B., Huang, L. C., Jo, B., Khoo, S. T., Goff, G., Novak, J., & Shih, J. (1995). Opportunity-to-learn effects on achievement: analytical aspects. *Educational Evaluation and Policy Analysis*, 17(3), 371-403.

National Council of Teachers of Mathematics. (1989). *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics (2000). *Principles and Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence*. Reston, VA: Author.

National Research Council (1999). *How people learn: Brain, mind, experience, and school*. Committee on developments in the science of learning, Commission on Behavioral and Social Sciences and Education (J. Bransford, A. Brown, & R. Cocking, Eds.). Washington, DC: National Academy Press.

Opdenakker, M. C. & Van Damme, J. (2000). The importance of identifying levels in multilevel analysis: an illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. *School Effectiveness and School Improvement*, 11(1), 103-130.

Porter, A. C. (1995). The uses and misuses of opportunity to learn standards. *Educational Researcher*, 24(4), 21-27.

Porter, A. C. (1998). The effects of upgrading policies on high school mathematics and science. In D. Ravitch (Ed.), *Brookings Papers on Education Policy* (pp.123-164). Washington, DC: Brookings.

Porter, A. C., Kirst, M. W., Osthoff, E., Smithson, J. L., and Schneider, S. A. (1994). *Reform of high school mathematics and science and opportunity to learn*. New Brunswick, NJ: Rutgers University, Consortium for Policy Research in Education.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and data analysis methods*. 2nd ed. Thousand Oaks: Sage.

Raudenbush, S., Bryk, A., Cheong, Y.F., & Congdon, R. (2004). *HLM 6: Hierarchical Linear and Nonlinear Modeling*. Lincolnwood, IL: Scientific Software International.

Ravitz, J. L., Becker, H. J., & Wong, Y. T. (2000). *Constructivist-compatible beliefs and practices among U.S. teachers*. Irvine, CA: Center for Research on Information Technology and Organizations.

Saxe, G. B., Gearhart, M., & Seltzer, M. (1999). Relations between classroom practices and student learning in the domain of fractions. *Cognition and Instruction*, 17(1): 1-24.

Schenker, N., and Gentleman, J.F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55, 182-186.

- Schulz, W. (2006). *Testing parameter invariance for questionnaire indices using confirmatory factor analysis and item response theory*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.
- Schmidt, W. E., McKnight, C. C., & Raizen, S. A. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Dordrecht, Netherlands: Kluwer.
- Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 1066-1101). Washington, DC: American Educational Research Association.
- Shouse, R. (2001). The impact of traditional and reform-style practices on student mathematics achievement. In T. Loveless (Ed.), *The great curriculum debate: How should we teach reading and math?* (pp.108-133). Washington, DC: Brookings.
- Smith, J., Lee, V., & Newmann, F. (2001). *Instruction and achievement in Chicago elementary schools: Improving Chicago's schools*. Chicago, IL: Consortium on Chicago School Research.
- Smithson, J. L., & Porter, A., C. (1994). *Measuring classroom practice: Lessons learned from the efforts to describe the enacted curriculum—The Reform Up Close study*. Madison, WI: Consortium for Policy Research in Education, University of Wisconsin.
- Spillane, J. P., & Zeuli, J. S. (1999). Reform and teaching: Exploring patterns of practice in the context of national and state mathematics reforms. *Educational Evaluation and Policy Analysis*, 21, 1-27.
- Stevenson, H. W., Lee, S., & Stigler, J. W. (1986). Mathematics achievement of Chinese, Japanese, and American children. *Science*, 231(4739), 693-699.
- Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS videotape classroom study: Methods and findings from an exploratory research project on eight-grade mathematics instruction in Germany, Japan, and the United States* (NCES 1999-074). Washington, DC: U. S. Department of Education, National Center for Education Statistics.
- Thompson, D., & Senk, S. (2001). The effects of curriculum on achievement in second year algebra: The example of the University of Chicago mathematics project. *Journal for Research in Mathematics Education*, 32(1), 58-84.
- Walker, M. (2006). *The choice of Likert or dichotomous items to measure attitudes across culturally distinct countries in international comparative educational research*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.
- Wenglinsky, H. (2002). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives*, 10(12). Available at <http://epaa.asu.edu/epaa/v10n12/>.

Willms, J.D; Smith, T (2005). A manual for conducting analyses with data from TIMSS and PISA. Montreal, Canada: United Nations Educational, Scientific, and Cultural Organization; Institute for Statistics (UIS). Available at http://www.unb.ca/crisp/pdf/Manual_TIMSS_PISA2005_0503.pdf

Wood, T., & Sellers, P. (1997). Deepening the analysis: Longitudinal assessment of a problem-centered mathematics program. *Journal for Research in Mathematics Education*, 28(2), 163-186.