



TIMSS Primary and Middle School Data: Some Technical Concerns

by Jianjun Wang

Student scores from the *Third International Mathematics and Science Study* (TIMSS) have been employed as a benchmark to measure school effectiveness in many districts across the United States. Bracey (2000) discussed the TIMSS results at the high school level. This article focuses on an in-depth examination of student performance from the TIMSS primary and middle school studies. The data analyses show substantial uncertainty of the country rankings from the released TIMSS reports. Other technical issues hinge on the instrument construction, curricular inequivalency, and statistical outliers. These analyses raise concerns about the use of the TIMSS benchmark for school reforms in the U.S. Taken together, these concerns have important implications for the interpretation and use of the TIMSS results.

The Third International Mathematics and Science Study (TIMSS) is the largest and most ambitious international assessment of primary, middle, and high school student performance. In the 1994/95 school year, TIMSS tested over half a million students in more than 40 nations. The results showed that in the U.S. eighth graders performed slightly above the international average in science and slightly below the average in mathematics (Beaton et al., 1996a, 1996b). Using results from primary and middle schools, Schmidt and McKnight (1998) reported “a decline in the relative standing of U.S. students from fourth to eighth grade in both mathematics and science, as compared to those in other countries” (p. 1830).

The middle school findings were confirmed by a repeat of the TIMSS project (TIMSS-R) four years later in 1999 (Martin et al., 2000; Mullis, et al., 2000). The recent release of the TIMSS-R reports has rekindled public interest in TIMSS because the two projects together provide an opportunity for a trend analysis between the fourth and eighth grades during the period of 1995–1999 (Martin, Gregory, & Stemler, 2000).

In the May 2000 issue of *Educational Researcher*, Gerald Bracey examined the TIMSS findings at the high school level. He noted that

TIMSS in general is a rich source of information in comparison to earlier international comparisons. The videotapes, curriculum analyses, case studies, and other forthcoming aspects of TIMSS provide much food for thought and many pointers for the reform of mathematics education. The test data, at all grades, much less

so. The test data from the Final Year component least of all. (Bracey, 2000, p. 9)

The purpose of this article is to extend the discussion of TIMSS findings to the primary and middle school levels.

In the existing literature, the TIMSS project has been valued for its rich comparative information about educational systems, curriculums, school characteristics, and instructional practices. The strength of TIMSS also hinges on its attempt to link student performance scores with these components of school improvement (Schmidt & McKnight, 1998). On the other hand, Rotberg (1998) has questioned the validity of the score ranking “because countries differ substantially in such factors as student selectivity, curriculum emphases, and the proportion of low-income students in the test-taking population” (p. 1030). Evidence has been adduced to justify these concerns at the high school level (Bracey, 2000; Rotberg, 1998).

At the primary and middle school levels scrutiny of the TIMSS findings was mainly confined to the interpretation of TIMSS results. One researcher argued that, because TIMSS was not a controlled scientific study and did not measure the effectiveness of one teaching method versus another (<http://indigo.col-ed.org/mine/timssran.htm>), the results cannot support specific reforms at a local school. Hu (2000) further noted that the TIMSS reports did not examine the results for different ethnic groups. He argued:

America actually compares favorably with other nations with Caucasians, especially considering that 25% of the population is of under-performing African and Latino descent. The top nations are all East Asian. This study does not break down Americans by race, if they did, Asian Americans would likely score as high as Asians in their home countries, and Whites would rank near top of the European nations. (Hu, 2000, p. 8)

These critiques largely addressed design features that were missing in TIMSS. LeTendre and Baker (1998) categorized these kinds of attacks as “an ideological reaction against international comparisons, rather than a serious examination of the TIMSS methods” (p. 46).

Instead of continuing this line of inquiry, this article switches to an in-depth analysis of the TIMSS test scores. In my examination of the released data, I was able to verify the results of mathematics and science achievement in the existing TIMSS reports. However, I found that conflicting country ranks can be derived from the TIMSS database. In addition, I found several other technical problems that can alter the comparative results, undercutting the reliability of the TIMSS benchmarking. Al-

though none of these issues is fatal in its own right, taken together, they raise important methodological concerns about TIMSS. I share my observations in this article so that the U.S. policy makers and the general public can make more mindful interpretations of these TIMSS findings.

The Country Rankings Released in TIMSS Reports Are Not the Only Ones Supported by the TIMSS Database

Because students tested in TIMSS had been enrolled in school for several years, measuring the cumulative achievement in a short testing period represented a considerable challenge. As with any large-scale assessment, a short test may not sufficiently cover what students have learned so far. As it usually demands more time and effort from the test takers, a lengthy test can cause a low response rate.

To resolve the conflict between time allocation and content coverage, researchers at the Educational Testing Service proposed a balanced incomplete block (BIB) design that assigned subsets of a lengthy test to a random sample of students. Whereas the entire test instrument represented sufficient content coverage, assigning a subset of questions to each student substantially reduced the burden on test takers. On the basis of the subtest scores students' overall performance on the larger test can be projected through data imputations (see Mislevy, 1991; Mislevy, Johnson, & Muraki, 1992). The imputed total scores are plausible values drawn from an estimated ability distribution of students with similar item response patterns and background characteristics (Gonzalez & Smith, 1997).

This imputation method was first introduced in the mid 1980s for the National Assessment of Educational Progress (NAEP). When five imputed plausible scores were calculated in NAEP, no single score was used exclusively in the nation's report card (Allen, Carlson, & Zelenak, 1999). Instead, an average score was computed from all five imputed plausible scores to represent student academic performance.

Strictly speaking, "no statistical methods will fully compensate for missing units and data" (Madow, Nisselson, & Oklin, 1983, p. 6). Therefore, statisticians generally recommend the method of multiple imputations to triangulate the results (e.g., Little & Rubin, 1987; Rubin, 1987; Schafer, 2000; Wang, Sedransk, & Jinn, 1992). "Because of the error involved in the imputation process, TIMSS produced not one but five imputed values for each student in mathematics and science" (Gonzalez & Smith, 1997, ch. 6, p. 3). The method of computing multi-

ple plausible scores was "based on Rubin's (1987) 'multiple imputation' procedures for missing data" (Mislevy, 1991, p. 177).

Sande (1982) cautions, "If one wants to explore relationships between variables, the use of imputed data could be prejudicial, not to mention misleading" (p. 147). However, when TIMSS researchers checked the inter-correlations between the five plausible scores, they decided that "the imputation error could be ignored" (Gonzalez & Smith, 1997, ch. 6, p. 3). Accordingly, only the first imputed plausible score was used in TIMSS reports (e.g., Beaton et al., 1996a, 1996b). Arguing that "one set of the imputed plausible scores can be considered as good as another" (Gonzalez & Smith, 1997, ch. 6, p. 3), TIMSS researchers did not justify any superiority of the first plausible score over the other four scores. Nor did they explain the reason for not calculating the average plausible scores from the multiple imputations.

The impact of the imputation error can be illustrated through a comparison of the five imputed plausible scores in the released TIMSS database. Inspection of the data suggests that the score difference between some countries may fluctuate as much as seven points.¹ In many international comparisons, the score fluctuation is twice as much as the standard error of the TIMSS average score in each nation.² Therefore, the choice of other imputed plausible scores can result in alternative findings different from those in the released TIMSS reports (e.g., Frase et al., 1997; Peak, 1996). The percentage of countries changing at least one position in the TIMSS rank ranges from 17% to 59% in the primary and middle school surveys (see Table 1).

The imputation error directly affects assessment of U.S. education in a cross-national context. On the basis of the results from the eighth grade science test, Hungary scored significantly higher than did the U.S., whereas an insignificant difference was found between the U.S. and England (Beaton et al., 1996a). The use of different plausible scores can switch the ranks between England and Hungary and thus introduce inconsistency in the TIMSS reporting. Still, no effort has been made by TIMSS researchers to produce empirical findings that fit the majority of the plausible scores.

Given the existence of score variations, "nations have been grouped into broad bands according to whether their performance is higher than, not significantly different from, or lower than the U.S." (Peak, 1996, p. 19). Boundaries of these categories were delineated with cut-off lines from statistical testing. Inspection of the eighth grade science scores suggests that Spain's average performance was higher than that of Scotland in four out of the five plausible scores. But the plausible scores employed in a

Table 1
Percent of Countries Changing At Least One Position in the TIMSS Rank List

Population	Level	Mathematics	Science	Total number of countries
Primary School	Lower Grade	17%	42%	24
	Upper Grade	38%	50%	26
Middle School	Lower Grade	38%	41%	39
	Upper Grade	44%	59%	41

Note: A country that can be switched either up or down is counted as one country in the percentage computation.

TIMSS report placed Spain and Scotland at the same level of science performance (Beaton et al., 1996a). Moreover, a cut-off line has been set between Spain and Scotland that dropped Spain into a group of nations scoring significantly below the United States (Peak, 1996). This statistical artifact may have led the general public to believe that there was an indisputable difference in the eighth grade achievement between Spain and Scotland.

It is clear that the selection of plausible scores was somewhat arbitrary in the TIMSS reporting. Therefore, the exact number of countries outperforming or under-performing the U.S. cannot be conclusively determined from the arbitrary use of one of five different plausible scores. This discrepancy highlights an inconsistency problem within the TIMSS reports.

The Format of TIMSS Test Items Is Not Consistent With the Goals of Some Education Reforms

In the United States, educators and policy makers frequently use the TIMSS and TIMSS-R results to make judgments about the condition of mathematics and science education. Included in the use of these results is an implicit assessment of the effectiveness of ambitious national reform initiatives that have begun to take hold in U.S. schools in the last decade. Many of these initiatives stress higher order thinking, complex scientific and mathematical reasoning, and hands-on experience (American Association for the Advancement of Science, 1990, 1993; National Council of Teachers of Mathematics, 1989, 1998). Similarly, reforms in Japan, which have commanded more attention in that country than TIMSS rankings, call for problem solving and first-hand, original investigations (Atkin & Black, 1997).

However, the TIMSS test “measures mostly lower learning outcomes by means of predominantly multiple choice format” (Lange, 1997, p. 3). The distribution of TIMSS items was 429 multiple-choice, 43 short-response, and 29 extended-response items (Lange, 1997). Because students were tested on subsets of questions from this item pool, the TIMSS scores cannot reflect the kinds of outcomes that are emphasized in reform initiatives that focus on higher order thinking and hands-on experiences.

It should be acknowledged that the introduction of any free-response items in TIMSS was a technical improvement over the previous international studies (Martin & Kelly, 1997). The free-response items not only asked for final correct answers, but also solicited individual explanations for those answers. For instance, two of the TIMSS items read

1. A glass of water with ice cubes in it has a mass of 300 grams. What will the mass be immediately after the ice has melted? Explain your answer. (Lie, Taylor, & Harmon, 1996, ch. 7, p. 11)
2. The water level in a small aquarium reaches up to a mark A. After a large ice cube is dropped into the water, the cube floats and the water level raises to a new mark B. What will happen to the water level as the ice melts? Explain your reasoning. (Item # G11, <http://www.csteep.bc.edu/timss1/Items.html>)

Unfortunately, the grading rubrics maintained a clear feature of multiple-choice items, i.e., they accepted only one correct answer. Regardless of the free responses provided by students, they received no credit if their answers indicated that they believed there would be a change in mass or water level.

Such rubrics are also sources of potential scientific errors in the TIMSS scoring process. Awarding credits for an answer that indicated there would be no change in mass or water level may have penalized students who considered the ongoing effect of evaporation, particularly under a temperature close to the freezing point. Because neither item specified the experimental temperature, it was unclear whether it took one minute or several hours for the ice to melt. However, the careful examination of experimental conditions represents a fundamental component of scientific literacy highlighted in several key documents of science education, such as *Science for All Americans* (American Association for the Advancement of Science, 1990) and *Benchmarks for Science Literacy* (American Association for the Advancement of Science, 1993). The lack of alignment between these professional documents and the TIMSS benchmark raises doubts about the usefulness of the TIMSS results in measuring the effectiveness of various school reform initiatives in the United States, where those reforms have been implemented.

The TIMSS Test Booklets Have Discrepant Structures

The TIMSS test items were grouped into 26 different clusters labeled A through Z. Cluster A was the core cluster that was administered to all students. Clusters B through H were focus clusters that “appeared in at least three booklets” (Gonzalez & Smith, 1997, ch. 2, p. 4). The breadth clusters were labeled I through M for mathematics and N through R for science, and appeared in only one booklet. In comparison, the breadth clusters appeared less frequently than the focus clusters in most test booklets, and were designed to “contribute to the breadth of the tests” (Adams & Gonzalez, 1996, ch. 3, p. 6). Finally, the free-response clusters were labeled S through V for mathematics and W through Z for science, and were each assigned to two booklets (Gonzalez & Smith, 1997).

Because “the 26 clusters were assembled into eight booklets” (Gonzalez & Smith, 1997, ch. 2, p. 4), placement of these clusters deserves discussion. Gonzalez and Smith (1997) reported that a rotation design was used “to assign clusters B through H to booklets 1 through 7” (ch. 2, p. 6). Booklet eight was excluded from the rotation design.

Table 2 shows the structure difference between booklet eight and booklets one through seven. Booklet eight has more emphasis on the breadth clusters. Thus, this test booklet is likely to benefit those students who have studied a curriculum that is “a mile wide and an inch deep.” The other seven booklets include more focus clusters. Because of the difference in content coverage, a high achievement score from booklet eight may not provide a good prediction of student performance on the other test booklets. This structural discrepancy can introduce systematic errors during an indiscretionary imputation of test scores across the test booklets.

Because of Grade-Level Differences and Content-Differences Among Countries, the TIMSS Tests Might Not Align With What Students Have Learned

In most countries the TIMSS primary and middle school tests were administered to students in two adjacent grades. In the United States third and fourth graders were chosen to take the test for the primary school population. Although TIMSS re-

Table 2
Difference in the Number of Item Clusters Between
Booklet Eight and Booklets One through Seven

Type of Cluster	Booklet Eight	Booklets one through seven
Core	1	1
Focus	1	3
Breadth	3	1
Free Response	(2)	2

Note: () indicates that the free-response clusters were a part of the test booklet for primary schools only and not of that for middle schools.

searchers claimed to have an appropriate test for both grades, the grade gap represented 25% of students' schooling experience. Similarly, the period between grades seven and eight made up 12.5% of students' school lives. Given the difference in learning experiences from one grade to the next, it is unclear whether any single test can measure satisfactorily what students have learned in both the lower and upper grades.

Moreover, curriculum mismatch across countries can have a direct impact on the international comparison of student achievement. For instance, Silver (1998) reported, "Compared to many other countries, the content taught at grade eight in the United States is similar to the content taught at grade seven elsewhere" (p. 2). Because eighth graders in the U.S. did not have the opportunity to learn the eighth grade content taught in other nations, perhaps we should have used U.S. ninth graders to compare with the eighth graders in these nations. Similar grade adjustments were made in several English-speaking nations or regions, such as Australia, England, and Scotland. Although the grade gap seemed undesirable for U.S. schools, it was unfair to give students the on-grade test before a curriculum match had been made to cover the equivalent content in other countries. If the comparisons were made among students with a similar content exposure the TIMSS data would show that U.S. eighth graders outperformed Japanese seventh graders in science and scored above the seventh grade international average in mathematics (Beaton et al., 1996a, 1996b). These findings are essentially parallel to TIMSS results at the fourth grade level (Martin, Mullis et al., 1997; Mullis et al., 1997) but differ dramatically from the middle school findings in most TIMSS reports (e.g., Beaton et al., 1996a, 1996b; Peak, 1996). TIMSS researchers reported a drop of U.S. student performance from fourth to eighth grade in both mathematics and science (e.g., Frase et al., 1997; Peak, 1996).

Several Problematic Age Outliers in the TIMSS Database Are Not Adequately Explained

Martin and Kelly (1997) reported that

The question of whether student populations should be defined by chronological age or grade level in school is one that faces all comparative surveys of student achievement. TIMSS addressed this issue by defining (for Populations 1 and 2) the target population as the pair of adjacent grades that contains the largest proportion of a particular age group (9-year-olds for population 1, and 13-year-olds for Population 2). (p. 8)

Despite this agreement to use 13-year-olds to define the TIMSS middle school population, Colombia, Germany, Romania, and Slovenia opted to test seventh and eighth grade students "somewhat older than in the other countries" (Beaton, et al. 1996a, p. A-18).

In reality, students tested in those countries were much older. One seventh grader in Columbia, for instance, was 49.9 years old at the time TIMSS was administered. In addition, the TIMSS international data showed that 39% of participating countries had seventh or eighth graders over age 18. The U.S. survey did not include such adult students, but the U.S. data did include one 10-year-old eighth grade student and several six-year-old fourth graders. Because student age is a contributor to cognitive development, these outliers are potentially important in interpreting comparative studies such as TIMSS.

Summary

TIMSS does provide some rich information in terms of comparative education. TIMSS researchers adopted the latest plausible value methodology to shorten the testing time and extend the content coverage. They developed a testing framework on the basis of consensus and compromise among international educators, even including a small percentage of free-response items to solicit student explanations. And the entire TIMSS project was monitored by several quality control measures (Martin & Mullis, 1996). Still, without checking details of the released data, one may "have been relatively accepting of the TIMSS data from grades four and eight" (Bracey, 1998, p. 94).

As I have argued in this paper, the test score component of TIMSS has some technical problems that should be considered when interpreting TIMSS results. First, by including all five plausible values in the released database, TIMSS researchers have shown their academic integrity. Nonetheless, under an assumption that all five plausible values were equally representative of student performance, TIMSS researchers selected only one of those values. As I have shown, however, choosing other values can lead to the reordering of a large proportion of the nations in the existing fourth and eighth grade reports (see Table 1). Second, to justify a proper curriculum match, the U.S. could have followed Australia, England, and Scotland to raise its grade levels for the TIMSS testing. If the U.S. had done so the results could have been substantially different from the ones in the released TIMSS reports. Third, the few free-response items in the TIMSS assessment design were overshadowed by the predominance of multiple-choice questions. Finally, the quality control measures did not resolve booklet inequivalency, curriculum mismatch, and age outliers.

These issues are interconnected, and collectively can lead to a problematic interpretation of the TIMSS benchmark for U.S. education. In particular, it is worth re-iterating that one cannot accept the country rankings presented in the TIMSS reports at face value. Similarly, because the format of the test items does not sufficiently align with the goals of U.S. reforms the TIMSS benchmark is not an accurate measure of the success of these initiatives (e.g., American Association for the Advancement of Science, 1990, 1993; National Council of Teachers of Mathematics, 1989, 1998; National Research Council, 1996). For these reasons, the TIMSS findings should be scrutinized carefully before being made the basis for decision-making about policy, practice, or research in science and mathematics education.

NOTES

Bruce Smith, John R. Staver, David H. Ost, and James Laughner read this manuscript at the author's request and made insightful comments. The author also benefited from his fellowship experience at the National Center for Education Statistics. In addition, the author wishes to thank the *ER* Feature Editors and reviewers for their constructive suggestions. The author takes full responsibility for research findings presented in this article.

¹ For instance, in the eighth grade mathematics test, the mean score difference between Israel and Iran fluctuates as much as seven points. The same pattern exists in a comparison of the eighth grade science scores between Czech Republic and Iceland.

² In the following international science comparisons, the score fluctuation is twice as much as the standard error of the TIMSS average score in each nation: Canada vs. Portugal (the seventh grade test), Cyprus vs. Canada (the seventh grade test), France vs. Korea (the eighth grade test), Korea vs. Cyprus (the seventh grade test), and Korea vs. Slovenia (the seventh and eighth grade tests).

REFERENCES

- Adams, R. J., & Gonzalez, E. J. (1996). The TIMSS test design. In M. Martin & D. Kelly (Eds.), *Third international mathematics and science study: Technical report*. Chestnut Hill, MA: Boston College.
- Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The NAEP 1996 technical report*. Washington, DC: U.S. Department of Education.
- American Association for the Advancement of Science (1990). *Science for all Americans*. New York: Oxford University Press.
- American Association for the Advancement of Science (1993). *Benchmarks for science literacy*. New York: Oxford University Press.
- Atkin, J. M., & Black, P. (1997). *Policy perils of international comparisons: The TIMSS case*. [On line] Available: <http://www.enc.org/topics/timss/additional/documents/0,1946,CDS-000162-cd162,00.shtm> (October 11, 2000).
- Beaton, A., Martin, M., Mullis, I., Gonzalez, E., Smith, T., & Kelly, D. (1996a). *Science achievement in the middle school years*. Chestnut Hill, MA: Boston College.
- Beaton, A., Martin, M., Mullis, I., Gonzalez, E., Smith, T., & Kelly, D. (1996b). *Mathematics achievement in the middle school years*. Chestnut Hill, MA: Boston College.
- Bracey, G. (1998). The author responds. *Phi Delta Kappan*, 80(1), 94.
- Bracey, G. (2000). The TIMSS final year study and report: A critique. *Educational Researcher*, 29(4), 4-10.
- Frase, M., Peak, L., Jakworth, P., Schmidt, W., Martin, L., Suter, L., Orland, M., Takahira, S., Owen, E., & Williams, T. (1997). *Pursuing Excellence: A study of U.S. fourth-grade mathematics and science achievement in international context*. Washington, DC: U.S. Department of Education.
- Gonzalez, E. J., & Smith, T. A. (1997). *Users guide for the TIMSS international database*. Chestnut Hill, MA: TIMSS International Study Center.
- Hu, A. (2000, December). TIMSS: Arthur Hu's index. [On line] Available: <http://www.leconsulting.com/arthurhu/index/timss.htm>. (December 30, 2000).
- Lange, J. D. (1997). *Looking through the TIMSS mirror from a teaching angle*. [On line] Available: <http://www.enc.org/topics/timss/additional/documents/0,1341,CDS-000158-cd158,00.shtm> (October 11, 2000).
- LeTendre, G., & Baker, D. (1998). Why blame adolescents or middle schools for poor U.S. science performance? *Education Week*, XVII(40), 46, 93.
- Lie, S., Taylor, A., & Harmon, A. (1996). Scoring techniques and criteria. In M. Martin & D. Kelly (Eds.), *Third international mathematics and science study: Technical report*. Chestnut Hill, MA: Boston College.
- Little, R., & Rubin, D. (1987). *Statistical analysis with missing data*. NY: John Wiley & Sons.
- Madow, W., Nisselson, H., & Olkin, I. (1983). *Incomplete data in sample surveys: Report and case studies*. NY: Academic Press.
- Martin, M. O., Gregory, K. D., & Stemler, S. E. (2000). *TIMSS 1999: Technical report*. Chestnut Hill, MA: TIMSS International Study Center.
- Martin, M. O., & Kelly, D. L. (1997). *Technical report volume II: Implementation and analysis*. Chestnut Hill, MA: TIMSS International Study Center.
- Martin, M. O., & Mullis, I. V. S. (1996). *Quality assurance in data collection*. Chestnut Hill, MA: TIMSS International Study Center.
- Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Gregory, K. D., Smith, T. A., Chrostowski, S. J., Garden, R. A., & O'Connor, K. M. (2000). *TIMSS 1999: International science report*. Chestnut Hill, MA: TIMSS International Study Center.
- Martin, M., Mullis, I., Beaton, A., Gonzalez, E., Kelly, D., & Smith, T. (1997). *Science achievement in the primary school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Boston College.
- Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56, 177-196.
- Mislevy, R., Johnson, E., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, 17, 131-154.
- Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Gregory, K. D., Garden, R. A., O'Connor, K. M., Chrostowski, S. J., Smith, T. A. (2000). *TIMSS 1999: International mathematics report*. Chestnut Hill, MA: TIMSS International Study Center.
- Mullis, I., Martin, M., Beaton, A., Gonzalez, E., Kelly, D., & Smith, T. (1997). Mathematics achievement in the primary school years: IEA's Third International *Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Boston College.
- National Council of Teachers of Mathematics (1989). *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA: The Author.
- National Council of Teachers of Mathematics (1998). *Principles and Standards of School Mathematics (draft)*. Reston, VA: The Author.
- National Research Council (1996). *National science education standards*. Washington, DC: National Academy Press.
- Peak, L. (1996). Pursuing excellence: A study of U.S. eighth-grade mathematics and science achievement in international context. Washington, DC: U.S. Department of Education.
- Rotberg, I. (1998). Interpretation of international test score comparisons. *Science*, 280, 1030-1031.
- Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. NY: John Wiley & Sons.
- Sande, I. (1982). Imputation in surveys: Coping with reality. *The American Statistician*, 36(3), 145-152.
- Schafer, J. L. (2000). *Analysis of incomplete multivariate data*. London: Chapman & Hall/CRC.
- Schmidt, W. H., & McKnight, C. C. (1998). What can we really learn from TIMSS? *Science*, 282(5395), 1830-1831.
- Silver, E. A. (1998). Improving mathematics in middle school: Lessons from TIMSS and related research. [On line] Available: <http://www.ed.gov/inits/Math/silver.htm> (September 14, 2000).
- Wang, R., Sedransk, J., & Jinn, J. (1992). Secondary data analysis when there are missing observations. *Journal of the American Statistical Association*, 87, 952-961.

AUTHOR

JIANJUN WANG is a full professor of educational statistics and research design at California State University, Bakersfield, 9001 Stockdale Highway, Bakersfield, CA 93311; jwang@academic.csuak.edu. His areas of specialization include comparative education, multilevel analysis, structural equation modeling, and stochastic process.