Why Using International Comparative Math and Science Achievement Data from TIMSS Is Not Helpful

by William G. Holliday
and Berchie W. Holliday

Collecting and using international comparative-achievement data is not helpful, even if Canadian and U.S. schools, for instance, far surpass other assessed countries. To use such data reported by the Third International Mathematics and Science Study (TIMSS) researchers for policy-changing purposes is a distracting and misleading approach to assessing science and math teaching in any group of countries with widely differing characteristics (Wang 2001). Such global comparative assessments generally serve no productive purpose, because very little useful information can be discovered and reasonably applied to existing school programs. Researchers working for TIMSS-like organizations, such as the Netherlands-based International Association for the Evaluation of Education Achievement (IEA), have provided us with other policy products of much greater usefulness.

Canada and the United States reportedly have uninspiring science and math education programs that lack reasonable rigor, based on their students' performances on achievements tests when matched with many so-called comparative performances reported by other countries (Holliday 1999a). Middle and high school teachers continually fail to teach their students the skills, strategies, and conceptual understandings realistically needed for numerical and scientific literacy, especially when comparing students' test results to those in some other countries. At least that is what TIMSS suggested in 1995 to most political and educational commentators (Sanchez 1998). Investigators later repeated some of these comparative testing programs and found that some countries switched places, arguably because of changes made in schools (National Council of Teachers of Mathematics [NCTM] 2001). The data reported in news reports (National Science Teachers Association [NSTA] 1998) and other international-assessment studies implicitly argue for changing teaching practices and policies as well as increasing funding for projects that promise to resolve problems found in school math and science programs.

The TIMSS achievement data and the following discussion should not be confused with other attempts by TIMSS researchers to document math and science teaching in countries using case-study approaches, which include videotapings of selected teaching in countries such as Japan, Germany, and the United States. First, what is so wrong with such international comparison studies assessing achieve-

ment? Second, why do math and science professional organizations, such as NSTA and NCTM, run so many front-page articles in their newsletters about how informative TIMSS is? Similarly, why are excellent scholars in science education, such as Andy Porter (2002), addressing conventions and publishing in world-class journals with language that also seems to endorse these TIMSS findings as having potential value in altering the way we teach students? No one has provided a reasonable answer to these questions. Yet concerned educators arguably should not base their policy changes on TIMSS achievement data.

PROBLEMS WITH TIMSS

Simply put, researchers engaging in international comparative studies like TIMSS work with inadequate funding resources. Politically available methodologies cannot possibly compare, in a valid fashion, the science and math achievement of students living in different cultures with students residing in the United States and English-speaking Canada. To do so would take much more funding than was reportedly provided TIMSS researchers.

A much more important hurdle to overcome is the unique set of cultural factors situated in each country, such as differential national languages, social norms, cultural prides, ethical standards, political systems, educational goals, and school curricula. Comparing achievement among varied cultures, for example, is arguably much more difficult than merely comparing apples and oranges, or apples and sauerkraut—a specific reference made to faulty comparisons of two first-world countries, Germany and the United States (Noack 1999). Even a major director of a TIMSS project admitted that such fairness is difficult, saying that no one wants to watch sausage, legislation, and international test items being constructed, according to Bracey (1997). Likewise, one official of another major international-comparison study reportedly expressed an analogous desire that exam items are made equivalently unfair to all participating nations (Bracey 1997).

First, language is an important cultural factor when comparatively assessing students who speak, read, write, and listen using entirely different communication systems. How can anyone construct math or science test items and reasonably guarantee equivalency of meanings across countries? Specialists in language may declare words expressed in different languages used in TIMSS items to be identical, but students living in different countries have their own interpretation of "equivalent" words. Linguists, for example, may look at a sentence contained in a test item expressed in two languages and agree that the wording is identical. Yet students may interpret the message in quite different ways, rendering such items invalid. So language across countries and within nations like the United States and Canada is a variable that test developers must consider carefully (Holliday 1999a). Indeed, it is difficult to compare curricula within countries like the United States (Kilpatrick 2003). Realistically comparing typical students living in English-speaking Canada and corresponding students living in the Czech Republic—

William G. Holliday *is Professor of Science Education at the University of Maryland at College Park. His research interests include science learning and teaching strategies, education policy, and narrowing the gaps between research and practice.*

Berchie W. Holliday *is a retired high school math teacher and a textbook author. Her research interests address teaching and learning secondary school mathematics.*

a country that significantly outscored Canada on one major math test, according to the National Research Council (1999)—for instance, fails to make much common sense.

We were unable to find items expressed in different languages, but we found possible language problems included in math items administered to Canadian and U.S. students. For example, some of the TIMSS math items released to the public surely contained some odd-looking words to U.S. and Canadian students—for example, "centros" and "Zeds" (units of money) and "Zedland" (country). Why were these words appearing in TIMSS items chosen over more conventional words like "dollars" and "Canada or United States," respectively? One wonders what was measured in these two items: students' mathematical abilities or their ability to handle unfamiliar words placed in mathematical problem-solving contexts.

Second, the manner in which students were sampled by governments with limited funding for conducting and managing such sampling procedures is another troublesome factor. These sampling procedures are problematic, especially when a central methodological goal is to examine how comparable students performed on equivalent tests administered under equivalent and fair conditions. We learned that only 5 (Czech Republic, Hungary, New Zealand, Sweden, and Switzerland) of the 21 participating nations sampling their students in a major part of TIMSS actually met the standards set by the TIMSS researchers (U.S. Department of Education [USDE] 1998). Not even the United States met the TIMSS's relaxed standards. What kind of study is taken seriously when so few participants meet basic sampling standards? This practice is inconsistent with conventional scientific-reporting procedures, but it may be considered okay in some government documents.

Third, curricula differ across countries, with students enrolled in different courses at differing ages. In this regard, we were surprised to learn many things. First, many students schooled in the United States and Canada and administered the advanced mathematics test, which contains calculus items, had never taken calculus, unlike students residing in some other countries who outperformed Canadian and U.S. students. Indeed, 22 percent of the released TIMSS advanced-mathematics items were categorized as calculus and required an understanding of integration and differentiation (USDE 1998). As teachers for more than a decade, we know that, if we administered a test in which 22 percent of the items covered material not presented in class, we would not expect our students to perform well.

There is no possible reason why U.S. and Canadian students enrolled in pre-calculus courses should be expected to solve calculus problems. Yet TIMSS described this unique U.S. and Canadian sample as "students in Grade 12 who had taken or were taking Advanced Placement calculus, calculus, or pre-calculus." Students in other countries, according to TIMSS's *Pursuing Excellence* (USDE 1998, 90), apparently were enrolled in a wide variety of advanced mathematics courses, perhaps placing them at a significant performance advantage. Why would Canadian or U.S. officials make such a sampling decision? Some may wish to make curricular changes with an emphasis on increasing the number of high school students completing calculus at the secondary level. Yet no one should compare our students' abilities to perform calculus problems with students living in widely different cultures who likely have advanced schooling on these mathematical topics.

Several TIMSS countries outperformed the United States and Canada on the math-

ematics general-knowledge assessment. The amount of secondary schooling in years afforded Canadian and U.S. students was often less than that afforded students schooled in other TIMSS countries, as candidly discussed in TIMSS's Table A5.14 (USDE 1998, 115). Students schooled in other lands were enrolled in grades 13 and 14. The sampled students in countries with grades 13 and 14 arguably left U.S. and Canadian 12th-grade students at a developmental disadvantage. Clearly, on average, the older the students and the more mathematics courses such students take, the better their math performances are likely to be.

Even minor curricular differences can make a significant difference in scores.

Different countries place differing curricular emphasis on students working cooperatively, creatively, and strategically. Even minor curricular differences can make a significant difference in scores. For example, most U.S. students (unlike Canadians) lack familiarity with the metric system situated in everyday contexts, yet TIMSS items reportedly used the metric system. Some may argue that U.S. schools should adopt the metric system exclusively, as other major first-world countries like Canada have. However, the point of this discussion is about comparative cultural fairness in testing rather than a discussion on curricular issues.

Another example of differential curricula in a broad sense is whether most sampled students received extensive schooling during the day, before school, after school, and on Saturdays—and how many hours per week the students were in a mathematics class. There is no reasonable way of comparing the curricula among so many countries, because such comparisons are confounded by unknown differences among students selected to take TIMSS's science and math tests. These unknown differences also include regional and national patriotic attitudes within and among countries toward doing well on such international tests. Finally, TIMSS researchers did not address differential learning experiences of students selected by ministries of education sometimes placed under a great deal of political pressure to outperform countries like the United States (Bracey 1996). Yet, in fairness, the TIMSS researchers cannot be held responsible for these confounding variables. These researchers did the best they could with the funds they had.

Fourth, there is good reason to believe that some government officials in their respective countries placed in charge of administering the TIMSS tests unfairly selected students for testing and did not act in accordance with TIMSS student-sampling standards. For one thing, it is doubtful that many participating second- and third-world countries have the resources and, in a few cases, political courage necessary to produce representative samplings of young people and subsequent test data needed to compare with students living in first-world countries (Bracey 1996). Moreover, in 16 out of 21 participating TIMSS countries, students attend specialized schools that are differentiated by students' abilities and career goals. These students generally are enrolled in a focused, preset curriculum, according to the TIMSS (Bracey 1998). In many of these countries, a small portion of their students is placed in high-performance schools emphasizing mathematics and science (Bracey

1998). It is, perhaps, this select group of students who are assessed using international-comparative tests. How fair is it to compare students with privileged experiences, enrolled in these specialized programs, with sampled U.S. and Canadian students who generally are attending ordinary high schools without regard to "their ability, prior academic performance, and career goals" (USDE 1998, 61)?

Therefore, we must continue to ask whether all nations have fairly sampled their student populations—that is, honestly and completely reported their administrative procedures and follow TIMSS rules—or did some nations "modify" the sampling rules and procedures to their nation's clear advantage? No one will ever know the answer to this question. Yet psychometric and political experts surely must have their opinions based on other contexts where some foreign governments have failed to adhere to agreed-to conditions or failed to make honest reports to international watchdog bodies in other circumstances. Reasonable sampling techniques should be a major concern for readers of TIMSS, because unreasonable sampling procedures produce data that cannot be interpreted. Such data are never considered when contemplating scientific questions of any kind.

Separately, Japan usually did quite well on earlier international comparison tests, but their students' exceptional performances might be a result of a variety of factors. These factors may include the large amounts of time students spent on schoolwork throughout the year, and the small amounts of time spent engaging in extracurricular activities such as earning money by working in part-time jobs. One reviewer of Japanese schools calculated that students of the same age in Japan and in the United States vary by two years of schooling because of Japan's emphasis on academics at early ages of their children. This and other cultural features of Japan may be praiseworthy and even worth emulation, but they shouldn't be confused with changing today's U.S. school policies because of data reported in any international comparative study. Berliner and Biddle (1995) elaborate on these concerns in the award-winning *The Manufactured Crisis*.

Fifth, students in different countries may have varying amounts of pride when administered an international test of achievement such as TIMSS. Students in some countries perhaps work hard to prepare for such examinations and spend a great deal of time practicing tasks similar to the ones encountered in the actual exams. In contrast, many U.S. and Canadian students may not take the same tests with equivalent seriousness and conscientiousness. Controlling for pride and practice is practically impossible, especially given the limited resources received by educators conducting such complex international studies as TIMSS.

Source of the Real Problems Linked to School Achievement

No one argues that schools are without serious problem, especially schools underfunded by state and provincial governments and schools that struggle to serve poverty-stricken students. Yet what are causes of the real problems linked to science and math achievement? No one knows. What would be logically helpful is additional public support, including much-needed parental participation, real help from corporations, and financial support for schools designed to serve all children. Indiscriminately blaming U.S. and Canadian schools and teachers for problems based in complex social issues is simply irresponsible (Barlow and Robertson 1994; Berliner and Biddle 1995).

TIMSS has fueled some of the fire focused on our teachers, suggesting that their performances in classrooms is where the real achievement problem exists while sel-

dom mentioning other more important sources of statistical variance factors linked with home environments, state and local government funding allocated to public schools, educational and income levels of parents, and other characteristics of student learning unrelated to schools. Recent research, for example, reveals strong relationships between school funding, household income, and science and math achievement (Biddle 1997; Holliday 1999b; Payne and Biddle 1999). School funding and household income account for much of the variance in scores on the 1996 U.S. National Assessment for Educational Progress (NAEP) science test given to eighth graders. Scores on the test correlate positively with school funding at the state level and negatively with household income. Moreover, statistical models have accurately predicted science-achievement scores using school-funding and household-income data. Funding and income together account for 53 percent of the variation in average U.S. science achievement—and 55 percent of the variance in the 1996 NAEP math test scores (Biddle 1997). In the Second International Mathematics Study (SIMS), school funding and household income were important predictors of eighth graders' scores, even when statistically controlling for the disproportionate numbers of minorities in low-income brackets (Payne and Biddle 1999). This correlational research supports the perception that science and math teachers situated in fortunate communities teach students who generally succeed academically. In contrast, teachers working in struggling communities are typically overburdened with teaching many less-fortunate students freighted with problems originating at home and in communities (Payne and Biddle 1999). Experienced educators are certainly not surprised by these sources of statistical variances.

Two other researchers, a former Stanford University dean and an established researcher born and living in England both specializing in international curricular reforms in science education, found that countries around the world ridicule their own nation's science educational programs, demanding major policy changes and basing their opinions on flimsy data and weak logic (Atkin and Black 1997). They expressed concern that some commentators closely aligned with the TIMSS study claim that TIMSS data supports the notion that the U.S. curriculum is "a mile wide and an inch deep," a slogan that has caught fire and been heard increasingly by science educators (Atkin and Black 1997, 26). This phrase is indeed flashy—and may be true—but it is not supported by TIMSS data (Atkin and Black 1997). These and other irresponsible comments, many of which aren't even based on TIMSS's published methodology, data, or discussion, spuriously argue why teachers of science and their colleagues in today's public K–12 schools are inadequate and incompetent, and why the public must make teachers change their teaching ways (Atkin and Black 1997).

A Popular Item for TIMSS's Math Test

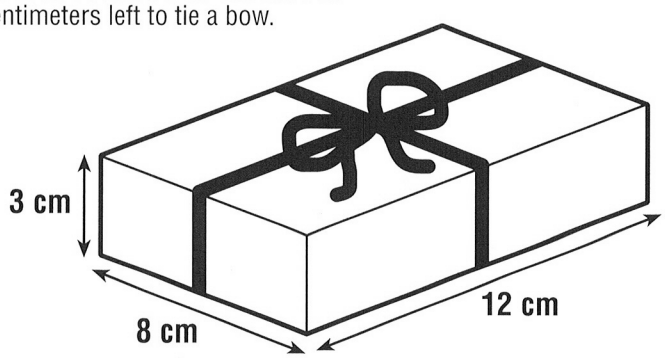
A popular TIMSS item among supporters of TIMSS appears in Figure One on page 256. This question, designed to assess 12th-grade U.S. math students, perhaps illustrates some subtle problems with comparing students' responses across nations. This particular TIMSS item, described as a Mathematics General Knowledge Item, has been cited often (Burrill 1998; Sanchez 1998).

We recommend that readers read and solve the problem and contemplate how U.S. students performed (U.S. average: 32 percent) in comparison to students schooled in other cultures (average: 45 percent). Specifically, we asked ourselves four

FIGURE 1

MATHEMATICS GENERAL KNOWLEDGE ITEM

Stu wants to wrap some ribbon around a box as shown below and have 25 centimeters left to tie a bow.



How long a piece of ribbon does he need?

A. 46 cm
B. 52 cm
C. 65 cm
D. 71 cm
E. 77 cm

Correct answer: E
U.S. Average: 32 percent
International Average: 45 percent
SOURCE: Third International Mathematics and Science Study, 1994–1995.

subtle questions after reading this item, based on our experiences dealing with U.S. high school math and science students.

First, how many U.S. students have a good sense of how to the measure 25 centimeters? Though Canadian students would not be disadvantaged by this issue, most U.S. students would be. We suspect that most U.S. students are clueless about exactly the length of 25 centimeters of ribbon, because this system of measurement is not regularly emphasized in schools and in society—unlike in Canada, where some students recognize that it is about 12 inches. Sure, the metric system is covered in math and science classes, but how much of this information is retained and useable by the average U.S. high school student, who very seldom uses metric measures beyond the classroom compared to students living in other countries?

Second, how many students "wrap some ribbon around a box"? Third, how many "tie a bow"? When wrapping packages, most U.S. and Canadian students would place a self-adhesive bow on the package without placing additional ribbon on the package. Fourth, how many students would want to know with any precision "how long a piece of ribbon" is needed to wrap a box with "25 centimeters left to tie a bow"? Thus, this word problem may be viewed as irrelevant to many students. Many of the Canadian and U.S. students we have taught and observed, even those enrolled in advanced pre-calculus and advanced-placement science courses, perhaps would call this a trick or unfair question that fails to assess students' real understanding of their achievement in math. Yet our queries about this popular item represent an empirical question. The past-president of NCTM and editors of *The Washington Post* thought the item was clearly appropriate and reasonably valid. Major U.S. and Canadian mathematics programs and textbooks typically include math problems of much greater relevancy to students than this particular item. This item isn't awful, but it does not deserve the praise heaped on it by non-practitioners.

REACTING TO TIMSS

When it comes to reacting to the TIMSS reports that have questioned by implication science and math teachers' abilities, the

public and policy makers have not appeared reluctant to blame today's teachers for problems evident in schools. Since the publication of *A Nation at Risk* (National Commission on Excellence in Education 1983)—an oft-cited report expressing lots of opinions but based on very little data—many U.S. and Canadian citizens seem ready to swallow almost any achievement-performance scores condemning schools and bashing teachers. How the data was collected does not seem to influence some readers who seem bent on blaming schools (Berliner and Biddle 1995).

Millions of dollars have been spent on TIMSS's attempt to compare achievement scores, money perhaps better spent on beginning efforts to improve teachers' salaries, working conditions, in-service courses, classroom materials, and parents' understanding about how math and science teachers are trying to teach their children. Moreover, the public seems uninformed about the increased need for classroom teachers competent in mathematics and science, as well as supplies and services required to do the job of producing mathematically and scientifically literate graduates. Of course, TIMSS researchers spent a relatively paltry sum on the comparative-achievement portion of their study, but we hate to see any money wasted on government-supported studies that fail to provide a basis for needed school improvement.

REFERENCES

Atkin, J. M., and P. Black. 1997. Policy perils of international comparisons: The TIMSS case. *Phi Delta Kappan* 79(1): 22–28.

Barlow, M., and H.-J. Robertson. 1994. *Class warfare: The assault on Canada's schools*. Toronto, Ontario: Key Porter Books..

Berliner, D. C., and B. J. Biddle. 1995. *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. Reading, Mass.: Addison-Wesley.

Biddle, B. J. 1997. Foolishness, dangerous nonsense, and real correlates of state differences in achievement. *Phi Delta Kappan* 79(1): 8–13.

Bracey, G. W. 1996. International comparisons and the condition of American education. *Educational Researcher* 25(1): 5–11.

Bracey, G. W. 1997. The seventh Bracey report on the condition of public education. *Phi Delta Kappan* 79(2): 120–36.

Bracey, G. W. 1998. Tinkering with TIMSS. *Phi Delta Kappan* 80(1): 32–38.

Bruer, J. T. 1997. Education and the brain: A bridge too far. *Educational Researcher* 26(8): 4–16.

Burrill, G. 1998. Changes in your classroom: From the past to the present to the future. *Journal of Research in Mathematics Education* 29(5): 583–96.

Holliday, G. W. 1999a. Questioning the TIMSS: Why international comparison studies like TIMSS say nothing to science teachers. *Science Teacher* 66(4): 38–41.

Holliday, W. G. 1999b. The bottom line in science. *Science Scope* 23(3): 8–9.

Kilpatrick, J. 2003. What works? In *Standards-based school mathematics curricula: What are they? What do students learn?* ed. S. L. Senk and D. R. Thompson, 471–88. Mahwah, N.J.: L. Erlbaum.

National Commission on Excellence in Education. 1983. *A nation at risk: The imperative for educational reform*. Washington, D.C.: U.S. Department of Education. ERIC ED 226 006.

National Council of Teachers of Mathematics. 2001. What can we learn from TIMSS-Repeat? *News Bulletin* 37(6): 1, 6–7.

National Research Council. 1999. Global perspectives for local action: Using TIMSS to improve U.S. mathematics and science education, Professional development guide. Washington, D.C.: National Academy Press.

National Science Teachers Association. 1998. U.S. twelfth graders rank poorly in latest TIMSS study: Students score below international average in math, science. *NSTA Report* 1, 10.

Noack, E. G. 1999. Comparing U.S. and German education: Like apples and sauerkraut. *Phi Delta Kappan* 80(10): 773–76.

Payne, K. J., and B. J. Biddle. 1999. Poor school funding, child poverty, and mathematics achievement. *Educational Researcher* 28(6): 4–13.

Porter, A. C. 2002. Measuring the content of instruction: Uses in research and practice. *Educational Researcher* 31(7): 3–14.

Sanchez, R. U. S. 1998. High school seniors rank near bottom: Europeans score higher in math, science test. *The Washington Post*, 25 February, A1, A12.

U.S. Department of Education. 1998. *Pursuing excellence: A study of U.S. twelfth-grade mathematics and science achievement in international context, initial findings from the Third International Mathematics and Science Study*. Washington, D.C.: National Center for Education Statistics, Office of Educational Research and Improvement, USDE.

Wang, J. 2001. TIMSS primary and middle school data: Some technical concerns. *Educational Researcher* 30(6): 17–21.