An Analysis of Item Score Difference Between 3rd and 4th Grades Using the TIMSS Database

Jianjun Wang
Department of Advanced Educational Studies
California State University, Bakersfield, CA 93311

ABSTRACT

Primary school data from the Third International Mathematics and Science Study (TIMSS) are analyzed in this article to examine performance difference between 3rd and 4th grades. Score comparisons are determined across all TIMSS items in each of the participating countries, using computer technology and programming to complete the thousands of score subtractions. The empirical findings indicate that not all TIMSS items have resulted in a higher mean score at the upper grade level. Item features are discussed to characterize part of the released TIMSS instrument that generates a higher average score at the lower grade. This research outcome may facilitate articulation of the TIMSS benchmark with specific patterns of item performance to enrich understanding of the test results among education stakeholders.

Introduction

The global market competition has led countries around the world to pay close attention to student academic preparation in mathematics and science. For instance, the U.S. federal government announced Goals 2000 demanding the best student performance in mathematics and science (U.S. Department of Education, 1999), and benchmark data have been gathered from local schools using the established methods from the *Third International Mathematics and Science Study* (TIMSS) (Martin, Mullis, Beaton, Gonzalez, Smith, & Kelly, 1998). Kelly (2002) observed, "By articulating performance at the TIMSS international benchmarks, 'world class' achievement has been defined" (p. 41).

At the level of elementary education, TIMSS researchers gathered student scores from 3rd and 4th grades in 24 nations. Schmidt and McKnight (1998) noted, "The use of adjacent grades in the third/fourth- and seventh/eighth-grade populations allow the estimation of differences between cross-section samples of grade pairs, which is a fair surrogate for gains that might have been measured by a true longitudinal design" (p. 1830). Interpretation of the score gap between the adjacent grades hinges on quality of the test items that are sensitive to the school learning process. The purpose of this investigation is to examine the item score difference between the 3rd and 4th grades from all 24 nations. Because similar items are used in a TIMSS trend study in 2003, results of this investigation may not only help disentangle patterns of student item performance related to the average scores in the existing TIMSS reports but also facilitate understanding of the TIMSS trend report in the near future.

Uploaded Nov 20, 2004

Literature Review

Given the importance of mathematics and science education to the world economy, many countries participated in comparative studies like TIMSS over the last 10 years. Besides the original TIMSS in 1995 and the TIMSS trend study in 2003, there was a repeat of the TIMSS project (TIMSS-R) conducted in 1999. But TIMSS-R did not cover primary schools (Mullis et al., 2000), and cannot be employed to examine the average score difference between the 3rd and 4th grades.

On the other hand, *Programme for International Student Assessment* (PISA) is another project initiated in 2000 by the Organisation for Economic Co-operation and Development (OECD) (Bracey, 2004). Prais (2003) noted that the PISA project was not designed to assess effectiveness of a school system, and thus, "the verdict of the previous IEA inquiries (TIMSS)—which were focused on the school curriculum—still needs to be accepted" (p. 144-145).

Importance of the Item Checking

Upheld as a credible benchmark in comparative education, TIMSS has been the only project that gathered a large-scale database at 3rd and 4th grade levels. Because of the designated focus on the school curriculum, TIMSS test items are expected to result in a higher average score at the upper grade of the same country. The chronicle counts show that the learning experience between these two grades consists of approximately 25% of students' school life up to the 4th grade. Thus, the additional schooling should not cause a drop of student performance at the upper grade. More importantly, failure to detect the between-grade difference also contradicts a postulation that TIMSS item scores are sensitive to the school curriculum coverage across the participating nations.

Despite contextual differences among various nations, it would be incomprehensible to observe a drop of academic achievement on the same set of test items as students move to a higher grade within that country. Although TIMSS is not a longitudinal study, the average difference in academic performance can be employed to measure a cross-sectional gap between adjacent grades in each nation.

Riley, McGuire, Inman, and Dorfman (1998) noted that "one basic use of TIMSS at the state and local level is as a benchmark" (p. 9). Accordingly, checking the item scores between two adjacent grades is not a trivial research undertaking. Items that cause a reverse of the score gap would collectively lead to an invalidation of the TIMSS benchmark. Despite cultural differences among various nations, no parents expect their children to become less knowledgeable as the children move to a higher grade. This type of abnormal results not only defies common sense in education but also contradicts the fundamental IRT premise (Baker, 2001) pivotal to the TIMSS assessment.

Need of the Item Checking

Inadvertently, the TIMSS researchers did not have a chance to investigate the link between student learning experience and the content covered by TIMSS items. Fensham (1998) recalled that he had raised this issue at an implementation stage of the TIMSS assessment but it was too late to gather the student information relevant to the item coverage at the 3rd and 4th grades (i.e., Population 1). Fensham (1998) reported,

> At the time of my question in 1995 only Population 3 in TIMSS, the final secondary year students, was still to be tested in Australia as Population 1 (9 year olds) and Population 2 (13 year olds) had been tested towards the end of the previous year. (p. 481)

Consequently, researchers were unable to collect the information directly from students regarding their learning experiences pertaining to the TIMSS items, and thus, it remains unclear whether the between-grade difference can be reflected in the test scores.

Thus far, no researcher has examined patterns of the mean score difference using the TIMSS achievement data between the 3rd and 4th grades. Instead, the performance comparisons have been largely confined within total scores or a set of subcategory scores (Martin et al., 1997; Mullis et al., 1997). Schmidt, McKnight, Cogan, Jakwerth, and Houang (1999) observed:

> TIMSS achievement reporting thus far has been limited to global mathematics and science scale scores and to reporting the national percentages of items correct in a set of six 'reporting categories' in both subjects. These reporting categories were still so broad—as the global scores obviously were—as to include somewhat disparate items. (p. 117)

Beyond these designated subject categories, more indepth investigations need to be conducted on test scores at the item level to examine relevancy of the TIMSS assessment to school learning processes. Whereas the item scores were gathered directly from students, the overall total scores were not. In fact, a matrix sampling technique was employed to assign part of the TIMSS instrument to each student, and the total score was estimated from data imputations (Gonzalez & Smith, 1997). Therefore, the total score comparison is built on an assumption of particular imputation models, and inevitably, additional variations could be attributed to the statistical artifact.

In this regard, results from the item score analyses may reveal findings that are not otherwise available from total score comparisons. Although school curricula may vary across different countries, the between-grade comparison is made within each country, and the item performance can be linked to the domestic condition of science and mathematics education. Schmidt et al. (1999) assert that "it is precisely these content-specific differences among items that make achievement assessments curricularly sensitive" (p. 116). On the basis of the achievement data, quantitative and qualitative inquiries have been incorporated in this study to examine and discuss student item scores that impact the TIMSS assessment.

Method

TIMSS has designated its Population 1 at two adjacent grades that have most students at age 9. The 3rd and 4th grades are the two levels investigated in most participating nations, except for Israel and Kuwait. The existing TIMSS reports are based on a representative sample of students at each grade in each nation (Martin et al., 1997; Mullis et al., 1997).

Although it has been claimed that the item selection was grounded on an international consensus (Martin, 1996), a group of TIMSS researchers noted that "due to the tremendous curricular variability across nations and the desire to over-sample some topic areas, the TIMSS test varied in its match to any particular curriculum" (Jakwerth et al., 1997, pp. 7-8). Consequently, an item score comparison across all the nations does not reflect variation of the curriculum match to the test instrument.

Instead, the item performance should be examined in each nation separately. For students studying within the same school system, those at the 4th grade should perform better than their peers at the 3rd grade on the average item scores. To examine the performance difference, a simple approach is to subtract the item mean scores between adjacent grades in each nation. If this issue involves only one item or a few items, the subtraction can be easily completed through hand-calculations. For a large number of items in the TIMSS instrument (Lange, 1997), the entire computing involves several thousand subtractions of the item mean scores. Without a computer program, no researchers have made an indepth comparison of the item scores in all participating countries (TIMSS International Study Center, 1999).

Fortunately, a solid comparison method has been developed in an analysis of the TIMSS item scores between 7th and 8th grades (Wang & Zhu, 2003). The first step of the computing is to output the mean item scores for each grade in each nation. In the SAS or SPSS software applications, the output average scores are laid out across columns, and manual coding is needed to handle the score subtraction between any adjacent columns. To complete the computing across all items over all nations, the program coding could be extraordinarily long.

One way to avoid the tedious programming is to transpose the output data matrix, and have the mean item scores listed in a column. A LAG function is available in SAS and SPSS to complete the needed subtractions between the adjacent rows. Because the subtraction is carried out automatically throughout the list of item mean scores from all items in various nations, Wang and Zhu (2003) developed computer codes to avoid the redundant subtractions over different items across the nation boundaries, and thus, retain the needed mean score difference for the same item within each nation.

This well-established method is adapted in this study to complete a similar computing of average score difference between the 3rd and 4th grades. Whereas most TIMSS reports did not cover item score comparisons across all participating nations, two reports (i.e., Martin et al., 1997; Mullis et al., 1997) have released a few item scores for discussion. These results have been reconfirmed in the data analysis to ensure a proper access to the TIMSS database. The essential program code in SAS is outlined in Table 1.

Table 1. **SAS Statements to Compute Item Score Difference Between Adjacent Grades**

```
* IDCNTRY – country names;
* IDGRADER – grades;
* TOTWGT – sampling weight;
* ASMMA01 – ASESZ03 (TIMSS item scores);

* (after reading the TIMSS data into SAS);

proc sort;
        by idcntry idgrader;

proc means noprint;
        class idcntry idgrader;
        var ASMMA01--ASESZ03;
        weight totwgt;
        output out=new(where=(_type_=3)) mean=;
```

Table 2 cont.

```
data two;
        set new;
        n=_n_;
proc transpose data=two out=three;
        by n idcntry idgrader;
proc sort;
        by _name_ idcntry idgrader;
data last;
        set three;
        drop n;
        by _name_ idcntry idgrader;
        mean_diff=dif(col1);
        if first.idcntry then mean_diff=.;
        if mean_diff=. then delete;
        if mean_diff<0;
proc sort;
        by _name_;
proc print;
        var IDCNTRY IDGRADER _NAME_ mean_diff;
        run;
```

Result

Outcomes of the item score analysis are listed in Table 2.

Table 2. **Number of Items Resulting in Higher Average Scores at the Lower Grade Level**

Country	Number of Items
Australia	4
Austria	7
Canada	1
Cyprus	5
Czech	4
England	5
Greece	7
Hong Kong	9
Hungary	4
Iceland	6
Iran	12
Ireland	2
Japan	7

Country	Number of Items
Korea	16
Latvia	7
Netherlands	2
New Zealand	4
Norway	2
Portugal	7
Scotland	8
Singapore	9
Slovenia	5
Thailand	15
U.S.	2
Total	**150**

Inspection of Table 2 suggests that not all TIMSS items have resulted in a higher mean score at the upper grade level. This pattern existed in all participating nations that involved two adjacent grades in the investigation. The issue seemed to be widespread and a total of 150 abnormal cases showed a reverse of student performance, i.e., having 4th grade students score lower than their peers at the 3rd grade in the same nation.

Discussion

The problem of having a higher average performance at a lower grade varies in its extent across nations (Table 2). In Canada, only one item has such a problem. On the other hand, 16 items demonstrate this problem in the Korean data. The number of the seemingly problematic items for the United States is 2, less than that of top-performing countries, such as Japan (7) and Singapore (9). The variation of empirical outcome is in line with the observation of Jakwerth and his colleagues, i.e., "the TIMSS test varied in its match to any particular curriculum [of the participating nations]" (Jakwerth et al., 1997, pp. 7-8).

To date, few researchers have examined the academic content of the TIMSS instrument, and this omission can be partly explained by endorsement of the items by the TIMSS Subject Matter Advisory Committee, which included "distinguished scholars from 10 countries" (Beaton et al., 1996, p. A-9).

On the other hand, Fensham (1998) recollected,

> At one of the later meetings of the Science Subject Matter Advisory Committee for TIMSS, I innocently asked members of the overall coordinating group whether they know if any country was investigating what the students in the sample thought about the tests and the testing as a whole.
>
> To my surprise, this question, though simple to conceive and to ask, created quite a stir. (p. 481)

With due respect to these scholars of the Subject Matter Advisory Committee, an examination is still needed to determine whether the expert endorsement guarantees a proper fit between the test items and student cognitive development levels.

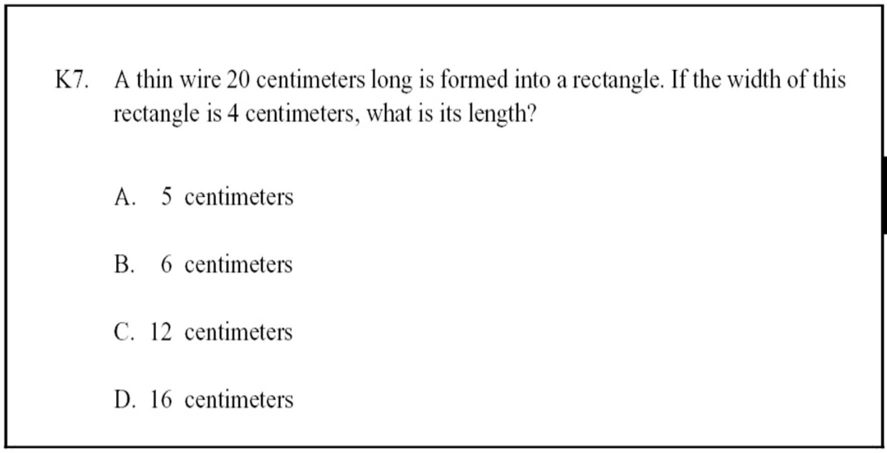
Based on developmental psychology, students' test-taking approaches could be differentiated from those of grownups. More specifically, students of the 3rd and 4th grades may have cognitive skills at a concrete operational level (Piaget, 1985). Therefore, their reasoning process often needs support from concrete examples. When a test item has conflicting answers, the mental equilibrium is disturbed. As a result, the confusing item may generate chaotic responses that inadvertently cover up the average achievement difference between these adjacent grades. For instance, one TIMSS item [item name: ASMSO02] reads:

> John kept some seeds on moist cotton in a dish. Mike put the same kind of seeds in a dish besides John's dish, and covered them with water. After two days, John's seeds sprouted, but Mike's did not.
>
> Which is the most likely reason?
>
> A. Mike's seeds needed more air.
>
> B. Mike's seeds needed more light.
>
> C. Mike did not put the dish in a warm enough place.
>
> D. Mike should have used a different kind of seed.

When *option A* was used as the correct answer to grade student performance, some countries, such as Singapore, Thailand, Iran, and Greece, had the 3rd graders receiving a higher mean score than the 4th graders on this item. In part, this could be because *option A* did not appear to be the only correct answer to this item. Through daily observations, students may have noted that some seeds covered with water can still sprout. Lotus seeds represent a simple example for such cases. Therefore, the answer could also be *option D*, *Mike should have used a different kind of seed.* Unfortunately, this reasoning process built on concrete experiences has led to no credit for these students.

The extent to which this issue might exist is not easy to assess because not all TIMSS items have been released to the public. Since two thirds of the items are saved for future use, a discussion of the assessment outcome should consider ongoing improvement made by the TIMSS team in recent years. One of the most significant changes is a switch of the item scoring from a one-parameter model in TIMSS to a three-parameter model in TIMSS-R (Mullis et al., 2000). An improved feature from this change is consideration of potential guessing effects in student response, which seems relevant to more than 90% TIMSS items that are in a multiple-choice format (Lange, 1997). According to Hambleton (1988), "with difficult multiple-choice tests, a researcher might anticipate considerable guessing on the part of examinees" (p. 154). The difficulty level can be represented by the overall percentage of correct responses. Thus, the additional consideration of the guessing effect could be effective if the issue of reversed score gaps is confined among items with a relatively low percentage of correct responses.

For example, the item below has four options. Thus, the probability of obtaining a correct answer through random guessing is 25%. Across all TIMSS participating nations, however, only 21% third graders and 23% fourth graders answered this question correctly (TIMSS International Study Center, 1999). The low rate of correct responses seemed to suggest that this item was too difficult for these students. Besides speculation of a potential guessing effect, this item also revealed 10 (Australia, Austria, Greece, Iceland, Ireland, Japan, New Zealand, Portugal, Scotland, and the United States) out of 26 participating countries having a lower average item score at the upper grade. In addition, seven other countries (Czech Republic, Hungary, Iran, Korea, Norway, Thailand, and Scotland) showed a gap of the item correct response less than 3% between the 3rd and 4th grades.



K7. A thin wire 20 centimeters long is formed into a rectangle. If the width of this rectangle is 4 centimeters, what is its length?

A. 5 centimeters

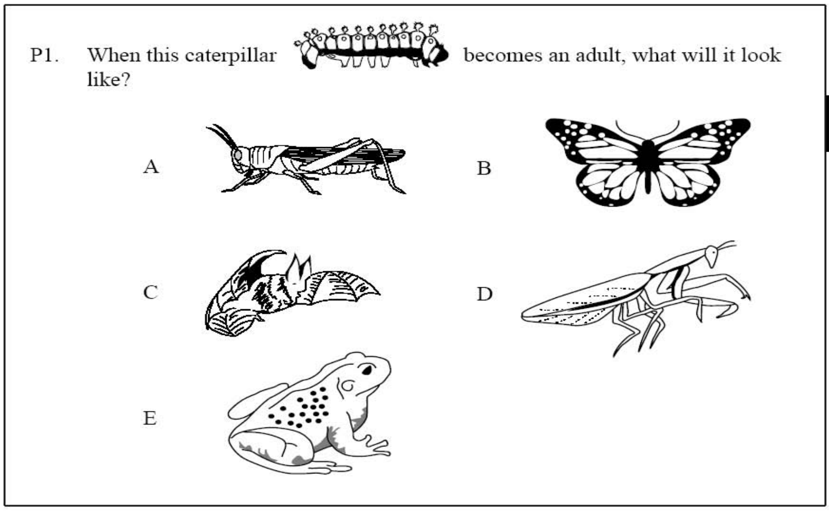
B. 6 centimeters

C. 12 centimeters

D. 16 centimeters

K-7

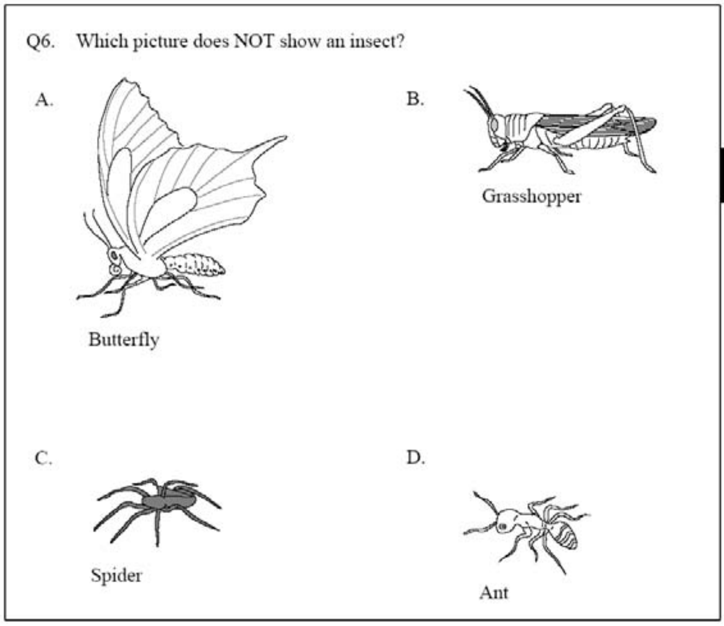
In contrast, the next item is apparently easy because 85% fourth graders and 82% third graders across all participating nations answered this question correctly. With not much concern on a guessing effect for this easy item, a comparison in each nation still shows 10 participating countries either *having the gap less than 3%* or *having the lower graders score higher than the upper graders* (e.g., Korea) on this item.



P1. When this caterpillar becomes an adult, what will it look like?

P-1

In between the previous two items, the following example dealt with a middle level of difficulty. Among all participating countries, 43% of the 4th graders and 41% of the 3rd graders answered this question correctly. Nevertheless, a total of 11 countries either showed the percentage of correct-response gap less than 3% between the two adjacent grades, or demonstrated a better average performance at the lower grade. These countries include top-performing nations, such as Japan, Korea, and Singapore.



Q6. Which picture does NOT show an insect?

Q-6

In summary, the item examination indicates that the problem of lowering the average score at the upper grade has been spread out among items of various difficulty levels. All 24 nations demonstrate this issue for a total of 150 times (Table 2). The two other nations, Israel and Kuwait, did not involve two adjacent grades in their TIMSS data collection, and thus, it remains unclear whether similar issues are relevant to them. Because some of the TIMSS items have been employed in the TIMSS 2003 data collection, for which the results are yet to be released, results from this investigation may benefit these research projects in the future.

About the Author

Jianjun Wang is a full professor of educational research and statistics at California State University, Bakersfield. He is interested in stochastic processes and comparative studies in mathematics and science education.

Correspondence concerning this article should be addressed to: Jianjung Wang, Department of Advanced Educational Studies, California State University, Bakersfield, 9001 Stockdale Highway, Bakersfield, CA 93311. Phone: (661) 664-3048. Fax: (661) 664-2016. E-mail: jwang@csub.edu.

References

Baker, F. B. (2001). *The basics of item response theory.* Washington, DC: ERIC Clearinghouse on Assessment and Evaluation.

Beaton, A., Mullis, I., Martin, M., Gonzalez, E., Kelly, D., & Smith, T. (1996b). *Mathematics achievement in the middle school years.* Chestnut Hill, MA: Boston College.

Bracey, G. W. (2004). International comparison: Less than meets the eye? *Phi Delta Kappan, 85*, 477-478.

Fensham, P. (1998). Student response to the TIMSS test. *Research in Science Education, 28* (4), 481-489.

Gonzalez, E. J., & Smith, T. A. (1997). *Users guide for the TIMSS international database.* Chestnut Hill, MA: TIMSS International Study Center.

Hambleton, R. K. (1988). Principles and selected applications of item response theory. In R. Linn (Ed), *Educational measurement* (3rd ed.). London: Collier Macmillan.

Jakwerth, P., Bianchi, L., Houang, R, Schmidt, W., Valverde, G., Wolfe, R., & Yang, W. (1997, April). *Validity in cross-national assessments: Pitfalls and possibilities.*

Paper presented at the 1997 American Educational Research Association annual meeting, Chicago, IL.

Kelly, D. L. (2002). The TIMSS 1995 international benchmarks of mathematics and science achievement: Profiles of world class performance at fourth and eighth grades. *Educational Research and Evaluation*, *8* (1), 041-054

Lange, J. D. (1997). *Looking through the TIMSS mirror from a teaching angle.* [Online] Retrieved from http://www.enc.org/topics/timss/additional/documents/0,1341,CDS-000158-cd158,00.shtm (April 10, 2003).

Martin, M. (1996). Third international mathematics and science study: An overview. In M. Martin & D. Kelly (Eds.), *Third international mathematics and science study: Technical report.* Chestnut Hill, MA: Boston College.

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.

Martin, M. O., Mullis, I. V. S., Beaton, A. E., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1998). *Science achievement in Missouri and Oregon in an international context: 1997 TIMSS benchmarking.* 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.

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.

Piaget, J. (1985). *The equilibration of cognitive structures: The central problem of intellectual development.* Chicago, IL: University of Chicago Press.

Prais, S. J. (2003). Cautions on OECD's recent educational survey (PISA). *Oxford Review of Education*, *29*, 139-164.

Riley, R. W., McGuire, C. K., Inman, D., & Dorfman, C. H. (1998). *What the Third International Mathematics and Science Study (TIMSS) means for systemic school improvement.* Washington, DC: The Government Printing Office.

Schmidt, W. H. & McKnight, C. C. (1998). What can we really learn from TIMSS? *Science, 282* (5395), 1830-1831.

Schmidt, W. H., McKnight, C. C., Cogan, L. S., Jakwerth, P. M., & Houang, R. T. (1999). *Facing the consequences: Using TIMSS for a closer look at U.S. mathematics and science education.* Boston, MA: Kluwer.

TIMSS International Study Center. (1999). *TIMSS IEA's Third InternationalMathematics and Science Study*. Retrieved from http://isc.bc.edu/timss1995i/TIMSSPDF/ BMItems.pdf.

U.S. Department of Education. (1999). *Archived information: section 102. national education goals.* Retrieved from http://www.ed.gov/legislation/GOALS2000/ The Act/sec102.html

Wang, J. & Zhu, C. (2003). An in-depth analysis of achievement gaps between 7th and 8th grades in the TIMSS database. *School Science and Mathematics, 103* (4), 186-191.