DOCUMENT RESUME

ED 475 358 TM 034 834

AUTHOR	Wang, Jianjun
TITLE	An Analysis of Item Score Difference between 3rd and 4th Grades Using the TIMSS Database.
PUB DATE	2003-04-00
NOTE	19p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Comparative Analysis; Elementary Education; Foreign Countries; *International Studies; *Mathematics Achievement; *Science Achievement; *Scores; *Test Items
IDENTIFIERS	*Third International Mathematics and Science Study

ABSTRACT

Primary school data from the Third International Mathematics and Science Study (TIMSS) were analyzed in this study to examine performance difference between third and fourth grades. The score comparison was conducted across all TIMSS items in each of the 26 participating countries. 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 finding may facilitate articulation of the TIMSS benchmark with specific patterns of item performance to enrich understanding of the test results among education stakeholders. (Contains 4 tables and 17 references.) (Author/SLD)

Running head: TIMSS Database

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

Jianjun Wang

Department of Advanced Educational Studies

School of Education

California State University

9001 Stockdale Highway

Bakersfield, CA 93311-1099

E-mail: jwang@csub.edu

Phone: (661) 664-3048

Fax: (661) 589-2466

Paper presented at the 2003 annual meeting of the American Educational Research Association (AERA), Chicago, IL.

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

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. The score comparison is conducted across all TIMSS items in each of the participating countries. 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.

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

The Third International Mathematics and Science Study (TIMSS) is the largest and most ambitious project in comparative education (Gonzalez & Smith, 1997). At the primary school level, TIMSS researchers gathered student scores from the 3rd and 4th grades in 26 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. 1030). The purpose of this investigation is to examine the item score difference between the 3rd and 4th grades. As more educators consider using the TIMSS benchmark to evaluate school effectiveness (Martin et al., 1998; Mullis et al., 1998), results of this study may help disentangle patterns of student item performance related to the average scores of TIMSS reports.

Literature Review

In general, the item score difference comprises variation of student achievement in specific subject domains. Accordingly, Fensham (1998) suggested the need of investigating impact of student learning experience on the contents covered by TIMSS items. At the time he raised this important topic, it was too late for TIMSS researchers to gather the information of student learning experiences relevant to the instrument 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 information directly from students regarding their learning experiences related to the TIMSS items, and thus, it remains unclear whether the between-grade difference can be reflected in the test scores.

Based on the chronicle counts, the learning experience between 3rd and 4th grades consists of approximate 25% of students' school life. Regardless of contextual differences among various nations, it is incomprehensible to observe a drop of academic achievement on the same set of test items as students move from a lower grade 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 the adjacent grades in each nation. If the test content is covered by a curriculum at the upper grade, the results should show an increase in academic achievement. In addition, maturation and cognitive development are also in favor of the senior students, causing higher scores at the upper grade (Peterson, 1986; Walker & Madhere, 1987).

To date, no researchers have confirmed this pattern of score difference using the TIMSS achievement data. Instead, 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 in-depth investigations need to be conducted on test scores at the item level. 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 basis of the achievement data, quantitative and qualitative inquiries have been incorporated in this study to examine students' item scores and their patterns of test taking that impact the TIMSS assessment.

Methods

Given the average scores of a test item, this study is to confirm if students at the upper grade have met a general expectation of outperforming their peers at the lower grade in each nation. A simple approach to checking the score difference is to subtract item mean scores between the 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 overall computing operation involves thousands subtractions across the nations. Without a computer program, no researchers have made the in-depth comparisons of item scores in all participating countries (http://www.timss.org/timss1995i /Items.html).

In this study, the computer-based data analysis involves two steps: (1) compute mean item scores for each grade; (2) subtract the mean item scores between the 3rd and 4th grades in each nation. Using standard statistical software packages, such as SPSS or SAS, one can easily complete the first step, and export the mean item scores into a new database. Because the mean scores from different grades and countries are listed under separate variables of the output file (see Table 1), the following SAS codes are employed to transpose the variable names in Table 1 into a column in Table 2:

```
proc transpose data=TIMSS95 out=new;

by n idcntry idgrader;

var ASMMA01--ASESZ03;
```

Insert Tables 1 & 2 around here

A *LAG* function is subsequently introduced to calculate the mean score difference between adjacent rows (SAS Institute, 1990).

As the item mean scores are arranged by *grade* and *country*, this operation inevitably includes score subtractions between the last record of the previous country and the first record of the next country. In the transposed data structure (Table 2), these records are linked to *different* items of adjacent countries. This portion of results is meaningless since a lower grader in Japan may not necessarily score lower than a higher grader from South Africa on *different* test items.

To remove these redundant subtraction results across the country borders, a SAS command "if first.idcntry then mean_diff=.;" is issued. In addition, a statement "if mean_diff < 0;" is employed to single out these items that have resulted in higher average scores at the lower grade in each country. For some readers who have an extensive interest in details of the computation, the author has included an actual SAS program for the aforementioned data analysis (see Table 3). Because the LAG function is also available in SPSS, this approach can be readily adapted in SPSS-based data analyses (e.g., see SPSS, 1988).

Insert Table 3 around here

Results

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 at an initial stage of the data analysis to ensure a proper access to the TIMSS database. Outcomes of the item score analysis are assembled in Table 4.

Insert Tables 4 around here

Inspection of Table 4 suggests that not all TIMSS items have resulted in a higher mean score at the upper grade level. This pattern exists in all participating nations except for those that gathered data from a single grade. The extent of having a higher item score at the lower grade level also varies among the nations. In Canada, only one item (item name: ASMMM07) has such a problem. On the other hand, sixteen items demonstrate this problem in the Korea data. The number of the seemingly problematic items for U.S. is 2, less than that of top performing countries, such as Japan (9) and Singapore (3).

Discussions

Riley, McGuire, Inman, and Dorfman (1998) maintain that "One basic use of TIMSS at the state and local level is as a benchmark" (p. 9). Quality of the TIMSS benchmark generally depends on two components: (1) item scores that truly reflect student achievements, and (2) imputation methods that appropriately compile the total scores for TIMSS reporting. Discussions of student performance are centered on these two points to enrich understanding of the TIMSS benchmark at the 3rd and 4th grades.

The aforementioned first component deserves a special attention because no evidence has been gathered to represent students' thought about the TIMSS instrument. In particular, 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)

It remains unclear whether the lack of student perspectives has contributed to reverse of the mean score difference between the 3rd and 4th grades. This general issue is pertinent to the original design of TIMSS projects across all countries. According to Fensham (1998), "Quite late in the planning of this very expensive study, it transpired that no country had considered gathering data on the students' sense of the relevance of the science topics in the achievement tests, of their science learning, or, their metacognitive awareness of the learning" (p. 481).

Based on developmental psychology, students' test-taking approaches could be different from those of grown-ups. 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 achievement difference between two adjacent grades. For instance, one of the TIMSS items 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.

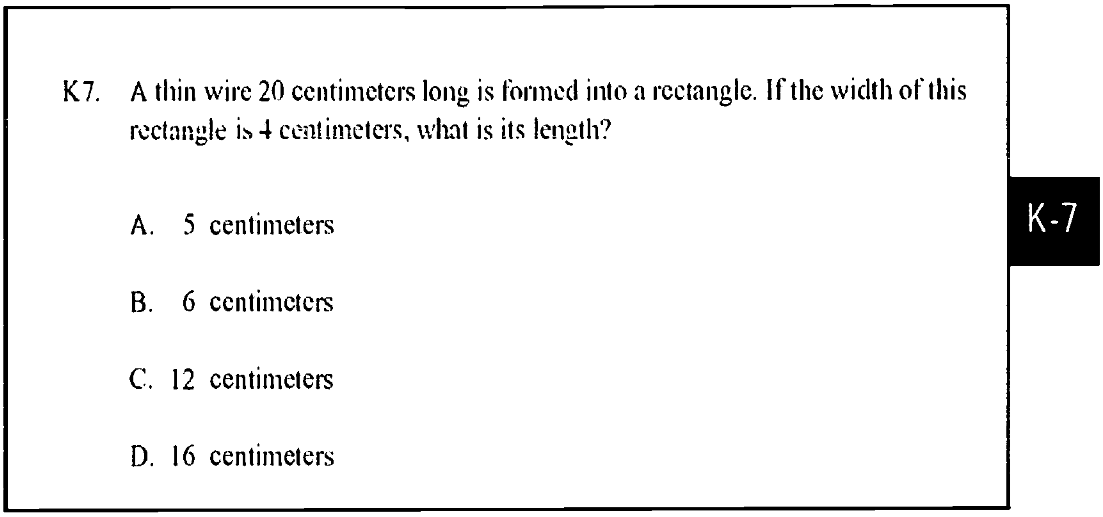
C. Mike did not put the dish in a warm enough place.

D. Mike should have used a different kind of seed. [item name: ASMSO02]

When option A was used as the correct answer to grade student performance, some countries, such as Singapore, Thailand, Iran, and Greece, had 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. In addition, the multiple-choice format did not allow students to explain the reasoning process.

Besides the concern on item scores, the total score imputation was also grounded on an assumption of *no guessing behavior* among the test takers (Gonzalez & Smith, 1997). Lange (1997) noted that more than 90% TIMSS items were in a multiple-choice format. Thus, the opportunity for guessing cannot be completely ruled out. As an example, the released TIMSS data contain student scores for the following item:



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

With the four options, 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 (http://isc.bc.edu/timss1995i/TIMSSPDF/BMItems.pdf). The low rate of correct responses seemed to suggest that this item was too difficult for these students. In general, "with difficult multiple-choice tests, a researcher might anticipate considerable guessing on the part of examinees. Needed, therefore, would be a model that could handle this situation" (Hambleton, 1988, p. 154).

Inadvertently, the imputation model employed in TIMSS did not contain a parameter to describe the effect of guessing (Gonzalez & Smith, 1997). In addition, the seemingly too difficult item could not differentiate student achievement in some countries. The United States was among 10 countries that had the third graders scoring higher than the fourth graders on this item[1]. Counterexamples from the response pattern analyses should caution users of the TIMSS benchmark to be more mindful when interpreting the final scores from TIMSS reports (e.g., Martin et al., 1997; Mullis et al., 1997).

In summary, despite curricular differences across nations, fourth graders are generally expected to outperform their peers at the 3rd grade due to *maturation* and *additional learning experiences* in the same country. Item scores that have reverted this pattern, however, do not reflect the impact of *mental growth* and *educational advancement*. The quantitative data analysis in this study has identified the item numbers in each country to assess extensiveness of this problem with TIMSS scores. Response patterns discussed at the item level represent an attempt to articulate the TIMSS benchmark with students' cognitive development to enrich understanding of the test results among education stakeholders.

Notes:

[1] Countries that show a reverse performance gap on item K-7 are: Australia, Austria, Greece, Iceland, Ireland, Japan, New Zealand, Portugal, Scotland, and United States.

References

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. L. Linn (Ed), Educational measurement (3rd ed.). London: Collier Macmillan.

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 (April 10, 2003).

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.

Martin, M. O., Gregory, K. D., & Stemler, S. E. (2000). TIMSS 1999: Technical 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.

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., & Smith, T. (1998). Mathematics achievement in Missouri and Oregon in an international context: 1997 TIMSS benchmarking. Chestnut Hill, MA: TIMSS International Study Center.

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.

Peterson, A. C. (1986, April). Early adolescence: A critical development transition? Paper presented at the 67th annual meeting of the American Educational Research Association, San Francisco, CA.

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

SPSS (1988). SPSS-X user's guide (3rd ed.). Chicago, IL: Author.

SAS Institute (1990). SAS procedures guide (3rd ed.). Cary, NC: Author.

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.

Walker, E. M., & Madhere, S. (1987). Multiple retentions: Some consequences for the cognitive and affective maturation of minority elementary students. Urban Education, 22 (1), 85-102.

Table 1

Mean score layout from SAS PROC MEANS

IDCNTRY	IDGRADER	ASMMA01	ASMMA02	ASMMA03
840	low	#	#	#
	up	#	#	#
890	low	#	#	#
	up	#	#	#

Note:

IDCNTRY – country codes
IDGRADER – grade codes
ASMMA01, … – TIMSS item names

Table 2

Partial transpose of mean score layout from SAS PROC MEANS

IDCNTRY	IDGRADER	_Name_	Col1
...			
840	low	ASMMA01	#
	up	ASMMA01	#
890	low	ASMMA01	#
	up	ASMMA01	#
...			
840	low	ASMMA02	#
	up	ASMMA02	#
890	low	ASMMA02	#
	up	ASMMA02	#
...			

Note:

Name a default variable created by SAS to contain item names;
Col1 a default variable created by SAS to contain item scores.

Table 3

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=;

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;
```

Table 4

Number of items resulting in higher average scores at the lower grade level

Country	Number of Items
Australia	4
Austria	6
Canada	1
Cyprus	5
Czech	4
England	5
Greece	7
Hong Kong	9
Hungary	5
Iceland	6
Iran	12
Ireland	3
Japan	3
Korea	16
Latvia	7
Netherlands	3
New Zealand	5
Norway	2
Portugal	7
Scotland	8
Singapore	9
Slovenia	5
Thailand	15
U.S.	2

TM034834

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

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

Author(s): Jianjun Wang

Corporate Source: American Educational Research Association, Chicago, IL

Publication Date: April, 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
☑	□	□
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature:	Printed Name/Position/Title: Jianjun Wang / Professor	
Organization/Address: California State University, Bakersfield	Telephone: (661) 664-3048	FAX: (661) 664-2016
	E-Mail Address: [illegible]	Date: [illegible]