The International Association
for the Evaluation of
Educational Achievement

The Second IEA International Research Conference: Proceedings of the IRC-2006

VOLUME ONE

Proceedings of the IEA IRC-2006 (Vol. 1)

The Second IEA International Research Conference: Proceedings of the IRC-2006

Volume 1: Trends in International Mathematics and Science Study (TIMSS)

The International Association
for the Evaluation of
Educational Achievement

ISBN: 978-90-806643-5-7

Copies of the *Proceedings of the IEA IRC-2006*, *Volumes 1 and 2*, can be obtained from:
IEA Secretariat
Herengracht 487
1017 BT Amsterdam, the Netherlands
Telephone: +31 20 625 3625
Fax: + 31 20 420 7136
Email: Department@IEA.nl
Website: www.iea.nl

Printing: Excel Digital Print, Freestyle Artworks, Wellington, New Zealand
Editing: Paula Wagemaker, Christchurch, New Zealand
Design and layout: Becky Bliss Design and Production, Wellington, New Zealand

Table of contents

Foreword

As part of its mission, the International Association for the Evaluation of Educational Achievement is committed to the development of the community of researchers who work in the area of assessment both nationally and internationally. The association also has a commitment to provide policymakers with the types of data and analyses that will further their understanding of student achievement and the antecedent factors that are implicated in student learning.

As part of a larger strategy to achieve these broad goals, the IEA sponsors a research conference every two years as a means of providing opportunities for new researchers and more experienced scholars to meet, discuss, and present the findings of their work as it relates to the secondary analysis of IEA studies. The proceedings of the Second IEA International Research Conference, which was held in Washington DC, November 2006, and hosted by the Brookings Institution, are published here in two volumes.

The papers in Volume 1 of the proceedings have as their central focus the Trends in Mathematics and Science Study (TIMSS). Volume 2 brings together papers that focus on the Progress in International Reading Literacy Study (PIRLS), the Second Information on Technology in Education Study (SITES), and the Civic Education Study (CivEd).

IEA is grateful to everyone who participated in this conference and hopes that the papers provided here will interest those who work in the various areas of educational research represented in these pages.

We look forward to future contributions to our conferences, and hope that these papers not only contribute to our understanding of educational achievement but also lead to the development of the community of researchers involved in international and national assessment.

Hans Wagemaker PhD
EXECUTIVE DIRECTOR, INTERNATIONAL ASSOCIATION FOR THE
EVALUATION OF EDUCATIONAL ACHIEVEMENT

Effects of science beliefs and instructional strategies on achievement of students in the United States and Korea: Results from the TIMSS 2003 assessment

J. Daniel House
Northern Illinois University
DeKalb, Illinios, USA

Abstract

Several instructional strategies have been designed to improve student achievement in science. In addition, longitudinal research shows that student beliefs are significant predictors of science achievement. The purpose of this study was to use data from the Trends in International Mathematics and Science Study 2003 (TIMSS 2003) assessment to identify relationships between the science achievement of students in the United States and Korea, the classroom instructional strategies they experienced, and the beliefs they held about their learning in science. Because of the complex sampling design of the TIMSS 2003 assessment, jackknife variance estimation procedures using replicate weights were used to compute appropriate standard errors for each variable in this study. Multiple regression procedures were used to simultaneously assess the relative contribution of each science belief variable and instructional strategy toward the explanation of science test scores. There were several significant findings from this study. Frequent use of active learning strategies related positively to achievement test scores for students in both countries. Students who frequently engaged in cooperative learning activities (worked in small groups on an experiment or investigation) tended to earn higher science test scores. Students from both countries who indicated positive self-appraisals of their science ability (usually did well in science and learned things quickly in science) earned higher achievement test scores. Conversely, students who compared themselves negatively to other students (science was more difficult for them than for many of their classmates) tended to earn lower science test scores. The findings also identified significant relationships between several instructional strategies and science achievement, as well as between several science beliefs and science achievement. These results emphasize the importance of simultaneously considering instructional strategies and student beliefs when assessing factors related to science achievement.

Introduction

There is considerable interest in the design of effective instruction for teaching and learning in science. Rillero (2000) observed that early success in science can be facilitated through hands-on experiences that develop an interest in science. Student success in science further develops skills (such as classification, measurement, understanding variables, and analyzing data) that lead to success in many academic subjects. Innovative programs have been developed to provide elementary and secondary school students with hands-on opportunities to learn science concepts and laboratory techniques (Doyle, 1999). A program coordinated by the University of California (Los Angeles) enables high school students and teachers to use integrated science learning and technology activities to improve student problem-solving skills and science knowledge (Palacio-Cayetano, Kanowith-Klein, & Stevens, 1999). In a program conducted by the University of California (San Francisco), medical students visit Grade 6 classrooms and provide instruction on topics related to health and biological sciences (Doyle, 1999).

Two instructional approaches shown to be effective for improving student achievement in science are active learning strategies and cooperative learning activities. Recent findings indicate that the use of active learning materials results in improved science achievement and more positive attitudes toward science (McManus, Dunn, & Denig, 2003). Kovac (1999) similarly found that the use of active learning strategies for a general chemistry course led to improved student achievement, while Lunsford and Herzog (1997) found use of student-centered classroom instructional strategies for a life sciences course resulted in positive student responses and was effective for student performance on standardized exams. Finally, Maheady, Michielli-Pendl, Mallette, and Harper (2002) found the use of active learning strategies for Grade 6 science associated with positive attitudes regarding learning gains and high levels of motivation to succeed in science.

With respect to cooperative learning, results from cross-cultural research indicate that the use of co-operative learning groups for earth science results in higher achievement test scores and more positive attitudes toward science for high school students in Taiwan (Chang & Mao, 1999). Results from meta-analyses of the effects of cooperative learning on science outcomes indicates that a cooperative web-based learning environment designed to foster intrinsic motivation for learning science (Wang & Yang, 2002) is an effective strategy for improving student performance, facilitating more positive attitudes toward science, and promoting persistence into more advanced science and mathematics courses (Bowen, 2000; Springer, Stanne, & Donovan, 1999). This body of work shows that several instructional strategies have the simultaneous goals of improved learning outcomes and increased student motivation for learning science.

A number of studies note that student beliefs are significantly associated with science achievement. Recent findings indicate that the motivational beliefs of high school students are significant predictors of science achievement test scores (Kupermintz & Roeser, 2002). Junior-high school students in Taiwan who expressed more positive attitudes toward science also tended to earn higher science test scores (Tuan, Chin, & Shieh, 2005). House (1996) found specific student beliefs (self-ratings of overall academic ability and drive to achieve, and expectations of graduating with honors) to be significant predictors of grade performance in introductory chemistry. In another study by House (2000a), academic self-concept and achievement expectancies significantly correlated with the grades earned by students in science, engineering, and mathematics disciplines. Singh, Granville, and Dika (2002) found a strong relationship between students' science attitudes and time spent on academic activities and science homework. DeBacker and Nelson (2000) observed that high school students who expressed high academic goals and held a high value for science were also more likely to show higher science achievement. Taken together, these results emphasize the importance of considering student beliefs when assessing factors related to science achievement outcomes.

Results from international assessments indicate that students in Korea typically score above international averages (Martin et al., 2000). Consequently, there has been a continuing interest in identifying factors related to science achievement for students in Korea. Commentators observe that education is an important part of Korean society and that parents play a critical role in emphasizing academic success (Ellinger & Beckham, 1997; Sorenson, 1994). In addition, several studies have identified classroom practices associated with science achievement in Korea. Results from an analysis of high school chemistry classrooms in Korea indicated that students need to understand theoretical models in order to facilitate reflective thinking and to incorporate new learning material (Cho, Park, & Choi, 2000). In a study by Lee and Fraser (2000), high school students in Korea reported the development of constructivist strategies in their high school classes through the use of cooperative learning activities and relevant materials. Korean high school students who expressed specific views about science tended to show lower self-efficacy toward science and tended to be passive learners (Park & Choi, 2000). Results from a case study of science instruction in Korea (Oh, 2005) indicated that teachers employed three primary roles during class sessions: presenting science knowledge to the students through various activities, coaching to enhance science achievement, and scaffolding. There is also evidence that the use of cooperative learning activities results in higher achievement levels and more positive attitudes toward science for middle school students in Korea (Chung & Son, 2000). Consequently, there is evidence of significant relationships between effective teaching strategies for science and the achievement outcomes of students in Korea.

Research studies drawing on data from the TIMSS assessments show significant relationships between science outcomes and various student characteristics and instructional strategies. With respect to science test scores, results from an analysis conducted by House (2000b) of data relating to TIMSS 1995 students from Hong Kong found significant relationships between certain classroom strategies and science achievement. Students who earned higher test scores were those who reported frequently doing experiments or practical investigations in class and working together in pairs or small groups. Similarly, students in Japan who earned higher science test scores indicated that they frequently used active learning strategies during their science lessons and discussed practical or story problems related to everyday life when learning new science topics (House, 2002). Findings from the TIMSS 1999 assessment indicated that students in Japan, Hong Kong, and Chinese Taipei who earned the higher test scores were also those students who frequently used things from everyday life when solving science problems and did experiments or practical investigations in class (House, 2005). Elementary school students in Japan who frequently did experiments in class, worked together in pairs or small groups, and used computers during science lessons also tended to earn higher science test scores (House, 2006a). In addition, results from the TIMSS science performance assessment indicate that student performance on both procedural and higher-order thinking items contributes to achievement (Harmon, 1999).

Other studies have examined the importance of student beliefs for influencing achievement. Results for students in Australia revealed significant relationships between attitudes and aspirations and achievement outcomes (Webster & Fisher, 2000). House's examinations of TIMSS data for students in Ireland and Hong Kong found that students who indicated they enjoyed learning science tended to earn higher test scores while students who attributed success in science at school to external factors (such as good luck) were more likely to earn lower science test scores (House, 2000c, 2003). Finally, an analysis of Grade 8 students from Cyprus indicated that teaching practices exerted a significant influence on attitudes toward science (Papanastasiou, 2002).

The purpose of this study was to use data from the TIMSS 2003 assessment to simultaneously identify relationships between the science achievement of adolescent students in the United States and Korea, the classroom instructional strategies they experienced in relation to science, and their beliefs about their learning of this subject. Data relating to students from these countries were examined for two reasons. First, students from Korea have shown high levels of science achievement on previous international assessments (Martin et al., 2000). Second, previous cross-cultural research has examined instructional strategies related to achievement for students from these countries, and this study provided opportunity to add to this body of work.

Method

The TIMSS 2003 assessment

The TIMSS 2003 assessment examined target populations that were the two adjacent grades containing the largest proportions of nine-year-old and 13-year-old students. Student assessments were conducted during the spring of the 2002/2003 school year. A matrix sampling procedure was used to compile test items into booklets because of the large number of science and mathematics test items on the assessment (Martin & Mullis, 2004). Eight test booklets were developed, and six blocks of items were included in each booklet (Smith Neidorf & Garden, 2004). Representative samples of students took each part of the assessment. The intention of the TIMSS 2003 assessment was to measure student performance on both mathematics and science at the Grade 4 and Grade 8 levels.

Several procedures were used to select the schools within the Korean and the United States samples. For the sample of schools from Korea, initial stratifications were made according to province and urban status (large city, middle, rural). Further stratification was made by student gender in the schools (boys, girls, mixed). Remote schools, special education schools, and sports schools were excluded from the sampling. This procedure resulted in a total of 151 schools in the sample. However, 149 schools actually participated in the TIMSS 2003 assessment. With regard to the United States sample, stratifications were made by school type (public/private) and region. Schools at the Grade 8 level were also stratified by minority status (more than 15% minority students/less than 15% minority students). This procedure resulted in 301 schools in the sample and 232 schools in the TIMSS 2003 assessment.

Students

The students included in these analyses were from the TIMSS 2003 Population 2 samples (13-year-olds) from the United States and from Korea. Of these students, 8,093 from the United States and 5,076 from Korea completed all of the measures regarding self-belief variables and instructional practices examined in this study.

Measures

As part of the TIMSS 2003 assessment, students were given a questionnaire that collected various data, including information regarding student beliefs about science and mathematics, classroom instructional activities, family characteristics, learning resources, out-of-school activities, and science achievement.

This present study examined the effects of several classroom instructional activities on science achievement. Students indicated how frequently the following activities happened during their science lessons. *"How often do you do these things in your science lessons?"*

1. We watch the teacher demonstrate an experiment or investigation
2. We formulate hypotheses or predictions to be tested
3. We design or plan an experiment or investigation
4. We conduct an experiment or investigation
5. We work in small groups on an experiment or investigation
6. We write explanations about what was observed and why it happened
7. We relate what we are learning in science to our daily lives
8. We review our homework
9. We listen to the teacher give a lecture-style presentation
10. We work problems on our own
11. We begin our homework in class.

For these items, the original codings were transformed so that the following values were used to indicate the frequency of each activity: (1) never, (2) some lessons, (3) about half the lessons, (4) every or almost every lesson.

Six specific measures were examined with respect to student beliefs about science:

1. I usually do well in science
2. Science is more difficult for me than for many of my classmates
3. I enjoy learning science
4. Sometimes when I do not initially understand a new topic in science, I know that I will never really understand it
5. Science is not one of my strengths
6. I learn things quickly in science.

For these items, original codings were transformed so that the following levels of agreement were indicated: (1) disagree a lot, (2) disagree a little, (3) agree a little, (4) agree a lot.

The dependent measure examined in this study was each student's science score on the TIMSS 2003 assessment. Because students in the TIMSS 2003 assessment were given relatively few test items in each specific content area, statistical procedures were developed to estimate student proficiency by generating plausible values for each student based on responses given (Gonzalez, Galia, & Li, 2004). Each plausible value provides an estimate of the performance of each student had they actually taken all possible items on the assessment. Five plausible score values were computed for each student because of error in the generation of these imputed proficiency values (Gonzalez et al., 2004). To provide consistency with the statistical procedures used for computing each national average score for mathematics achievement, the dependent measure used in this study was the average of the five plausible values generated for each student on the TIMSS 2003 science assessment.

Procedure

Statistical procedures applied to data collected using simple random sampling are inappropriate for data collected from assessments using complex sampling designs (Foy & Joncas, 2004). One potential problem of using statistical procedures for simple random sampling on data collected from complex sampling designs is the possibility of underestimation of the error (Ross, 1979). Underestimation of error can produce spurious findings of statistical significance in hypothesis testing (Wang & Fan, 1997). Consequently, it is critical when conducting appropriate statistical tests of significance that the design effect is considered and procedures are used that produce unbiased variance estimates.

Because the TIMSS 2003 assessment employed a two-stage stratified cluster sample design, jackknife variance estimation procedures using replicate weights were used to compute appropriate standard errors for each variable included in this study. Brick, Morganstein, & Valliant (2000) found that jackknife variance procedures are an effective method for providing full-sample estimates for data collected from cluster sample designs. This technique simulates repeated sampling of students from the initial sample according to the specific sample design (Johnson & Rust, 1992). Sometimes referred to as a re-sampling plan, the technique produces estimates of the population means and the standard errors of those estimates (Welch, Huffman, & Lawrenz, 1998). An advantage of using the jackknife replication statistic is that it increases the generalizability of research findings because it provides population estimates rather than findings from a single sample (Ang, 1998).

For this study, multiple regression procedures were used to simultaneously assess the relative contribution of each self-belief variable and classroom instructional strategy in explaining the science test scores. In each instance, analyses were conducted separately for the entire sample of students from each country.

Results

Table 1 presents a summary of the results from the multiple regression analysis of relationships between science beliefs, classroom instructional strategies, and achievement test scores for students in Korea.

Five science belief variables significantly entered the multiple regression equation. Students who earned the higher test scores tended to indicate that they learned things quickly in science and usually did well in science. Interestingly, students who expressed negative self-appraisals of their ability to learn new science topics ("Sometimes when I do not initially understand a new topic in science, I know that I will never really understand it") tended to earn higher science achievement test scores. Conversely, students who earned the lower science test scores tended to report that science was not one of their strengths. Similarly, students who earned lower test scores were the students most likely to express negative comparisons of their science ability relative to the ability of other students ("Science is more difficult for me than for many of my classmates").

With respect to instructional strategies, students who earned the higher test scores were those most likely to report that they frequently worked problems on their own and that they related what they were learning in science to their daily lives. Students who said they frequently listened to the teacher give a lecture-style presentation also tended to earn higher science test scores. Frequent use of cooperative learning strategies ("We work in small groups on an experiment or investigation") was positively associated with science test scores. Three instructional strategies showed significant negative relationships with science test scores. Students who earned lower test scores reported that they frequently conducted an experiment or investigation and watched the teacher demonstrate an experiment or investigation. Students who reported higher amounts of class time spent reviewing homework also tended to earn the lower test scores.

The overall multiple regression equation that assessed the joint significance of the complete set of science belief variables and classroom instructional strategies was significant ($F(17,59) = 70.53$, $p < .001$) and explained 31.0% of the variance in science achievement test scores for adolescent students in Korea.

Findings from the multiple regression analysis of relationships between science beliefs, classroom instructional strategies, and science achievement test scores for students in the United States are presented in Table 2. Five science belief variables and seven instructional strategies significantly entered the multiple regression equation. Students who earned the higher test scores tended to indicate they learned things quickly in science and usually did well in science. Conversely, students who earned lower science test scores were more likely to report that science was not one of their strengths. However, students who reported that they enjoyed learning science actually earned lower test scores. In addition, students who earned lower test scores expressed lower self-appraisals of their ability to learn new science topics ("Sometimes when I do not initially understand a new topic in science, I know that I will never really understand it").

With respect to classroom instructional strategies, students who earned higher test scores reported that they more frequently conducted an experiment or investigation in class and worked problems on their own. Students who earned higher test scores also reported that they frequently engaged in cooperative learning activities (in small groups on an experiment or investigation). Four classroom instructional strategies showed significant negative relationships with science test scores. For instance, those students who more frequently watched the teacher demonstrate an experiment or investigation and who formulated hypotheses or predictions to be tested earned lower test scores. Similarly, students who earned lower test scores reported that they frequently designed or planned an

Table 1: Relationships between Science Beliefs, Classroom Instructional Strategies, and Science Achievement Test Scores (Korea)

Self-belief/Instructional activity	Parameter estimate	Standard errors of estimate	*Z*-score
Science beliefs			
I usually do well in science	27.197	2.131	12.76**
Science is more difficult for me than for many of my classmates	-5.576	1.509	-3.69**
I enjoy learning science	-1.446	1.429	-1.01
Sometimes when I do not initially understand a new topic in science, I know that I will never understand it	4.523	1.355	3.34**
Science is not one of my strengths	-4.792	1.617	-2.96**
I learn things quickly in science	5.961	1.638	3.64**
Instructional strategies			
We watch the teacher demonstrate an experiment or investigation	-7.525	1.198	-6.28**
We formulate hypotheses or predictions to be tested	0.779	1.624	0.48
We design or plan an experiment or investigation	0.039	1.811	0.02
We conduct an experiment or investigation	-5.178	1.866	-2.78**
We work in small groups on an experiment or investigation	7.921	1.313	6.03**
We write explanations about what was observed and why it happened	1.973	1.223	1.61
We relate what we are learning in science to our daily lives	4.603	1.248	3.69**
We review our homework	-5.436	1.261	-4.31**
We listen to the teacher give a lecture-style presentation	8.526	1.385	6.16**
We work problems on our own	15.373	1.199	12.82**
We begin our homework in class	-1.730	1.408	-1.23

Note: **p < .01.

experiment or investigation and related what they were learning in science to their daily lives.

The overall multiple regression equation that assessed the joint significance of the complete set of science beliefs variables and classroom instructional strategies was significant ($F(17,59) = 30.32$, $p < .001$) and explained 14.6% of the variance in science achievement test scores for adolescent students in the United States.

Discussion

Several significant findings emerged from this study. A number of specific science beliefs were significantly associated with science achievement test scores for students in the United States and Korea. Students from both countries who expressed positive beliefs about their science abilities (usually did well in science and learned things quickly in science) also tended to earn higher science test scores. Students from both countries who held negative appraisals of their science ability (considered science was not one of their strengths) were more likely to earn lower achievement test scores. Differences between students in the United States and Korea were also noted for relationships between science beliefs and test scores. Students in the United States who indicated they enjoyed learning science actually earned lower test

Table 2: Relationships between Science Beliefs, Classroom Instructional Strategies, and Science Achievement Test Scores (United States)

Self-belief/Instructional activity	Parameter estimate	Standard errors of estimate	*Z*-score
Science beliefs			
I usually do well in science	13.075	1.920	6.81**
Science is more difficult for me than for many of my classmates	-2.188	1.381	-1.58
I enjoy learning science	-4.280	1.544	-2.77*
Sometimes when I do not initially understand a new topic in science, I know that I will never understand it	-11.008	1.520	-7.24**
Science is not one of my strengths	-5.595	1.387	-4.04**
I learn things quickly in science	8.167	1.908	4.28**
Instructional strategies			
We watch the teacher demonstrate an experiment or investigation	-4.076	1.719	-2.37*
We formulate hypotheses or predictions to be tested	-4.362	1.791	-2.44*
We design or plan an experiment or investigation	-12.330	1.633	-7.55**
We conduct an experiment or investigation	14.379	1.948	7.38**
We work in small groups on an experiment or investigation	6.960	1.995	3.49**
We write explanations about what was observed and why it happened	-1.366	1.514	-0.90
We relate what we are learning in science to our daily lives	-4.607	1.373	-3.36**
We review our homework	-0.025	1.478	-0.02
We listen to the teacher give a lecture-style presentation	1.826	1.500	1.22

Note: $^{**}p < .01$; $^{*}p < .05$.

scores, but this relationship was not significant for students in Korea. Students in the United States who expressed negative appraisals of their ability to learn new science information ("Sometimes when I do not initially understand a new topic in science, I know that I will never really understand it") tended to earn lower science test scores; the same relationship was positive for students in Korea. However, the results of this study indicate that student beliefs related significantly to science test scores and need to be considered when assessing factors associated with science achievement.

With respect to classroom instructional strategies, a number of specific activities related significantly to science achievement for students in the United States and Korea. More frequent use of cooperative learning activities (students working in small groups on an experiment or investigation) was positively associated with science achievement test scores for students from both countries. Similarly, active learning strategies (students working problems on their own) were positively related to science achievement for students from both countries. Conversely, students from both countries who reported that they frequently were passive learners during their science lessons (they watched the teacher demonstrate an experiment or investigation) tended to earn lower science test scores.

Several differences between students in the United States and Korea were also found in terms of relationships between instructional strategies and science achievement. For instance, students from the United States who earned lower science test scores reported that they frequently formulated hypotheses or predictions to be tested and designed or planned an experiment or investigation; these relationships were not significant for students from Korea. Students from the United States who reported frequently conducting experiments or investigations during science lessons also tended to earn higher test scores; the same relationship was negative for students from Korea. Students from Korea who reported frequently making real-world connections to their science material (they related what they were learning in science to their daily lives) also tended to earn higher science test scores. However, this relationship was negative for students in the United States. These findings indicate that the assessment of classroom instructional strategies is critical for understanding student outcomes. These results also suggest cross-cultural similarities and differences in the relationship between instructional activities and science achievement.

Several of the findings from this study are consistent with the results of previous research results. One important group of findings in the study concerned the significant relationships between a number of student self-beliefs and science achievement—relationships that held even after the effects of classroom instruction had been taken into account. For instance, students who expressed positive beliefs about learning science also tended to earn higher science test scores. House (2006b) similarly found, even after considering the effects of several types of instructional strategies, significant relationships between several mathematics beliefs of adolescent students in Japan and their algebra achievement. Results from a developmental analysis of the relationship between academic self-concept and achievement showed a significant causal connection between the self-beliefs of elementary school students and teacher ratings of performance (Guay, Marsh, & Biovin, 2003). Earlier work by House (1994) and by Vallerand, Fortier, and Guay (1997) showed beliefs to be significant predictors of withdrawal from high school and of grade performance in science courses. Beliefs are thus an important factor in determining student achievement outcomes.

Another important finding from this study was that more frequent use of cooperative learning activities related positively to science achievement for students in cross-cultural settings (in this case, the United States and Korea). Other research studies also demonstrate the effectiveness of cooperative learning strategies for science learning across cultures. Yu (1998) found that Grade 5 students who engaged in cooperative learning while using computers had positive attitudes toward science. The use of a cooperative learning program for elementary school students similarly produced improved student motivation and higher achievement outcomes (Janes, Koutsoppanagos, Mason, & Villaranda, 2000). A cooperative learning experience for high school students increased students' motivation to achieve in science at higher grade levels (Pearson, 1989).

The results of this present study identified several significant relationships between science beliefs, instructional practices, and science achievement for students from cross-cultural contexts where high levels of science achievement have been noted (see Martin et al., 2000, in this regard). In addition, the results extend previous related research because they are based on simultaneous examination of the effects of student self-beliefs and multiple instructional strategies on the science achievement of large national samples of students who were part of a comprehensive international assessment. The findings also provide several directions for further research. For example, would similar findings be observed for students from other countries that participated in the TIMSS 2003 assessment, and what are the unique effects of specific instructional strategies on student self-beliefs? In addition, further qualitative studies are needed to identify the specific facets of active learning strategies and cooperative learning activities that positively contribute to student achievement.

References

Ang, R. P. (1998). Use of the jackknife statistic to evaluate result replicability. *Journal of General Psychology, 125*, 218–228.

Bowen, C. W. (2000). A quantitative literature review of cooperative learning effects on high school and college chemistry achievement. *Journal of Chemical Education, 77*, 116–119.

Brick, J. M., Morganstein, D., & Valliant, R. (2000). *Analysis of complex sample data using replication*. Rockville, MD: Westat.

Chang, C. Y., & Mao, S. L. (1999). Comparison of Taiwan science students' outcomes with inquiry-group versus traditional instruction. *Journal of Educational Research, 92*, 340–346.

Cho, I. Y., Park, H. J., & Choi, B. S. (2000). *Conceptual types of Korean high school students and their influences on learning style*. Paper presented at the National Association for Research in Science Teaching annual meeting, New Orleans, LA.

Chung, Y. L., & Son, D. H. (2000). Effects of cooperative learning strategy on achievement and science learning attitudes in middle school biology. *Journal of the Korean Association for Research in Science Education, 20*, 611–623.

DeBacker, T. K., & Nelson, R. M. (2000). Motivation to learn science: Differences related to gender, class type, and ability. *Journal of Educational Research, 93*, 245–254.

Doyle, H. J. (1999). UCSF partnership to enrich science teaching for sixth graders in San Francisco's schools. *Academic Medicine, 74*, 329–331.

Ellinger, T. R., & Beckham, G. M. (1997). South Korea: Placing education on top of the family agenda. *Phi Delta Kappan, 78*, 624–625.

Foy, P., & Joncas, M. (2004). TIMSS 2003 sampling design. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 108–123). Chestnut Hill, MA: International Study Center, Boston College.

Gonzalez, E. J., Galia, J., & Li, I. (2004). Scaling methods and procedures for the TIMSS 2003 mathematics and science scales. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 252–273). Chestnut Hill, MA: International Study Center, Boston College.

Guay, F., Marsh, H. W., & Biovin, M. (2003). Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of Educational Psychology, 95*, 124–126.

Harmon, M. (1999). Performance assessment in the Third International Mathematics and Science Study: An international perspective. *Studies in Educational Evaluation, 25*, 243–262.

House, J. D. (1994). Student motivation and achievement in college chemistry. *International Journal of Instructional Media, 21*, 1–11.

House, J. D. (1996). Student expectancies and academic self-concept as predictors of science achievement. *Journal of Psychology, 130*, 679–681.

House, J. D. (2000a). Academic background and self-beliefs as predictors of student grade performance in science, engineering, and mathematics. *International Journal of Instructional Media, 27*, 207–220.

House, J. D. (2000b). Relationships between instructional activities and science achievement of adolescent students in Hong Kong: Findings from the Third International Mathematics and Science Study (TIMSS). *International Journal of Instructional Media, 27*, 275–288.

House, J. D. (2000c). Student self-beliefs and science achievement in Ireland: Findings from the Third International Mathematics and Science Study (TIMSS). *International Journal of Instructional Media, 27*, 107–115.

House, J. D. (2002). Relationships between instructional activities and science achievement of adolescent students in Japan: Findings from the Third International Mathematics and Science Study (TIMSS). *International Journal of Instructional Media, 29*, 275–288.

House, J. D. (2003). Self-beliefs and science and mathematics achievement of adolescent students in Hong Kong: Findings from the Third International Mathematics and Science Study (TIMSS). *International Journal of Instructional Media, 30*, 195–212.

House, J. D. (2005). Classroom instruction and science achievement in Japan, Hong Kong, and Chinese Taipei: Results from the TIMSS 1999 assessment. *International Journal of Instructional Media, 32*, 295–311.

House, J. D. (2006a). The effects of classroom instructional strategies on science achievement of elementary-school students in Japan: Findings from the Third International Mathematics and Science Study (TIMSS). *International Journal of Instructional Media, 33*, 217–229.

House, J. D. (2006b). Mathematics beliefs, instructional strategies, and algebra achievement of adolescent students in Japan: Results from the TIMSS 1999 assessment. *International Journal of Instructional Media, 33*, 443–462.

Janes, L. M., Koutsoppanagos, C. L., Mason, D. S., & Villaranda, I. (2000). *Improving student motivation through the use of engaged learning, cooperative learning and multiple intelligences*. Master's Action Research Project, St. Xavier University and SkyLight Field-Based Master's Program. (ERIC Document Reproduction Service No. ED443559)

Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17*, 175–190.

Kovac, J. (1999). Student active learning methods in general chemistry. *Journal of Chemical Education, 76*, 120–124.

Kupermintz, H., & Roeser, R. (2002). *Another look at cognitive abilities and motivational processes in science achievement: A multidimensional approach to achievement validation* (Center for the Study of Evaluation Technical Report No. 571). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Lee, S. S. U., & Fraser, B. J. (2000). *The constructivist learning environment of science classrooms in Korea*. Paper presented at the Australasian Science Education Research Association annual meeting, Fremantle, Western Australia.

Lunsford, B. E., & Herzog, M. J. R. (1997). Active learning in anatomy and physiology. *American Biology Teacher, 59*, 80–84.

Maheady, L., Michielli-Pendl, J., Mallette, B., & Harper, G. F. (2002). A collaborative research project to improve the academic performance of a diverse sixth grade science class. *Teacher Education and Special Education, 25*, 55–70.

Martin, M. O., & Mullis, I. V. S. (2004). Overview of TIMSS 2003. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 2–21). Chestnut Hill, MA: International Study Center, Boston College.

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: International Study Center, Boston College.

McManus, D. O., Dunn, R., & Denig, S. J. (2003). Effects of traditional lecture versus teacher-constructed and student-constructed self-teaching instructional resources on short-term science achievement and attitudes. *American Biology Teacher, 65*, 93–102.

Oh, S. (2005). Discursive roles of the teacher during class sessions for students presenting their science investigation. *International Journal of Science Education, 27*, 1825–1851.

Palacio-Cayetano, J., Kanowith-Klein, S., & Stevens, R. (1999). UCLA's outreach program of science education in the Los Angeles schools. *Academic Medicine, 74*, 348–351.

Papanastasiou, C. (2002). School, teaching and family influences on student attitudes toward science: Based on TIMSS data for Cyprus. *Studies in Educational Evaluation, 28*, 71–86.

Park, H. J., & Choi, B. S. (2000). *Korean high school students' views about learning and knowing science*. Paper presented at the National Association for Science Teaching annual meeting, New Orleans, LA.

Pearson, R. W. (1989). *Cooperative learning in an urban science classroom*. (ERIC Document Reproduction Service No. ED312225)

Rillero, P. (2000). Exploring science with young children. *Scholastic Early Childhood Today, 14*(5), 11–12.

Ross, K. N. (1979). An empirical investigation of sampling errors in educational survey research. *Journal of Educational Statistics, 4*, 24–40.

Singh, K., Granville, M., & Dika, S. (2002). Mathematics and science achievement: Effects of motivation, interest, and academic engagement. *Journal of Educational Research, 95*, 323–332.

Smith Neidorf, T., & Garden, R. (2004). Developing the TIMSS 2003 mathematics and science assessment and scoring guides. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 22–65). Chestnut Hill, MA: International Study Center, Boston College.

Sorensen, C. W. (1994). Success and education in South Korea. *Comparative Education Review, 38*, 10–35.

Springer, S., Stanne, M. E., & Donovan, S. S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of Educational Research, 69*, 21–51.

Tuan, H. L., Chin, C. C., & Shieh, S. H. (2005). The development of a questionnaire to measure students' motivation towards science learning. *International Journal of Science Education, 27*, 639–654.

Vallerand, R. J., Fortier, M. S., & Guay, F. (1997). Self-determination and persistence in a real-life setting: Toward a motivational model of high school dropout. *Journal of Personality and Social Psychology, 72*, 1161–1176.

Wang, L., & Fan, X. (1997). *The effect of cluster sampling design in survey research on the standard error statistic*. Paper presented at the American Educational Research Association annual meeting, Chicago, IL.

Wang, S. K., & Yang, C.C. (2002). *An investigation of a web-based learning environment designed to enhance the motivation and achievement of students in learning difficult mental models in high school science*. Paper presented at the ED-MEDIA 2002 World Conference on Educational Multimedia, Hypermedia, and Telecommunications, Denver, CO.

Webster, B. J., & Fisher, D. L. (2000). Accounting for variation in science and mathematics achievement: A multilevel analysis of Australian data, Third International Mathematics and Science Study (TIMSS). *School Effectiveness and School Improvement, 11*, 339–360.

Welch, W. W., Huffman, D., & Lawrenz, F. (1998). The precision of data obtained in large-scale science assessments: An investigation of bootstrapping and half-sample replication methods. *Journal of Research in Science Teaching, 35*, 697–704.

Yu, F. (1998). The effects of cooperation with inter-group competition on performance and attitudes in a computer-assisted science instruction. *Journal of Computers in Mathematics and Science Teaching, 17*, 381–395.

Relationships between student and instructional factors and algebra achievement of students in the United States and Japan: An analysis of TIMSS 2003 data

J. Daniel House
Northern Illinois University
DeKalb, Illinois, USA

James A. Telese
University of Texas
Brownsville, Texas, USA

Abstract

Algebra knowledge is a critical part of middle school mathematics achievement. Success in algebra is necessary for taking higher-level mathematics courses and leads to higher scores on standardized tests. The purpose of this study was to use data from the Trends in International Mathematics and Science Study 2003 (TIMSS 2003) assessment to identify relationships between the algebra achievement of adolescent students in the United States and Japan, the beliefs these students held about their learning in this subject, and the classroom instructional strategies they experienced in relation to it. Jackknife variance estimation procedures using replicate weights were used to compute appropriate standard errors for each variable in this study. Multiple regression procedures were used to simultaneously examine the relative contribution of each instructional activity and mathematics belief variable toward explaining the explanation of algebra achievement test scores. Students from both the United States and Japan who earned higher algebra test scores were more likely to indicate positive beliefs about their mathematical ability (they learned things quickly in mathematics and usually did well in mathematics). Students who earned lower algebra test scores compared themselves negatively to other students. With instructional practices, those students from both countries who had frequent opportunity to work problems on their own tended to earn higher algebra test scores. The study also found the mathematics beliefs of the United States and Japanese students and the classroom instructional practices they experienced to be significantly related to algebra achievement. Cross-cultural similarities and differences were also noted for these relationships. These results have implications for mathematics instruction and identify strategies to improve algebra achievement.

Introduction

There is increasing interest in identifying factors associated with mathematics achievement. Many career options are open only to students who have mastered mathematical skills and enrolled in advanced mathematics courses (House, 1993). Algebra knowledge is a critical part of middle school mathematics achievement. Student success in algebra is necessary for taking higher-level mathematics courses and leads to higher scores on standardized tests (Catsambis, 1994; Telese, 2000). Instructional strategies such as the use of appropriate problem-solving activities have been used to foster student achievement in algebra. An algebra curriculum centered on problem-solving and incorporating real-world applications provides students with opportunities to succeed in algebra. Farrell and Farmer (1998) have identified eight modes of instruction that can be applied to algebra teaching: lecture, question/answer, discussion, demonstration, laboratory, individual student projects, supervised practice, and technological activities. Research findings indicate that instruction in introductory algebra should include activities that incorporate the use of informal knowledge, application to real-world settings, and applications of mathematical thinking (Telese, 2000). These studies highlight the importance of examining student and instructional factors related to algebra achievement in order to improve opportunities for success in mathematics.

Several studies have examined the relationship between student beliefs and academic achievement. A longitudinal study by House (1997), for example, found the initial self-beliefs of a sample of Asian-

American students to be significant predictors of their grade performance. Marsh and Teung (1997) found that the academic self-concept of adolescent students exerts significant causal effects on their mathematics achievement. Similarly, earlier studies by House (1993, 1995) with a group of older adolescent students found their academic self-concept to be significantly related to higher mathematics course grades. Another study by House (2001a), this time with American Indian/Alaska Native students, found significant correlations between specific facets of these students' academic self-concept and achievement expectancies (self-ratings of overall academic ability and mathematical ability, and expectations of making at least a B average in college) and their mathematics achievement. Results from a longitudinal study of middle school students in Germany indicated that students who expressed higher initial levels of interest in mathematics were those students most likely to subsequently enroll in advanced mathematics courses (Koller, Baumert, & Schnabel, 2001). Significant relationships have also been found between the mathematics self-efficacy and the subsequent mathematics achievement of middle school students (Pajeres & Graham, 1999). Research findings for middle school students also indicate that students who express more focused learning goals tend to have a higher mathematics self-concept (Anderman & Young, 1993).

Examinations of cross-cultural differences in the relationship between student beliefs and mathematics achievement indicate the importance of considering student beliefs when assessing factors that influence mathematics achievement. A recent study by Ercikan, McCreith, and Lapointe (2005) found that self-confidence in mathematics was the factor most strongly associated with mathematics achievement for students in Norway and Canada, but not for students in the United States. Another study, by Tsao (2004), found Grade 5 students in Taiwan tended to have more positive beliefs about mathematics than did students in the United States.

Considerable interest has been directed toward the mathematics achievement of students in Japan, and research has examined instructional strategies used for mathematics teaching and learning. For instance, the Learner's Perspective Study (LPS), involving an analysis of mathematics classrooms in nine countries, found that students in Japan discuss their strategies for solving problems set during the lesson and make presentations to the rest of the class (Shimizu, 2002). Sawada (1999) notes that mathematics instruction in Japan focuses on the development of problem-solving strategies, with entire class sessions oriented toward a single problem. An observational analysis of classrooms conducted by Stigler, Lee, and Stevenson (1987) found a significantly high proportion of class time spent on mathematics instruction in classrooms in Japan, and more time spent on other activities such as classroom management in United States classrooms. Becker, Silver, Kantowski, Travers, and Wilson (1990) note that teachers in Japanese classrooms tend to present multiple strategies for solving mathematics problems, while Kroll and Yabe (1987) describe teaching strategies used in Japan that incorporate manipulative materials designed to help students develop flexible thinking about methods for solving mathematics problems. These strategies have led to observations that elementary school students in Japan explain solutions to problems in ways that incorporate more complex mathematical concepts (Silver, Leung, & Cai, 1995). Finally, according to Perry (2000), teachers in Japanese classrooms provide more extended explanations to their students. The results of these studies highlight cultural differences in mathematics classroom practices and problem-solving strategies.

Several studies have successfully used data from the TIMSS assessments to identify student and instructional factors associated with the mathematics outcomes of students in the United States and Japan. For example, results from the TIMSS Videotape Classroom Study indicate that students in Japan spend a considerable amount of time during mathematics lessons developing solutions to problems and examining a single problem (Stigler, Gallimore, & Hiebert, 2000). Japanese students also are likely during mathematics lessons to present alternative strategies for solving mathematics problems and to cover advanced mathematical content (Shimizu, 1999; Stigler, Gonzales, Kawanaka, Knoll, & Serrano, 1999). A case study of a geometry lesson in Japan as part of the TIMSS 1995 Videotape Classroom Study found incorporation during computer-based mathematics teaching of instructional activities enhanced student attention and interest (House, 2002).

Findings from the TIMSS 1995 assessment found significant associations between specific instructional activities and the mathematics achievement of students in Japan. Those students who earned the higher test scores were also those students frequently assigned homework, who used things from everyday life when solving mathematics problems, and who tried to solve problems related to new mathematics topics when learning new material (House, 2001b). The TIMSS 1999 assessment also found that the students in Japan who earned the higher mathematics test scores were those who tended to frequently receive homework (House, 2004). Conversely, those students who frequently spent time during mathematics lessons checking one another's homework and/or having the teacher check homework tended to earn lower test scores. Telese (2004) found that students in the United States who frequently used calculators during mathematics lessons also showed higher algebra test scores.

Research from the TIMSS assessments also highlights instructional practices associated with interest in learning mathematics for students in Japan. For example, students who expressed enjoyment when learning mathematics were the students most likely to report discussing practical or story problems related to everyday life, working on mathematics projects, and engaging in cooperative learning (working together in pairs or small groups on problems or projects) during their mathematics lessons (House, 2003, 2005). Students in Japan who attributed success in mathematics to controllable factors (hard work studying at home and memorizing the textbook or notes) tended to gain higher test scores while the students who attributed success to external factors (good luck) tended to show lower achievement levels (House, 2006a).

The purpose of this present study was to use data from the TIMSS 2003 assessment to simultaneously identify relationships between the algebra achievement of adolescent students in the United States and Japan, their beliefs about their learning of algebra, and the classroom instructional strategies they experienced in relation to this subject. Data relating to students from these two countries were examined for two reasons. First, students from Japan have scored above international averages on previous mathematics assessments (Kelly, Mullis, & Martin, 2000). Second, previous cross-cultural studies have examined factors associated with mathematics achievement for students in these two countries, and this study provided opportunity to add to this body of work.

Method

The TIMSS 2003 assessment

The TIMSS 2003 assessment examined target populations that were the two adjacent grades containing the largest proportions of nine-year-old and 13-year-old students. Student assessments were conducted during the spring of the 2002/2003 school year. A matrix sampling procedure was used to compile test items into booklets because of the large number of science and mathematics test items on the assessment (Martin & Mullis, 2004). Eight test booklets were developed and six blocks of items were included in each booklet (Smith Neidorf & Garden, 2004). Representative samples of students took each part of the assessment. The intention of the TIMSS 2003 assessment was to measure student performance on both mathematics and science at the Grade 4 and Grade 8 levels.

Several procedures were used to select the schools within the Japanese and the United States samples. For the sample of students from Japan, initial stratifications were made in order to exclude schools for educable mentally disabled students and functionally disabled students. Further stratification was made by level of urbanization (big city area, city area, and non-city area). This procedure resulted in a total sample of 150 schools, all of which participated in the TIMSS 2003 assessment. For the United States sample, stratifications were made by school type (public/private) and region. Schools at the Grade 8 level were also stratified by minority status (more than 15% minority students/less than 15% minority students). This procedure resulted in 301 schools in the sample, of which 232 schools participated in the TIMSS 2003 assessment.

Students

The students included in these analyses were from the TIMSS 2003 Population 2 samples (13-year-olds) from the United States and Japan. Of these students, 4,244 from Japan and 7,862 students from the United States completed all of the measures regarding classroom instructional strategies and mathematics beliefs examined in this study.

Measures

As part of the TIMSS 2003 assessment, students were given a questionnaire that collected various data, including information regarding student beliefs about science and mathematics, classroom instructional activities, family characteristics, learning resources, out-of-school activities, and science and mathematics achievement.

This present study examined the influence of several mathematics beliefs on mathematics achievement. The items included in these analyses were:

1. I usually do well in mathematics
2. I would like to take more mathematics in school
3. Mathematics is more difficult for me than for many of my classmates
4. I enjoy learning mathematics
5. Sometimes when I do not initially understand a new topic in mathematics, I know that I will never really understand it
6. Mathematics is not one of my strengths
7. I learn things quickly in mathematics
8. I think learning mathematics will help me in my daily life
9. I need mathematics to learn other school subjects
10. I need to do well in mathematics to get into the university of my choice
11. I would like a job that involved using mathematics
12. I need to do well in mathematics to get the job I want.

For these items, the original codings were transformed so that the following levels of student agreement were indicated: (1) disagree a lot, (2) disagree a little, (3) agree a little, or (4) agree a lot.

With respect to classroom instructional activities, students indicated how frequently the following strategies were used in their mathematics lessons:

1. We practice adding, subtracting, multiplying, and dividing without using a calculator
2. We work on fractions and decimals
3. We interpret data in tables, charts, or graphs
4. We write equations and functions to represent relationships
5. We work together in small groups
6. We relate what we are learning in mathematics to our daily lives
7. We explain our answers
8. We decide on our own procedures for solving complex problems
9. We review our homework
10. We listen to the teacher give a lecture-style presentation
11. We work problems on our own
12. We begin our homework in class
13. We have a quiz or test
14. We use calculators.

For each of these items, the original codings were transformed so that the following values were used to indicate the frequency of each activity: (1) never, (2) some lessons, (3) about half the lessons, (4) every or almost every lesson.

The dependent measure examined in this study was each student's algebra score on the TIMSS 2003 assessment. Because students in the TIMSS 2003 assessment were given relatively few test items in each specific content area, statistical procedures were developed to estimate student proficiency by generating plausible values for each student based on responses given (Gonzalez, Galia, & Li, 2004). Each plausible value provides an estimate of the performance of each student had they actually taken all possible items on the assessment. Five plausible score values were computed for each student because of error in the generation of these imputed proficiency values (Gonzalez et al., 2004). To provide consistency with the statistical procedures used for computing each national average score for mathematics achievement, the dependent measure used in this study was the average of the five plausible values generated for each student on the TIMSS 2003 algebra assessment.

Procedure

Statistical procedures applied to data collected using simple random sampling are inappropriate for data collected from assessments using complex sampling designs (Foy & Joncas, 2004). One potential problem of using statistical procedures for simple random sampling on data collected from complex sampling designs is the possibility of underestimation of the error (Ross, 1979). Underestimation of error can produce spurious findings of statistical significance in hypothesis testing (Wang & Fan, 1997). Consequently, it is critical when conducting appropriate statistical tests of significance that the design effect is considered and procedures are used that produce unbiased variance estimates.

Because the TIMSS 2003 assessment employed a two-stage stratified cluster sample design, jackknife variance estimation procedures using replicate weights were used to compute appropriate standard errors for each variable included in this study. Brick, Morganstein, and Valliant (2000) found that jackknife variance procedures are an effective method for providing full-sample estimates for data collected from cluster sample designs. This technique simulates repeated sampling of students from the initial sample according to the specific sample design (Johnson & Rust, 1992). Sometimes referred to as a re-sampling plan, this technique produces estimates of the population means and the standard errors of those estimates (Welch, Huffman, & Lawrenz, 1998). An advantage of using the jackknife replication statistic is that this method increases the generalizability of research findings because it provides population estimates rather than findings from a single sample (Ang, 1998).

For this study, multiple regression procedures were used to simultaneously assess the relative contribution of each self-belief variable and classroom instructional strategy in explaining the algebra test scores. In each instance, analyses were conducted separately for the entire sample of students from each country.

Results

Table 1 presents a summary of the results from the multiple regression analysis of relationships between mathematics beliefs, classroom instructional strategies, and algebra test scores for students in Japan.

Seven mathematics belief variables significantly entered the multiple regression equation. Students who earned the higher algebra test scores were those most likely to indicate they usually did well in mathematics and enjoyed learning mathematics. Similarly, students who indicated they learned things quickly in mathematics also tended to earn higher algebra test scores. Students who earned higher test scores were also the students most likely to indicate they needed to do well in mathematics to gain entry to the university of their choice. Conversely, students who earned the lower algebra test scores tended to report that mathematics was not one of their strengths. In addition, students who expressed negative comparisons of themselves in terms of their mathematics ability relative to the ability of other students ("Mathematics is more difficult for me than for many of my classmates") tended to earn the lower test scores. Students who reported that they would like to take more mathematics in school actually earned lower algebra test scores.

Seven instructional strategies also significantly entered the multiple regression equation. Students who earned higher test scores were those who reported that they spent time practicing mathematical operations (adding, subtracting, multiplying, and dividing) without using a calculator. Students who showed higher algebra test scores also reported that they frequently decided on their own procedures for solving complex problems and explained their answers during mathematics lessons. Similarly, students who said they frequently worked problems on their own tended to earn the higher algebra test scores.

Three instructional strategies showed significant negative relationships with algebra test scores. Frequent use of cooperative learning activities (students working together in small groups) was negatively related to algebra achievement. Students who showed lower algebra test scores also reported frequently relating what they were learning in mathematics to their daily lives. In addition, students who reported that they frequently used calculators during mathematics lessons tended to earn lower test scores.

The overall multiple regression equation that assessed the joint significance of the complete set of mathematics beliefs and instructional strategies was significant ($F(26,49) = 64.84$, $p < .001$) and explained 33.7% of the variance in algebra test scores for adolescent students in Japan.

Findings from the multiple regression analysis of relationships between mathematics beliefs, classroom instructional strategies, and algebra test scores for students in the United States are summarized in Table 2.

Ten mathematics belief variables significantly entered the multiple regression equation. Students who showed the higher algebra test scores were those students most likely to indicate that they usually did well in mathematics and learned things quickly in mathematics. Students who reported that they would like a job that involved using mathematics and that they needed to do well in mathematics to get into the university of their choice also earned higher test scores. Conversely, students who earned the lower algebra test scores were the students most likely to report that mathematics was not one of their strengths, and students who expressed negative comparisons of themselves relative to other students ("Mathematics is more difficult for me than for many of my classmates") tended to obtain lower algebra test scores. Similarly, students who expressed negative self-appraisals of their ability to learn new material ("Sometimes when I do

Table 1: Relationships between Mathematics Beliefs, Classroom Instructional Strategies, and Algebra Test Scores (Japan)

Self-belief/Instructional activity	Parameter estimate	Standard errors of estimate	Z-score
Mathematics beliefs			
I usually do well in mathematics	25.359	2.300	11.03**
I would like to take more mathematics in school	-5.275	2.123	-2.48*
Mathematics is more difficult for me than for many of my classmates	-2.977	1.482	-2.01*
I enjoy learning mathematics	7.377	1.842	4.01**
Sometimes when I do not initially understand a new topic in mathematics, I know that I will never understand it	0.404	1.169	0.34
Mathematics is not one of my strengths	-5.600	1.685	-3.32**
I learn things quickly in mathematics	9.987	2.535	3.94**
I think mathematics will help in my daily life	-3.700	1.952	-1.89
I need mathematics to learn other school subjects	1.123	1.959	0.57
I need to do well in mathematics to get into the university of my choice	5.126	1.423	3.60**
I would like a job that involved using mathematics	1.008	2.134	0.47
I need to do well in mathematics to get the job I want	1.971	1.526	1.29
Instructional strategies			
We practice adding, subtracting, multiplying, and dividing without using a calculator	11.323	1.413	8.01**
We work on fractions and decimals	1.199	1.672	0.72
We interpret data in tables, charts, or graphs	-1.612	2.460	-0.65
We write equations and functions to represent relationships	-0.260	1.743	-0.15
We work together in small groups	-6.500	2.191	-2.97**
We relate what we are learning in mathematics to our daily lives	-8.923	2.453	-3.64**
We explain our answers	5.166	1.648	3.13**
We decide on our own procedures for solving complex problems	4.038	1.656	2.44*
We review our homework	-0.388	1.830	-0.21
We listen to the teacher give a lecture-style presentation	3.420	2.928	1.17
We work problems on our own	18.063	1.772	10.19**
We begin our homework in class	3.469	2.052	1.69
We have a quiz or test	-3.078	1.833	-1.68
We use calculators	-15.904	2.394	-6.64**

Note: $^{**}p < .01$; $^{*}p < .05$.

Table 2: Relationships between Mathematics Beliefs, Classroom Instructional Strategies, and Algebra Test Scores (United States)

Self-belief/Instructional activity	Parameter estimate	Standard errors of estimate	Z-score
Mathematics beliefs			
I usually do well in mathematics	14.840	1.450	10.23**
I would like to take more mathematics in school	1.544	1.237	1.26
Mathematics is more difficult for me than for many of my classmates	-5.329	1.136	-4.69**
I enjoy learning mathematics	-6.093	1.325	-4.60**
Sometimes when I do not initially understand a new topic in mathematics, I know that I will never understand it	-9.369	0.980	-9.56**
Mathematics is not one of my strengths	-9.384	1.329	-7.40**
I learn things quickly in mathematics	5.036	1.290	3.90**
I think mathematics will help in my daily life	-11.844	1.348	-8.79**
I need mathematics to learn other school subjects	-0.394	1.384	-0.28
I need to do well in mathematics to get into the university of my choice	12.115	1.718	7.05**
I would like a job that involved using mathematics	6.092	1.409	4.32**
I need to do well in mathematics to get the job I want	-6.423	1.192	-5.39**
Instructional strategies			
We practice adding, subtracting, multiplying, and dividing without using a calculator	-0.263	1.067	-0.25
We work on fractions and decimals	-3.305	1.652	-2.00*
We interpret data in tables, charts, or graphs	-6.820	1.567	-4.35**
We write equations and functions to represent relationships	18.309	1.443	12.69**
We work together in small groups	-3.411	1.638	-2.08*
We relate what we are learning in mathematics to our daily lives	-10.582	1.154	-9.17**
We explain our answers	-1.635	1.553	-1.05
We decide on our own procedures for solving complex problems	-1.241	1.251	-0.99
We review our homework	12.708	1.530	8.31**
We listen to the teacher give a lecture-style presentation	-2.243	0.912	-2.46*
We work problems on our own	6.693	1.543	4.34**
We begin our homework in class	1.597	1.641	0.97
We have a quiz or test	-8.253	1.402	-5.89**
We use calculators	7.845	1.648	4.76**

Note: ***p* < .01; * *p* < .05.

not initially understand a new topic in mathematics, I know that I will never really understand it") also obtained lower test scores. Interestingly, students who earned lower algebra test scores also reported that they enjoyed learning mathematics, needed to do well in mathematics to get the job they wanted, and thought learning mathematics would help in their daily lives.

Ten instructional strategies also significantly entered the multiple regression equation. Students who earned the higher test scores reported that they frequently wrote equations and functions to represent the relationships during their mathematics lessons. Students who showed higher algebra achievement also indicated that they frequently reviewed their homework and used calculators during mathematics lessons. Students who said they often worked problems on their own during mathematics class also tended to earn higher algebra test scores.

Six instructional strategies showed significant negative relationships with algebra achievement. For instance, students who earned the lower test scores reported that they frequently worked on fractions and decimals and had a quiz or test during class. Similarly, students who indicated that they frequently engaged in cooperative learning activities (worked together in small groups) and interpreted data in tables, charts, and/or graphs also tended to earn lower algebra test scores. In addition, students who showed lower algebra achievement indicated they frequently listened to the teacher give a lecture-style presentation. These students also said they related what they were learning in mathematics to their daily lives.

The overall multiple regression equation that assessed the joint significance of the complete set of mathematics beliefs and instructional strategies was significant ($F(26,50) = 63.15$, $p < .001$) and explained 35.3% of the variance in algebra test scores for adolescent students in the United States.

Discussion

Several significant findings emerged from this study. For instance, a number of mathematics beliefs were significantly associated with algebra achievement for students in Japan and the United States. Students from both countries who indicated that they learned things quickly in mathematics and usually did well in mathematics tended to be those students who earned the higher algebra test scores. Students from both countries who earned higher test scores also indicated that they needed to do well in mathematics to get into the university of their choice. Conversely, students from both countries who earned the lower algebra test scores were more likely than students who earned higher scores to report that mathematics was not one of their strengths. Further, students from both countries who compared their ability in mathematics negatively with the ability of other students ("Mathematics is more difficult for me than for many of my classmates") showed lower test scores.

Differences were also noted for students from the United States and Japan. Students in Japan who reported they enjoyed learning mathematics tended to earn higher algebra test scores, while a negative relationship was found for students in the United States. In addition, students from the United States who expressed negative self-appraisals of their ability to learn new mathematics information ("Sometimes when I do not initially understand a new topic in mathematics, I know that I will never really understand it") tended to earn lower algebra test scores; the same relationship was not significant for students in Japan.

Several classroom instructional strategies were significantly associated with algebra achievement for students in both countries. Similarities and differences between students in the United States and in Japan were also found for these relationships. In regard to similarities, students from both countries who reported frequently working problems on their own during mathematics lessons also earned higher algebra test scores. Conversely, students from both countries who indicated they frequently engaged in cooperative learning activities (working together in small groups) also tended to earn lower test scores. Similarly, students from both countries who frequently related what they were learning in mathematics to their daily lives earned lower algebra test scores.

Several differences in the relationship between instructional practices and algebra test scores also were noted for students in the United States and Japan. For instance, the more often students in Japan practiced basic mathematical operations (adding, subtracting, multiplying, and dividing) without using a calculator, they more likely they were to earn higher algebra test scores; the same relationship was not significant for students in the United States. In addition, students in Japan who frequently explained their answers during mathematics lessons earned higher test scores. This relationship was not significant for students in the United States. However, students in the United States who earned higher algebra test scores reported that they frequently wrote equations and functions to represent relationships. This association was not significant for students in Japan. Also, those students in the United States who frequently reviewed their homework during mathematics lessons were those who earned higher algebra test scores; the same relationship was not significant for students in Japan.

Finally, students from the United States and Japan showed opposite trends for the relationship between calculator use during mathematics lessons and algebra test scores. Those students in the United States who reported frequently using calculators during mathematics lessons were also those more likely to earn higher test scores. Conversely, students in Japan who more frequently used calculators tended to earn lower test scores.

Several of the relationships between self-beliefs and mathematics achievement found in this study are consistent with results from previous research. For example, House (1993) and Wheat, Tunnell, and Munday (1991) found significant relationships between academic self-concept and algebra course grades. Ma (2001) reported that students' future expectations of success exerted a significant influence on enrollment in advanced mathematics. Similarly, House (2000) found achievement expectancies and academic self-concept to be significant predictors of student achievement in science, engineering, and mathematics. Also of interest is that the relationship between self-concept and mathematics achievement appears to become stronger as students reach higher class levels in school (Ma & Kishor, 1997). With respect to findings from the TIMSS assessments, Hammouri (2004) reported that student beliefs exerted significant direct effects on the mathematics test scores of students in Jordan. These findings emphasize the importance of considering student beliefs when assessing factors related to mathematics achievement.

The results of the present study, in concert with the findings of other research, also have implications for mathematics teaching practices and the learning of mathematics. For instance, in the present study, students from both countries who reported frequent use of active learning ("We work problems on our own") also tended to earn higher algebra test scores. Various classroom strategies exist that have the aim of developing mathematical connections, particularly through use of concrete models and practical examples. Work by Crocker and Long (2002) and Hines (2002) focuses on the use of practical exercises and physical models to teach students elementary functions and exponents. Smith (1999) notes that giving students the opportunity to reflect after engaging in active learning experiences helps them integrate the information with their existing knowledge.

A second application of the results of the present study is selection of learning examples to foster positive attitudes toward mathematics. The results of this study show significant relationships between self-beliefs and algebra test scores for students in both countries. Boyer (2002) proposes several classroom activities that develop students' self-management of their learning and improve their motivation for learning mathematics. Reimer and Moyer's (2005) assessment of the effectiveness of using computer-based manipulatives for Grade 3 mathematics indicated improvement in students' enjoyment for learning mathematics. Mathematics teachers have also reported that the use of hands-on projects during class results in increased motivation and participation (DeGeorge & Santoro, 2004). Similarly, the use of learner-centered classrooms appears to improve intrinsic motivation for mathematics learning (Heuser, 2000).

The significant associations found in this present study between the algebra achievement of students in the United States and Japan, the beliefs these students held about mathematics, and the classroom instructional strategies they experienced provide several directions for further research. For instance, longitudinal research is needed to assess the effects of mathematics beliefs and instructional practices on other measures of student achievement, such as enrollment in advanced mathematics courses in high school. Findings from a recent study conducted by House (2005) show mathematics beliefs (perceptions that mathematics is an easy subject and students reporting that they enjoy learning mathematics) significantly related to "fractions" and "number sense" scores for adolescent students in Japan. Research is also needed to determine whether the effects of the self-beliefs and instructional activities noted in the present study hold for other measures of student achievement. For example, the United States and Japanese students in the present study who frequently used cooperative learning strategies were significantly more likely than were the other students in the two samples to have a low level of achievement in algebra. This finding differs from more recent results that show a positive significant relationship between frequent use of cooperative learning activities in science lessons and science tests scores for elementary-school students in Japan (House, 2006b). Further study therefore is needed to clarify the relationship between cooperative learning and algebra achievement for students in the United States and Japan. Finally, further research is needed to determine if similar findings would be observed for students from other countries that participated in the TIMSS 2003 assessment. Despite this need for further investigation, the results from this study have usefully extended previous research by providing a simultaneous assessment of the effects of several specific mathematics beliefs and classroom instructional strategies on the algebra achievement of large national samples of students in cross-cultural settings.

References

Anderman, E. M., & Young, A. J. (1993). *A multilevel model of adolescents' motivation and strategy use in academic domains.* Paper presented at the American Educational Research Association annual meeting, Atlanta, GA.

Ang, R. P. (1998). Use of the jackknife statistic to evaluate result replicability. *Journal of General Psychology, 125*, 218–228.

Becker, J. P., Silver, E. A., Kantowski, M. G., Travers, K. J., & Wilson, J.W. (1990). Some observations of mathematics teaching in Japanese elementary and junior high schools. *Arithmetic Teacher, 38*(2), 12–21.

Boyer, K. R. (2002). Using active learning strategies to motivate students. *Mathematics Teaching in the Middle School, 8*, 48–51.

Brick, J. M., Morganstein, D., & Valliant, R. (2000). *Analysis of complex sample data using replication.* Rockville, MD: Westat.

Catsambis, S. (1994). The path to math: Gender and racial-ethnic differences in mathematics participation from middle school to high school. *Sociology of Education, 67*, 199–225.

Crocker, D. A., & Long, B. B. (2002). Rice + technology = an exponential experience. *Mathematics Teaching in the Middle School, 7*, 404–407.

DeGeorge, B., & Santoro, A. M. (2004). Manipulatives: A hands-on approach to math. *Principal, 84*(2), 28.

Ercikan, K., McCreith, T., & Lapointe, V. (2005). Factors associated with mathematics achievement and participation in advanced mathematics courses: An examination of gender differences from an international perspective. *School Science and Mathematics, 105*, 5–14.

Farrell, M., & Farmer, W. (1998). *Secondary mathematics teaching: An integrated approach.* Needham, MA: Janson Publications.

Foy, P., & Joncas, M. (2004). TIMSS 2003 sampling design. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 108–123). Chestnut Hill, MA: International Study Center, Boston College.

Gonzalez, E. J., Galia, J., & Li, I. (2004). Scaling methods and procedures for the TIMSS 2003 mathematics and science scales. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 252–273). Chestnut Hill, MA: International Study Center, Boston College.

Hammouri, H. A. M. (2004). Attitudinal and motivational variables related to mathematics achievement in Jordan: Findings from the Third International Mathematics and Science Study. *Educational Research, 46*, 214–257.

Heuser, D. (2000). Mathematics workshop: Mathematics class becomes learner centered. *Teaching Children Mathematics, 6*, 288–295.

Hines, E. (2002). Exploring functions with dynamic physical models. *Mathematics Teaching in the Middle School*, 7, 274–278.

House, J. D. (1993). Achievement-related expectancies, academic self-concept, and mathematics achievement of academically under-prepared adolescent students. *Journal of Genetic Psychology, 154*, 61–71.

House, J. D. (1995). The predictive relationship between academic self-concept, achievement expectancies, and grade performance in college calculus. *Journal of Social Psychology, 135*, 111–112.

House, J. D. (1997). The relationship between self-beliefs, academic background, and achievement of adolescent Asian-American students. *Child Study Journal, 27*, 95–110.

House, J. D. (2000). Academic background and self-beliefs as predictors of student grade performance in science, engineering, and mathematics. *International Journal of Instructional Media, 27*, 207–220.

House, J. D. (2001a). Predictive relationships between self-beliefs and mathematics achievement of American Indian/Alaska Native students. *International Journal of Instructional Media, 28*, 287–297.

House, J. D. (2001b). Relationships between instructional activities and mathematics achievement of adolescent students in Japan: Findings from the Third International Mathematics and Science Study (TIMSS). *International Journal of Instructional Media, 28*, 93–105.

House, J. D. (2002). The use of computers in a mathematics lesson in Japan: A case analysis from the TIMSS Videotape Classroom Study. *International Journal of Instructional Media, 29*, 113–124.

House, J. D. (2003). The motivational effects of specific instructional strategies and computer use for mathematics learning in Japan: Findings from the Third International Mathematics and Science Study (TIMSS). *International Journal of Instructional Media, 30*, 77–95.

House, J. D. (2004). The effects of homework activities and teaching strategies for new mathematics topics on achievement of adolescent students in Japan: Results from the TIMSS 1999 assessment. *International Journal of Instructional Media, 31*, 199–210.

House, J. D. (2005). Motivational qualities of instructional strategies and computer use for mathematics teaching in the United States and Japan: Results from the TIMSS 1999 assessment. *International Journal of Instructional Media, 32*, 89–104.

House, J. D. (2006a). Mathematics beliefs and achievement of elementary school students in Japan and the United States: Results from the Third International Mathematics and Science Study. *Journal of Genetic Psychology, 167*, 31–45.

House, J. D. (2006b). The effects of classroom instructional strategies on science achievement of elementary-school students in Japan: Findings from the Third International Mathematics and Science Study. *International Journal of Instructional Media, 33*, 217–229.

Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17*, 175–190.

Kelly, D. L., Mullis, I. V. S., & Martin, M. O. (2000). *Profiles of student achievement in mathematics at the TIMSS international benchmarks: U.S. performance and standards in an international context.* Chestnut Hill, MA: International Study Center, Boston College.

Koller, O., Baumert, J., & Schnabel, K. (2001). Does interest matter? The relationship between academic interest and achievement in mathematics. *Journal for Research in Mathematics Education, 32*, 448–470.

Kroll, D. L., & Yabe, T. (1987). A Japanese educator's perspective on teaching mathematics in the elementary school. *Arithmetic Teacher, 35*(2), 36–43.

Ma, X. (2001). Participation in advanced mathematics: Do expectation and influence of students, peers, teachers, and parents matter? *Contemporary Educational Psychology, 26*, 132–146.

Ma, X., & Kishor, N. (1997). Attitude toward self, social factors, and achievement in mathematics: A meta-analytic review. *Educational Psychology Review, 9*, 89–120.

Marsh, H. W., & Teung, A. S. (1997). Causal effects of academic self-concept on academic achievement: Structural equation models of longitudinal data. *Journal of Educational Psychology, 89*, 41–54.

Martin, M. O., & Mullis, I. V. S. (2004). Overview of TIMSS 2003. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 2–21). Chestnut Hill, MA: International Study Center, Boston College.

Pajares, F., & Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. *Contemporary Educational Psychology, 24*, 124–139.

Perry, M. (2000). Explanations of mathematical concepts in Japanese, Chinese, and U.S. first- and fifth-grade classrooms. *Cognition and Instruction, 18*, 181–207.

Reimer, K., & Moyer, P. S. (2005). Third-graders learn about fractions using virtual manipulatives: A classroom study. *Journal of Computers in Mathematics and Science Teaching, 24*, 5–25.

Ross, K. N. (1979). An empirical investigation of sampling errors in educational survey research. *Journal of Educational Statistics, 4*, 24–40.

Sawada, D. (1999). Mathematics as problem-solving: A Japanese way. *Teaching Children Mathematics, 6*(1), 54–58.

Shimizu, Y. (1999). Studying sample lessons rather than one excellent lesson: A Japanese perspective on the TIMSS Videotape Classroom Study. *International Reviews on Mathematical Education, 31*, 190–194.

Shimizu, Y. (2002). *Capturing the structure of Japanese mathematics lessons: Some findings of the international comparative studies.* Paper presented at the ICMI Second East Asia Regional Conference on Mathematics Education and Ninth Southeast Asian Conference on Mathematics Education, Nanyang Technological University, Singapore.

Silver, E. A., Leung, S. S., & Cai, J. (1995). Generating multiple solutions for a problem: A comparison of the responses of US and Japanese students. *Educational Studies in Mathematics, 28*, 35–54.

Smith, J. (1999). Active learning of mathematics. *Mathematics Teaching in the Middle School, 5*, 108–110.

Smith Neidorf, T., & Garden, R. (2004). Developing the TIMSS 2003 mathematics and science assessment and scoring guides. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 22–65). Chestnut Hill, MA: International Study Center, Boston College.

Stigler, J. W., Gallimore, R., & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from the TIMSS video studies. *Educational Psychologist, 35*, 87–100.

Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *Findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States.* Washington, DC: National Center for Education Statistics Report No. NCES99-074.

Stigler, J. W., Lee, S.Y., & Stevenson, H. W. (1987). Mathematics classrooms in Japan, Taiwan, and the United States. *Child Development, 58*, 1272–1285.

Telese, J. A. (2000). *School algebra reform: Meeting the grade?* Paper presented at the American Educational Research Association annual meeting, New Orleans, LA.

Telese, J.A. (2004). Middle school mathematics classroom practices and achievement: A TIMSS-R analysis. *Focus on Learning Problems in Mathematics, 26*(4), 19–30.

Tsao, Y. L. (2004). A comparison of American and Taiwanese students: Their math perception. *Journal of Instructional Psychology, 31*, 206–213.

Wang, L., & Fan, X. (1997). *The effect of cluster sampling design in survey research on the standard error statistic.* Paper presented at the American Educational Research Association annual meeting, Chicago, IL.

Welch, W. W., Huffman, D., & Lawrenz, F. (1998). The precision of data obtained in large-scale science assessments: An investigation of bootstrapping and half-sample replication methods. *Journal of Research in Science Teaching, 35*, 697–704.

Wheat, J., Tunnell, J., & Munday, R. (1991). Predicting success in college algebra: Student attitude and prior achievement. *College Student Journal, 25*, 240–244.

Capturing the dynamics that led to the narrowing achievement gap between Hebrew-speaking and Arabic-speaking schools in Israel: Findings from TIMSS 1999 and 2003

Ruth Zuzovsky
School of Education, Tel Aviv University
Tel Aviv, Israel

Abstract
Closing achievement gaps between sub-populations in Israel, and amongst them between students in the Hebrew-speaking schools and the Arabic-speaking schools, continues to be one of the priorities of the country's education system. TIMSS 2003 findings provide first evidence that efforts made during the 1990s to close these gaps were in the right direction and that, although inequality in input between the two sectors still remains, gaps in learning outcomes have narrowed. This paper highlights the dynamics that led to the narrowing achievement gap. Instead of focusing on school conditions at one point in time (Zuzovsky, 2005), this paper looks for changes that occurred from 1999 to 2003 in certain school/class-level variables in the two ethnic sectors and relates them to the uneven increase in achievement of these sectors.

Background

Closing achievement gaps between sub-populations in Israel, especially those related to socioeconomic, ethnic, or gender factors, continues to be one of the most prioritized goals of the national education system. The first two acts of education were enacted primarily to ensure that all students in Israel, without any discrimination on the basis of race, religion, or gender, would have the right to education (Compulsory Education Act 1949) and that the state would provide equal educational opportunities regardless of political or any other organizational affiliation (State Education Act 1953). These legislative acts resulted in a centralistic educational policy that issued in a strategy of equity among schools regarding both school inputs (e.g., a unified curriculum, equal numbers of students per class, equal weekly learning hours per grade level, equal number of learning days per year, similar teaching methods, etc.) and school outputs (e.g., equal percentage of students entitled to a matriculation certificate, equal levels of achievement, etc.).

Toward the end of the 1960s, it became clear that this policy had not had the expected results, as large gaps in educational outcomes between students from different ethnic origins, usually also associated with socioeconomic status (SES), were persistently evident. The discourse on inequality in Israel's education system in the 1960s, and even in the 1970s, focused mainly on inequality within the Jewish sector, between students of Sephardi and Ashkenazi[1] origin or between high- and low-SES students, and ignored the inequality between the Hebrew- and Arabic-speaking student population groups. The Arab population of Israel has what amounts to a separate education system within the larger Israeli one, in which students, teachers, and principals are all Arab citizens of Israel and the language of instruction is Arabic, not Hebrew. Only rarely during the 1960s and 1970s did policymakers and researchers deal with inequality between the Arab and Jewish populations in Israel (Mari, 1978; Mari & Dahir, 1978; Peled, 1976).

The failure of the equity policy, as it was practiced within the Jewish sector, gave way to a policy of differential treatment and affirmative action to schools hosting "disadvantaged" students, which included financial benefits, differential curricula, special entrance thresholds on to higher education institutions, special placement tests, and graded tuition fees for students from low-income families (Gaziel, Elazar, & Marom, 1993). This affirmative policy, introduced in the late

[1] *Sephardi:* A Jewish person of Spanish or Portuguese origin, now used loosely to refer to any Jewish person who is not of Northern and Eastern European (*Ashkenazi)* descent (also known as "Mizrachi").

1960s, was implemented at a much later stage in Arabic-speaking than in Hebrew-speaking schools.

During the 1980s, the general director of the Ministry of Education appointed several committees that generated plans to improve the education system in the Arab sectors (1981, 1985, mentioned in Gaziel, Elazar, & Marom, 1993). Their critique concerning discrimination (Al Haj, 1995; Bashi, Kahan, & Davis, 1981; Kahan & Yelnik, 2000; Mazawi, 1996; Shavit, 1990; Zarzure, 1995) led the Ministry of Education in the 1990s to announce two five-year plans for the Arabic-speaking sector that were mainly affirmative in nature. The first one started in the early 1990s and the second was launched in 1999. These projects aimed to improve all aspects of educational activity by increasing the number of students entitled to matriculation certificates, reducing the percentage of drop-outs, adding study hours, increasing auxiliary staff in schools, enhancing science and technology education, promoting special education services, and providing professional support for teachers and principals. The plans also included construction and development of school buildings.

The affirmative five-year plans in the Arab sector resulted in several improvements. For instance, from 1990 to 2001, enrolment rates of 14- to 17-year-olds in the upper elementary schools increased in the Arab sector by 26% compared to only 6% in the Jewish sector. More study time was allocated to Arabic-speaking schools, with an increase in the average hours per class and the average hours per student at all school levels, but especially at the upper secondary level, and more so in the Arab sector than in the Jewish sector (Sprinzak, Bar, Levi-Mazloum, & Piterman, 2003).

Despite these improvements, inequalities between the two education systems in both inputs and outputs continued to appear (Lavi, 1997). An official publication (Sprinzak et al., 2003) on allocation of inputs and on outputs in the Jewish and Arab sectors in the wake of the Third International Mathematics and Science Study (TIMSS) 2003 revealed gaps between the two systems in favor of the former (see Table 1).

Inequalities between the Jewish and Arab population groups go beyond the education system, encompassing other social aspects. According to official sources (Knesset Research and Information Center, 2004), the Arab sector is characterized by larger families, lower levels of parental education, lower income levels, higher ratio of families living below the poverty line, and lower percentage of employment (see Table 2).

The data presented demonstrate the ongoing inequality between the two population groups, which is in line with the persisting achievement gaps between the two populations as intensively reported in recent national and international studies (Aviram, Cafir, & Ben Simon, 1996; Cafir, Aviram, & Ben Simon, 1999; Karmarski & Mevarech, 2004; Sprinzak et al., 2003; Zuzovsky, 2001, 2005).

More than ever, the public education system is called upon to eliminate this inequality. The recent National Task Force for the Advancement of Education (Dovrat Committee, 2005) specifically refers to this inequality in some of its recommendations:

> Arab education shall have full budgetary equality based on uniform, differential per-student funding, like the rest of the educational system. Disparities between Jewish and Arabic schools with regard to buildings and physical infrastructure shall be eliminated in an effort to reduce, as quickly as possible, the disparities in educational achievement, including the percentage of students who graduate from high school and the percentage of

Table 1: Input and Output Indicators in Hebrew and Arabic Education

	Hebrew sector	Arab sector
Average no. of hours per student*	1.97	1.64
No. of students per FTP**	8.6	11.6
Average no. of students per class	26	29
Enrolment rates—ages 14–17 (%)	96	80.5
Percentage of Grade 12 students entitled to matriculation certificates	52.3	45.6
Annual student dropout rate between Grades 9 and 12 (%)	4.9	9.8

Note: * Total school hours/number of students; ** FTP = Full-time teacher's position.
Source: Sprinzak et al. (2003): Tables C9, C11, C23, F1.

Table 2: Demographic Characteristics of Jewish and Arab Population Groups in Israel (2003)

	Hebrew population	Arab population
Percentage of women with upper secondary education	42.0	22.8
Percentage of men with upper secondary education	37.9	23.4
Average no. of children per household	3.13	5.06
Percentage of families below poverty line	17.7	44.7
Percentage of employed persons	57.0	39.0

Source: Knesset Research and Information Center (2004). Background report on school outcomes in the Arab sector presented to the Child Rights Committee in the Knesset.

students eligible for matriculation certificates. (*Dovrat Summary Report*, Dovrat Committee, 2005, p. 29)

Surprisingly, despite the inequalities reported that could justify the continuing achievement gap, findings from the last TIMSS study (2003) indicated not only that there was a significant increase in Israel's mean scores in mathematics and science since 1999 (30 points score gain in mathematics and 20 points in science), but that this increase was much more profound in the Arabic-speaking schools than in the Hebrew-speaking ones (three times more in mathematics (68 score points vs. 23) and five times more in science (64 score points vs. 12)). As a result of this uneven increase, the achievement gap between the two sectors in favor of the Jewish sector narrowed from about one standard deviation in 1999 to 0.5 *SD* in mathematics and 0.45 *SD* in science in 2003 (Zuzovsky, 2005).

This uneven increase was the trigger for a previous study (Zuzovsky, 2005) that aimed to identify the web of effects that narrowed the achievement gap between the two sectors in 2003. Higher frequencies of certain instructional variables also found to be more positively associated with achievement in the Arab sector, and relatively higher schooling effects in less economically developed sectors such as the Arab sector in Israel, explained why students from this sector performed better in 2003.

The present study is a further step in studying the narrowing achievement gap between the two sectors. Rather than looking for school-level variables that have a differential effect on achievement in the two sectors at one point in time, the present work aims to look for changes that occurred in these variables in the two sectors from 1999 to 2003, and to relate them to changes in achievement that occurred in the two sectors during this period.

A possible interaction effect that the two major independent variables in this study—the ethnic affiliation of the school ("sector") and the year of testing ("year")—might have on the variability of the achievement scores was explored and confirmed at an initial phase of the study. Findings from a two-way, between-group analysis of variance, and later findings from a multi-level regressions analysis, revealed a significant interaction effect of *sector*year* on achievement. It remained for this study to look for the variables that affect this interaction effect in a way that favors the achievement gains of students in the Arabic-speaking schools.

In line with this aim, the following question was raised: *Which are the relevant school/class contextual variables whose frequency and/or association with achievement changed differently in the two sectors from 1999 to 2003 in a way that results in higher achievement gains of students in Arabic-speaking schools as compared to the gains of students in Hebrew-speaking schools?*

Method

Answering this question required several steps. The first of these delineated those contextual school-level variables whose frequency changed differently in the two sectors over the years. A series of two-way, between-groups analyses of variance were employed to identify significant joint effects (two-way interaction effects) of *sector* and *year* on the frequency or means of selected school/class contextual variables. The contextual school-level variables that were delineated in this first step served the second step of the analysis, which aimed to look at whether these variables affect the interaction effect *sector* and *year* have on the variability of students' outcomes (looking for a three-way interaction effect).

Acknowledging the hierarchical nature of the education system and the simultaneous, non-additive dependence of achievement on variables acting at each level of this hierarchy, I have adopted a multilevel and interactive approach to study these effects (Aitkin & Zuzovsky, 1994), applying HLM5 (hierarchical linear and non-linear modeling) software (Bryk & Raudenbush, 1992; Raudenbush, Bryk, Cheong, & Congdon, 2000).

As the sampling procedure of TIMSS allowed the sampling of at least one class from each sampled school (as was done in Israel), the school and class-level effects are confounded. Thus, the models that were specified for the analyses were two-level models of students nested in schools/classes. The possibility of considering year of testing as an additional hierarchical level, as was done in the case of cross-sectional studies with repeated observation for schools or students (Aitkin, 1988; Raudenbush, 1989), was rejected, as the sampling design in IEA studies provides neither repeated measures of schools nested in time, nor of students nested within schools over time. The alternative models specified contained, at their first level, five variables describing student characteristics (determined as having fixed effects on the variability of achievement).[2]

The second-level variables included (in all alternative models specified) both sector and year as well as their interaction terms. These two variables were treated as school-level dummy variables. *Year* describes the global functioning and conditions at two points in time: 1999 (0) and 2003 (1), and *sector* indicates whether the school is Arabic-speaking (0) or Hebrew-speaking (1).[3] In addition to these two "constantly appearing" school-level variables, only one of a series of selected school/class-level contextual variables and all its possible two- and three-way interaction terms with the former two predictors were introduced into the model. The reason for employing only one additional school/class-level variable in each model stemmed from computational constraints on the number of model parameters allowed to be included in the regression equation. In interactive models that include several school-level variables and all possible relevant interaction terms between them, there is a limit to the number of the school-level variables allowed.[4] The criteria used to select the school/class contextual variables and the process of selecting them are described further on in this paper.

To avoid problems of multi-colinearity and to maximize interpretability, the school-level contextual predictors were standardized around their grand mean (see Aiken & West, 1991, p. 43). The regression models specified, for each outcome score in an equation format for all HLM analyses, were the following:

Level 1 Model

Y = β_0 + β_1(Number of books in student's home – BSBGBOOK) + β_2(Level of education student aspires to finish – BSBGHFSG) + β_3(Academic education – "fathers" – ACADEM_F) + β_4(Academic education – "mothers" – ACADEM_M) + β_5(Number of people in student's home – BGPLHO1) + R.

Level 2 Model

β_0 = γ_{00} + γ_{01}(Sector) + γ_{02}(Year) + γ_{03}(Sector*Year) + γ_{04}(standardized relevant school variable) + γ_{05}(year*standardized relevant school variable) + γ_{06} (sector*standardized relevant school variable) + γ_{07}(year*sector*standardizedrelevant school variable) + U.

Models that exhibited a significant three-way interaction were then used to predict mathematics and science achievements in the two sectors at two points in time, conditioned on different values of the relevant school variable involved. Three possible values of these variables on their standardized scale were used for these predictions: 1—*minimal value*, 2—*actual mean value* set to 0, and 3—*maximal value*. Plotting the predicted outcomes highlighted the role of these contextual variables in the narrowing of the achievement gap that occurred from 1999 to 2003 between students in Hebrew-speaking schools and students in Arabic-speaking schools.

Data source

The data source that served the analyses was obtained from the TIMSS 1999 and TIMSS 2003 studies in Israel. A total of 4,195 Grade 8 students from 139 schools participated in the study in 1999. Of these students, 3,383 studied in 112 Hebrew-speaking

[2] The regression coefficients of all first-level student variables employed in the analyses were found to have a non-significant random effect over schools; thus, they were specified as having fixed effects.

[3] This distinction has several implications in terms of socioeconomic and cultural differences between the two populations.

[4] As a general principle, models should contain a reduced set of variables by at least an order of magnitude from the number of cases to the number of model parameters. With about 245 schools participating in the 1999 and 2003 studies, the second-level model parameter should not exceed more than 24 regression terms.

schools and 812 in 27 Arabic-speaking schools. In 2003, these numbers were 4,318 Grade 8 students, of which 3,163 studied in 108 Hebrew-speaking schools and 1,098 in 38 Arabic-speaking schools.

Most of the school/class contextual variables were derived from student questionnaires. I chose this data source in order to limit missing data problems. These variables described students' background characteristics and students' perceptions regarding the instructions they received. When these variables are aggregated or averaged at the class level, they provide us with valid class-level measures of student body composition and prevalent modes of instruction. Some data were obtained from principal and teacher questionnaires. As more than one teacher taught the sampled classes, responses of teachers teaching the same class were also averaged.

As the HLM5 software (Raudenbush et al., 2000) assumes a complete set of data, it provides two options for handling missing data problems for Level 1: pairwise and listwise deletion of cases. In this study, pairwise deletion was applied to Level 1 (student) variables. However, at the second level (school/class), in order to use all available data, listwise deletion of cases was needed. At the end of this process, full sets of school-level data relevant to the study were obtained describing about 246 schools when regressing mathematics scores and about 243 schools when regressing science scores.

The regressed outcome scores were estimates of the first plausible score[5] on the mathematics and on the science scales.

Selection of potential explanatory contextual school/class variables

Since this study deals with changes that occurred from 1999 to 2003 in the frequency and effectiveness of certain school/class-level contextual variables, the analyses conducted were limited only to those variables that appeared in both studies and were identical or only slightly adjusted.

This set of variables (about 60 in total), which described students and their home characteristics, student class composition, and instruction and school characteristics, already reflects theoretical as well as empirical input generated in earlier cycles of TIMSS studies. Out of this set of variables, a restricted set—33 variables of theoretical importance and with evidence of being related to achievement in Hebrew-speaking schools or in Arabic-speaking schools—was selected to go through the next screening process, which aimed to look for those variables whose frequency changed differently over time in the two sectors. This process was based mainly on the results of the two-way, between-group analyses of variance where significant joint effects of *sector* and *year* on measures of frequency of several school/class-level variables were detected. Additional data on changes that occurred in the association of these variables with achievement supported the choice of a final set of 17 school/class-level variables that fulfilled both or at least one of the two requirements: showing a significant differential change in their association with science or mathematics achievement over the years in the two sectors and/or showing a significant differential change in their frequency over time in the two sectors. These variables were then fitted, along with *sector* and *year*, into two-level interactive regression models, thus serving the last step in the analyses, aimed at revealing significant three-parameter interaction effects on achievement. The appendix to this paper presents a description of all variables used in the advanced stages of the analyses.

Results

The results are presented below with reference to the different steps of the analysis.

Delineating school-level variables whose frequency changed differently over time in the two sectors

Tables 3 and 4 display combined findings from a two-way analysis of variance of the frequency of school contextual variables (dependent variables) by *sector* and *year* and correlational findings on their association with achievement in the two sectors over the years (Pearson correlations and significance).

In general, the frequency of many school/class variables changed significantly in the two sectors over time. This is indicated by the significant interaction effect of sector*year on their frequency. These significant interactions, when coupled with changes in the association of these variables with achievement, provide us with some clues regarding their role in narrowing the achievement gap between the sectors.

[5] Proficiency score used in IEA studies. For more details, see the TIMSS 2003 technical report.

The increase over time in the percentage of mothers with an academic education (ACADEM_M) (positively associated with achievement) in Arabic-speaking schools and the decline in the number of people living at the student's home (BGPLHO1) (negatively associated with achievement) in Arabic-speaking schools point to an improvement, over the years, of home conditions associated with achievement in the Arab sector.

An increase in the frequency of several modes of mathematics instruction occurred in the Arabic-speaking schools, while in the Hebrew-speaking schools, their frequency either did not change or slightly decreased. This is also the case regarding the opportunities given to students to review their homework (CMHROH1), to work out problems on their own (CMHWPO1), and to relate what is learnt to daily life (CMHMDL1). Reviewing homework (CMHROH1), which was negatively associated with achievement in 1999, turned out, in 2003, to be positively and significantly associated with achievement in both sectors.

Relating what is learnt to daily life (CMHMDL1) and working out problems on student's own (CMHWPO1), which were negatively associated with mathematics achievement in 1999 in both sectors, continued unchanged in Hebrew-speaking schools, but became positively associated with achievement in 2003 only, in the Arabic-speaking schools.

Similarly, in science, from 1999 to 2003, students in Arabic-speaking schools had increased opportunities to practice inquiry-oriented modes of instruction (watch their teachers demonstrate experiments (CSDEMO1), conduct experiments on their own (CSEXPER1), and work in small groups on experiments or investigations (CSSGRP1)). These modes of instruction became less frequent over the years in the Jewish sector. In 2003, they were all slightly, although not significantly, positively associated with achievement in the Arab sector.

Changes that occurred from 1999 to 2003 in the frequency of conducting experiments in the Arabic-speaking schools illustrate this trend. While there was no change in the frequency with which students in Hebrew-speaking schools conducted experiments, this variable did become more frequent in the Arabic-speaking schools. The positive association this mode of instruction had in 1999 with science achievement in Hebrew-speaking schools disappeared in 2003, while in Arabic-speaking schools, it was the negative association of this mode of instruction with science achievement that disappeared. By 2003, frequently conducting experiments in the class was no longer improving Hebrew-speaking students' achievement, and no longer adversely affecting student achievement in Arabic-speaking schools.

The frequency of testing in mathematics (CMHHQT1) and science classes (CSTEST1) was similar in both sectors and remained constant over the years. In 1999, frequent testing was negatively associated with mathematics and science achievement, and it became more so in the Jewish sector in 2003. In the Arab sector, it became less negatively associated with science achievement, and even positively associated (albeit not statistically significantly) with mathematics achievement. Toward 2003, frequent testing seemed to undermine student achievement in Hebrew-speaking schools whereas it harmed students in Arabic-speaking schools less, or even slightly benefited them.

Improvement in school conditions from 1999 to 2003, as indicated by a decrease in the mean of the index of shortages that affect instruction (scale 1—*does not affect*; 4—*affects a lot*) occurred in both sectors, although more so in the Arab sector. In 1999, high levels of these indices (poor resources) when related to general school resources (RESOUR1) (especially in Arabic-speaking schools) or in mathematics resources (RESOUR2) (in both types of schools) were negatively associated with achievement. In 2003, this situation remained evident only in the Hebrew-speaking schools. In the Arabic-speaking schools, shortage of general school resources turned out, in 2003, not to be significantly associated with achievement, while high levels of shortage in mathematics or science resources even became positively associated with achievement. It seems that the improved conditions for science and mathematics learning in Arabic-speaking schools toward 2003 were not necessarily associated with improved achievement.

Changes in two indicators of school climate that limit instruction, both negatively associated with achievement, occurred differentially in the two sectors. The severity of late arrivals at school (BCBGSPO1) (1—*not at all*; 3—*very severe*) decreased from 1999 to 2003 in the two sectors and especially in Arabic-speaking schools. Teachers' assessment regarding

*Table 3: Descriptive Statistics of School/Class Contextual Variables, their Correlation with Mathematics Achievement in 1999 and 2003 in the Two Sectors, and the Interaction Effect of Sector*Year on their Measures*

		Hebrew-speaking schools						Arabic-speaking schools					
School/class variables	*F*-Values & sig. of interaction term *Sector* Year*	1999 *n*=100–112			2003 *n*=100–108			1999 *n*=24–27			2003 *n*=26–38		
		Mean	(*SD*)	Corr.	Mean	(*SD*)	Corr.	Mean	(*SD*)	Corr.	Mean	(*SD*)	Corr.
Student body composition:													
ACADEM_M Acad. edu. mother	4.75*	.43	(.21)	.70***	.40	(.21)	.61***	.12	(.11)	.13	.20	(.11)	.35*
BGPLHO1 No. people at home	17.17***	5.1	(.58)	-.34***	5.2	(.70)	.00	7.0	(.49)	-.32	6.3	(.43)	-.09
CGHFSG Stud. asp. level of edu.	6.46*	4.3	(.33)	.56***	4.0	(.51)	.63***	4.2	(.42)	.25	3.5	(.42)	.44***
Instruction:													
CMHROH1 Reviewing homework	33.3***	3.2	(.39)	-.03	3.1	(.44)	.19*	2.8	(.32)	-.19	3.4	(.27)	.49**
CMHWPO1 Solve problems on own	6.9**	3.2	(.32)	-.20*	3.3	(.26)	-.43***	2.6	(.35)	-.44**	2.9	(.32)	.37
CMHMDL1 Learnt in daily life	2.8	2.7	(.31)	-.32**	2.5	(.28)	-.17	2.6	(.28)	-.43**	2.7	(.31)	.33*
CMHHQT1 Quizzes or tests	0.2	2.8	(.27)	-.18*	2.7	(.31)	-.41***	2.7	(.28)	-.49**	2.7	(.31)	.10
School resources:													
RESOUR1 Short. gen. sch. res.	3.8*	1.7	(.61)	-.03	1.6	(.64)	-.15	2.4	(.77)	-.41*	1.9	(.64)	.05
RESOUR2 Short. math sch. res.	5.1*	2.0	(.84)	-.15	1.8	(.74)	-.14	2.9	(.70)	-.29	2.2	(.73)	.34*
School climate:													
BCBGSPO1 Late school arrival	5.8*	2.2	(.58)	-.10	2.1	(.52)	-.23*	2.2	(.72)	-.50*	1.7	(.53)	-.10
CTMGLTO6 Disrupt. stud. behav.	4.8***	2.0	(.60)	-.11	2.5	(.88)	-.36**	2.3	(.83)	.20	2.3	(.85)	-.18

Note: *$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$.

*Table 4: Descriptive Statistics of School/Class Contextual Variables, their Correlation with Science Achievement in 1999 and 2003 in the Two Sectors, and the Interaction Effect of Sector*Year on their Measures*

School/class variables	*F*-Values & sig. of interaction term *Sector* Year*	Hebrew-speaking schools						Arabic-speaking schools					
		1999 *n*=100–112			2003 *n*=100–108			1999 *n*=24–27			2003 *n*=26–38		
		Mean	(*SD*)	Corr.	Mean	(*SD*)	Corr.	Mean	(*SD*)	Corr.	Mean	(*SD*)	Corr.
Student body composition:													
ACADEM_M Acad. edu. mother	4.75*	.43	(.21)	.69***	.40	(.21)	.63***	.12	(.11)	.23	.20	(.11)	.42***
BGPLHO1 No. people at home	17.17***	5.1	(.58)	-.37***	5.2	(.70)	-.11	7.0	(.49)	-.23	6.3	(.43)	-.21
CGHFSG Stud. asp. level of edu.	6.46*	4.3	(.33)	.52***	4.0	(.51)	.69***	4.2	(.42)	.26	3.5	(.42)	.34***
Instruction:													
CSDEMO1 Teacher demonstrates	35.1***	3.2	(.39)	.09	3.0	(.39)	.01	3.0	(.33)	.09	3.4	(.45)	.09
CSEXPER1 Students experimenting	15.2***	2.7	(.43)	.22*	2.7	(.38)	.03	2.7	(.39)	-.20	3.2	(.45)	.07
CTSCSWE Stud. provides explana.	1.2	2.9	(.62)	.03	2.7	(.74)	.15	2.8	(.59)	-.17	2.8	(.64)	.12
CSSGRP1 Stud. work in small grps.	22.2***	2.7	(.40)	.06	2.5	(.49)	.06	2.6	(.31)	.06	2.9	(.40)	.05
CSTEST1 Having quizzes or tests	9.5**	2.7	(.32)	-.38***	2.6	(.27)	-.49***	2.8	(.31)	-.20	2.9	(.30)	-.08
School resources:													
RESOUR1 Short. gen. sch. res	3.8*	1.7	(.61)	.00	1.6	(.69)	-.23	2.4	(.77)	-.27	1.9	(.64)	-.00
RESOUR3 Short. science sch. res	3.61	2.0	(.80)	.00	1.8	(.77)	-.03	3.0	(.79)	-.07	2.3	(.82)	.25
School climate:													
BGSPO1 Late school arrival	5.8*	2.2	(.57)	-.07	2.1	(.52)	-.12	2.2	(.60)	-.54**	1.7	(.61)	-.16
CTSGLTO6 Disrupt. stud. behavior	.01	2.1	(.63)	-.23*	2.7	(.89)	-.33**	2.2	(.56)	-.25	2.6	(.71)	-.11

Note: p =.058; *$p \leq$ 0.05; **$p \leq 0.01$; ***$p \leq 0.001$.

limits on teaching due to disruptive student behavior (CTMGLTO6, CTSGLTO6) (on a scale of 1—*not at all* to 4—*a lot*) increased toward 2003 in the Jewish sector, but there was no change in the Arab sector. These are signs of improvement in school climate indicators in the Arab sector. In addition, these variables also became, toward 2003, more negatively associated with achievement in the Hebrew-speaking schools, while less negatively associated with achievement in the Arabic-speaking schools. These changes also indicate that the negative aspects of school climate have less effect on the achievement of students in Arabic-speaking schools.

School-level variables involved in a three-way interaction with sector and year

School-level variables that changed in the two sectors in their frequency and/or association with achievement from 1999 to 2003 were fitted along with *sector* and *year* into two-level HLM models. Tables 5 and 6 summarize the results of these analyses. The models are those that exhibited statistically significant ($p \leq .05$ or near) two- or three-way interactions between the relevant school-level variable and the other two school-level predictors. The models, in mathematics, contained variables describing instruction: students working in groups (ZCMHWSG1); reviewing homework (ZCMHROH1); relating what is learned to daily life (ZCMHMDL1); working out problems on their own (ZCMHWPO1); and having quizzes or tests (ZCMHHQT1). In the case of science, the models contained variables describing instruction: students conducting experiments (ZCSEXPER1) and providing explanations (ZTSCSWE). As all school-level variables employed in these analyses were standardized, their mean values were set to 0. Negative values of these variables indicate less frequent use of these modes of instruction while positive values indicate frequent use.

Other variables included in the models describe the levels of effect that shortage in general school resources has on school capacity to provide instruction (ZRESOUR1); the levels of effect that shortage in mathematics resources (ZRESOUR2) or in science resources (ZRESOUR3) has on school capacity to provide instruction; severity of late arrivals at school as reported by principals (1—*not a problem*; 3—*a serious problem*) (ZGSPO1); and "limit to teaching" due to disruptive student behavior as reported by mathematics teachers (ZTMGLTO6), and as reported by science teachers (ZTSGLTO6) (1—*does not limit teaching*; 4—*limits teaching a lot*). Here, too, the mean values of these variables were set to 0. Negative values on the standardized scale of these school variables indicate good school climate; positive values indicate bad school climate.

The most important output of these analyses were estimates of the regression coefficients of all terms involved in the models, and their standard error of measurement. In the presence of interactions, the regression coefficients of first-order terms (i.e., of sector, year, and the relevant school variables) do not represent "main effects" or constant effects across all values of the other variables. Rather, they represent the effects of the predictor at the mean of the other predictors (Aiken & West, 1991, pp. 38–39). As all school-level variables were standardized, the coefficients of the interaction terms indicate an increase or decrease in achievement due to change by one standard deviation on the scale of the school variable, above or below its mean value. The interpretation of these regression parameters and coefficients should be as follows:

- *The intercept* in each of the models indicates the mean achievement attained by students in Arabic-speaking schools at the mean level of the relevant standardized school variables.
- γ_{01} represents additional achievement score points attained by students in Hebrew-speaking schools in 1999 at the mean level of the relevant standardized school variables, as compared to the achievement of students in Arabic-speaking schools (sector gap in favor of Hebrew-speaking schools).
- γ_{02} represents additional achievement score points attained by Arabic-speaking schools in 2003 at the mean level of the standardized relevant school variables as compared to their achievement in 1999 (year gain in favor of Arabic-speaking schools).
- γ_{03} indicates an interaction effect between *sector* and *year*, and it represents the change in achievement score points from 1999 to 2003 of students in Hebrew-speaking schools at the mean level of the *standardized* relevant school variables compared to change in achievement score points of students in Arabic-speaking schools during this period under the same conditions (year gains of Hebrew-speaking schools compared to those of Arabic-speaking schools).

Table 5: Summary of Regression Coefficients Obtained from HLM Models Predicting Mathematics Scores, each Containing Three School-Level Predictors: Sector, Year and One Standardized School/Class Level Variable

Variables	Intercept	Sector γ01	Year γ02	IntS/Y γ03	B. Coef γ04	Int2S γ05	Int2Y γ06	Int3 γ07	Book	*SE*	ACADF1	ACADM	HO1	BSV	WSV
ZCGBOOK	357*** (9)	76.8*** (8)	78.8***(10)	–49.9*** (11)	–1.7 (8)	27.0** (9)	9.4 (14)	–11.2 (15)	3.5*** (1)	10.1*** (1)	8.2*** (2)	4.5* (2)	–2.9*** (1)	1064	5190
ZACADM_M	350** (17)	72.1*** (16)	97.4***(19)	–64.4** (20)	-1.8 (14)	30.4* (14)	28.0 (19)	–32.2 (20)	4.1*** (1)	10.0***(1)	7.8*** (2)	3.2 (2)	–2.4** (1)	1056	5188
ZCGHFSG	352*** (10)	62.4***(8)	95.6*** (13)	–47.7** (14)	0.8 (8)	29.2**(10)	15.5 (10)	–20.4 (13)	4.2*** (1)	9.6*** (1)	8.2*** (2)	4.6 *(2)	–2.3** (1)	1155	5189
ZCMHWSG1	363*** (11)	68.8*** (10)	69.5***(13)	–44.5** (14)	–17.6* (8)	7.3 (9)	27.4* (12)	–19.0 (13)	4.2***(1)	10.0*** (1)	8.7*** (2)	4.9* (2)	–2.5** (1)	1495	5189
ZCMHROH1	347*** (11)	80.8*** (9)	64.6***(11)	–33.8** (13)	–5.9 (8)	1.4 (10)	34.1* (12)	–21.4 (15)	4.2*** (1)	10.0*** (1)	8.8*** (2)	5.0* (2)	–2.4** (1)	1469	5190
ZCMHWPO1	331*** (12)	99.2*** (10)	111.0***(12)	–90.4*** (14)	–15.7* (6)	7.4 (7)	36.1**(12)	–3.8 (14)	4.1*** (1)	10.0*** (1)	8.7*** (2)	5.1* (2)	–2.6*** (1)	1338	5190
ZCMHMDL1	354*** (10)	73.2*** (8)	69.0***(10)	–38.5** (12)	–12.6* (7)	0.6 (9)	26.8** (10)	–20.5* (12)	4.2*** (1)	10.0***(1)	8.6*** (2)	4.8* (2)	–2.5** (1)	1441	5189
ZCMHHQT1	350*** (9)	80.1***(7)	82.0*** (9)	–54.7*** (11)	–21.6*** (5)	14.8* (7)	27.7* (7)	–38.5*** (10)	4.2*** (1)	10.0*** (1)	8.7*** (2)	4.8* (2)	–2.4** (1)	1354	5190
ZRESOUR1	362 *** (10)	66.1*** (9)	66.7*** (1)	–37.7** (13)	–11.3* (5)	11.4 (7)	17.4* (9)	–23.2*(11)	4.2*** (1)	10.0*** (1)	8.7*** (2)	5.0* (2)	–2.4** (1)	1528	5189
ZRESOUR2	361*** (11)	65.7 ***(9)	65.5*** (11)	–36.8** (12)	–10.3 (6)	8.2 (8)	28.2** (10)	–34.2** (9)	4.2*** (1)	10.0*** (1)	8.7*** (2)	4.9* (2)	–2.4** (1)	1494	5189
ZGSPO1	370*** (9)	71.1*** (8)	55.3*** (11)	–27.3* (13)	–14.3** (5)	10.8 (6)	7.1 (10)	–13.7 (12)	3.5*** (1)	7.2*** (1)	11.2*** (2)	6.1** (2)	–2.4** (1)	1429	5262
ZTMGLTO6	362*** (10)	76.9*** (9)	72.7*** (10)	–38.2** (12)	–10.0* (4)	4.6 (7)	8.7 (8)	–20.1* (11)	3.3*** (1)	7.3*** (1)	10.7*** (2)	6.7** (2)	–2.4**(1)	1370	5222

Notes: *p* ≤ 0.07; **p* ≤ 0.05; ***p* ≤0.01; ****p* ≤0.001.

Table 6: Summary of Regression Coefficients Obtained from HLM Models Predicting Science Scores, each Containing Three Predictors: Sector, Year and One Standardized School/Class Level Variable

Variables	Intercept	Sector γ01	Year γ02	IntS/Y γ03	B. Coef γ04	Int2S γ05	Int2Y γ06	Int3 γ07	Book	*SE*	ACADF1	ACADM	HO1	BSV	WSV
ZCGBOOK	338*** (12)	89.3*** (10)	81.4*** (11)	–69.7***(12)	–3.6 (14)	31.8* (15)	11.5 (16)	–24.8 (17)	7.9*** (1)	19.3*** (1)	4.7* (2)	2.9 (2)	–3.0*** (1)	1139	5852
ZACADM_F	347*** (13)	68.7*** (12)	75.5*** (16)	–59.2** (17)	23.7* (10)	3.0 (12)	-16.5(15)	17.2 (16)	8.3*** (1)	10.0*** (1)	3.1 (2)	2.4 (2)	–2.5** (1)	976	5850
ZCSEXPER1	333*** (12)	90.7*** (11)	76.0*** (13)	–63.9*** (14)	–13.7 (10)	27.2** (10)	18.1 (12)	–26.7* (13)	8.5*** (1)	10.2*** (1)	5.1* (2)	3.4** (2)	–2.6** (1)	1439	5851
ZTSCSWE	359*** (13)	80.8*** (12)	62.7*** (12)	–52.1*** (14)	–21.1** (6)	16.0* (9)	29.5** (8)	–21.8* (12)	6.1*** (1)	7.4*** (1)	7.4** (2)	6.3** (2)	–2.1* (1)	1456	5914
RESOUR1	344*** (14)	77.7*** (13)	68.6*** (13)	–57.3*** (15)	–11.8 (10)	13.9 (10)	14.7 (11)	–26.9* (13)	8.6*** (1)	10.1*** (1)	5.1* (2)	3.4 (2)	–2.5** (1)	1464	5852
ZGSPO1	360*** (11)	79.4*** (10)	58.7*** (13)	–48.2** (14)	–21.5*** (6)	19.0* (8)	18.3* (9)	–22.3* (12) 1	6.1*** (1)	7.3*** (1)	7.3** (2)	6.3** (2)	–2.1* (1)	1423	5915
ZTSGLTO6	350*** (13)	84.7*** (13)	70.0*** (13)	–54.2*** (15)	–27.1 (11)	18.9 (13)	27.7* (12)	–33.3* (15)	6.3*** (1)	7.7*** (1)	6.7** (2)	6.0** (2)	–2.2** (1)	1326	5935

Notes: *p* ≤ 0.07; **p* ≤ 0.05; ***p* ≤0.01; ****p* ≤0.001.
BSV = Between school variance; WSV = Within school variance.

- γ_{04} indicates change in achievement score points of students in Arabic-speaking schools in 1999 due to an increase of 1 *SD* above the standardized mean value of the relevant standardized school variables.
- γ_{05} indicates an interaction effect of the relevant school variable and year, and it represents a change in achievement score points from 1999 to 2003 in the Arab sector due to an increase of 1 *SD* above the standardized mean value of the relevant school variables.
- γ_{06} indicates an interaction effect of the relevant school variable and *sector*, and it represents additional score points that students in Hebrew-speaking school achieved in 1999 due to an increase of 1 *SD* above the standardized mean of the *relevant* school variable compared to the gains of students in Arabic-speaking schools under the same conditions.
- γ_{07} indicates a three-level interaction effect of *sector*, *year*, and the *relevant* school variable, and represents change in achievement score points from 1999 to 2003 in the Hebrew-speaking sector versus that in the Arab sector due to an increase of 1 *SD* above the mean of the standardized relevant school variable.

Adopting these interpretations to the results presented in Tables 5 and 6 reveals some similarities.

Sector and *year* (γ_{01} γ_{02}) were found to have significant positive effects in all analyses, pointing to higher achievement gains in 1999 in Hebrew-speaking schools than in Arabic-speaking schools at the mean of all *standardized* school variables employed in the analyses, and higher year achievement gains in 2003 than in 1999 of Arabic-speaking schools at the standardized mean of all school variables.

The significant negative two-way interaction effects of *sector* and *year (sector*year)* (γ_{03}) that were found in all models analyzed indicated lower gains in achievement from 1999 to 2003 of students in the Hebrew-speaking schools compared to the gain of students in Arabic-speaking schools due to the changes that occurred in the relevant school-level variables during this period.

The positive two-way interaction effects between the selected *school variables* and *year* (γ_{05}) indicated an increase in the achievement from 1999 to 2003 in the Arab sector due to an increase of 1 *SD* above the mean of the standardized school variable.

The positive two-way interaction effect between the selected *school variable* and *sector* (γ_{06}) represented the additional score points students in Hebrew-speaking schools achieved in 1999 due to an increase of 1 *SD* above the mean of the standardized school variables compared to gains of students in Arabic-speaking schools under the same conditions.

All coefficients of the three-way interaction effect (γ_{07}) in Tables 5 and 6 were found to have a negative effect, indicating a decrease in the achievement gains,[6] from 1999 to 2003, of students in Hebrew-speaking schools compared to gains of students in Arabic-speaking schools (higher year gains in the Arab sector).

The significant three-way interaction terms found in the mathematics models were with testing (ZCMHHQT1) and, to a lesser significance level ($p \leq .08$), with relating what is learnt to daily life (ZCMHMDL1). Similarly, significant three-way interaction terms in the mathematics models were found with the indices of the effect on school capacity to provide instruction due to shortage in general school resources (ZRESOUR1), and in mathematics school resources (ZRESOUR2), as well as with variables describing the limit to teaching due to disruptive student behavior (ZTMGLTO6).

In the science models, the school variables involved in significant three-way interaction with sector and year were the frequency with which students conducted experiments (ZCSEXPER1), how often their teachers asked them to provide explanations (ZTSCSWE), shortage in general school resources (ZRESOUR1), and the limit to teaching due to disruptive student behavior in science classes (ZTSGLTO6). An increase of 1 *SD* above the standardized mean of these variables resulted, in both subject areas, in lower gains from 1999 to 2003 in the achievement of students in Hebrew-speaking schools compared to the achievement gains of students in Arabic-speaking schools.

Probing the three-way interaction terms

Regressing science and mathematics scores by means of hierarchical models that contain beyond student-level variables, three second-level variables (i.e., *sector*, *year*, and one meaningful school contextual variable),

[6] Due to an increase of 1 *SD* above the mean value of the standardized selected school variables involved in the analyses.

as well as including all interactions amongst them allowed the computation of predicted outcomes in 1999 and 2003 for students in Arabic-speaking schools and for students in Hebrew-speaking schools. These computations are based on three values of the relevant school-level variables—maximal, mean, and minimal.

The following tables present the predicted outcomes in mathematics (Table 7) and in science (Table 8). I have chosen to present findings from these simulations related to variables that describe instruction or school resources and learning climate, as these variables are relatively easy to manipulate and probing their effect can enable the formulation of policy recommendations.

The predicted outcomes show, for both school subjects, that at the mean and the maximal values of the selected variables, student achievement gains from 1999 to 2003 (*year* gains) were much more profound in Arabic-speaking schools. As a result, the achievement gap between the sectors in favor of Hebrew-speaking schools (*sector* gains) narrowed over time and even changed direction to be in favor of Arabic-speaking schools.

In mathematics at the mean and maximal values of some instructional variables (testing, relating what is learnt to daily life, working out problems on one's own, reviewing homework), the year gains of students in the Arabic-speaking schools were much higher than those of students in the Hebrew-speaking schools. As a result, the sector gap narrowed and, in the case of frequent testing and frequent relating what is learnt to daily life, Arabic-speaking students outperformed their peers in Hebrew-speaking schools.

The same happened in science at the mean and maximal values of all instructional variables involved. The year gains of students in Arabic-speaking schools were higher than are those of their peers in Hebrew-speaking schools. This caused the *sector* gap to narrow. Here, too, in the case of frequent testing, Arabic-speaking students outperformed students in Hebrew-speaking schools.

Predicted outcomes in both mathematics and science, due to levels of the effect that shortage in general school resources have on school capacity to provide instruction, exhibit the same pattern. At the maximal level of effect (schools with low resources), the year gains of Arabic-speaking students were higher and the *sector* gap narrowed and almost disappeared.

At the low levels of effect, due to shortage in school resources (affluent school conditions), the *year* gains of students in Arabic-speaking schools were similar to those of their peers in Hebrew-speaking schools, and so the *sector* gap remained.

At low levels of limit to instruction caused by disruptive student behavior (positive school climate), the *year* gains of the two sectors were similar and the *sector* gap in favor of students in Hebrew-speaking schools remained the same from 1999 to 2003. However, at the mean and maximal levels of limit to teaching due to disruptive student behavior, the year gains of students in Arabic-speaking schools were much bigger than the *year* gains of students in the Hebrew-speaking schools, and so the *sector* gap declined. The negative impact on achievement of disruptive student behavior appeared in Hebrew-speaking schools, but it did not affect the achievement of students in Arabic-speaking schools.

Plotting these predicted outcomes for the two school subjects in three separate graphs dependent on the level of the relevant variable and on its association with achievement at the two points in time, for Hebrew-speaking schools and for Arabic-speaking schools, allows us to visualize the narrowing achievement gap between the sectors that occurred from 1999 to 2003.

The plots chosen to illustrate the narrowing achievement gaps in mathematics are those related to models that contain variables describing instruction and school conditions that affect learning: students working out problems on their own (ZCMHWPO1); frequent relating of what is learnt to daily life (ZCMHMDL1); frequent testing (ZMHHQT1); shortage in general school resources which affects instruction (ZRESOUR1); and disruptive student behavior (ZTMGLTO6) (Figures 1, 2, 3, 4, and 5). Changes in these variables represent either efforts made by the teachers in their classes or efforts made at the school level in line with some national interventions.

In science, the plots are those related to models that contain variables describing the frequency of students providing explanations (ZTSCSWE), frequency of working in small groups on investigations (ZCSSGRP1), having quizzes or tests (ZCSTEST1), and that contain the variable that describes shortage in general school resources which affects instruction (ZRESOUR1) as well as the negative effect of students' disruptive behavior (ZTSGLTO6) (Figures 6, 7, 8, 9, and 10).

Table 7: Predicted Mathematics Outcomes at Three Levels of Instruction and School Variables

		Minimum level			Mean level			Maximal Level		
		1999	2003		1999	2003		1999	2003	
		$\bar{x}$	$\bar{x}$	*Year* gain	$\bar{x}$	$\bar{x}$	*Year* gain	$\bar{x}$	$\bar{x}$	*Year* gain
Working out problems	HS	517.6	399.0	-118.6	483.6	497.0	13.4	466.9	545.1	78.2
on their own	AS	370.5	393.5	23.0	333.7	477.0	143.3	343.1	518.1	175.0
(ZCMHWPO1)	*Sector* gap	147.1	5.5		149.9	20.0		129.8	27.0	
Reviewing homework	HS	494.6	483.3	-11.3	482.2	505.9	23.7	473.9	521.1	47.3
	AS	408.2	370.4	-34.8	346.6	447.8	101.2	380.9	500.1	119.2
(ZCMHROH1)	*Sector* gap	86.4	112.9		135.6	58.1		93.0	21.1	
Relating what is learnt	HS	513.6	521.5	7.9	481.4	504.8	23.4	438.2	482.3	44.1
to daily life	AS	429.6	424.4	-5.2	353.7	459.1	105.4	357.5	505.5	148.0
(ZCMHMDL1)	*Sector* gap	84.2	97.1		127.2	45.7		80.7	-23.2	
Having quizzes or tests	HS	503.2	554.7	51.2	483.9	504.1	20.2	456.9	434.2	-22.7
(ZCMHHQT1)	AS	456.7	450.3	-6.4	349.5	467.8	118.3	308.4	492.0	183.6
	Sector gap	46.8	104.4		134.4	36.8		148.5	-57.8	
Effect of shortage in	HS	482.1	510.8	28.7	482.3	504.1	21.8	482.6	488.5	5.9
general school	AS	419.9	457.6	37.7	361.6	464.7	103.1	375.7	481.2	105.5
resources (ZRESOUR1)	*Sector* gap	62.2	53.2		120.7	39.4		106.9	7.3	
Effect of disruptive	HS	489.2	537.1	47.9	480.2	509.1	28.9	468.3	472.1	-14.2
student behavior	AS	410.1	462.0	52.9	362.0	459.8	97.8	371.3	456.9	85.6
(ZTMGLTO6)	*Sector* gap	79.2	75.1		118.2	49.3		115.0	15.2	

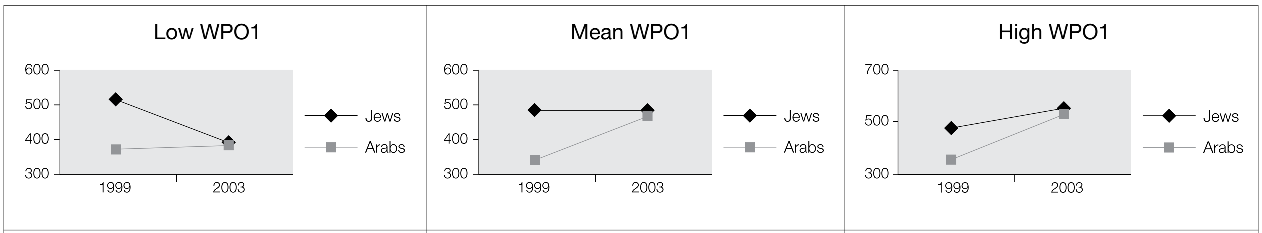
Notes: HS = Hebrew-speaking schools; AS = Arabic-speaking schools.
Year gain = 2003–1999; *Sector* gap = HS–AS.

Table 8: Predicted Science Outcomes at Three Levels of Instruction and School Variables

		Minimum level			Mean level			Maximal level		
		1999	2003		1999	2003		1999	2003	
		$\bar{x}$	$\bar{x}$	*Year* gain	$\bar{x}$	$\bar{x}$	*Year* gain	$\bar{x}$	$\bar{x}$	*Year* gain
Students watch teacher	HS	454.1	481.3	27.2	487.7	496.6	8.9	500.2	502.3	2.1
demonstrate experiments	AS	391.3	430.1	61.2	335.4	459.0	123.6	405.8	468.8	64.0
(ZCSDEMO1)	*Sector* gap	62.8	51.2		152.3		37.6	94.4	32.5	
Students conduct	HS	440.3	477.6	37.3	491.2	495.5	4.7	521.9	506.9	-15.0
experiments	AS	443.7	442.4	-1.3	333.3	459.0	125.7	360.1	468.9	108.9
(ZCSEXPER1)	*Sector* gap	-3.4	35.4		158.2	36.9		161.8	38.0	
Students provide	HS	499.9	491.0	-8.9	490.5	495.5	5.0	482.3	499.5	17.2
explanations	AS	439.2	442.2	3.0	358.8	457.3	98.5	367.0	470.6	103.6
(ZTSCSWE)	*Sector* gap	60.7	48.9		131.7	38.2		115.3	28.9	
Students have quizzes	HS	537.1	542.7	5.6	492.1	487.6	-4.5	452.5	439.0	-13.5
or tests	AS	421.8	477.7	55.9	337.3	465.7	128.4	371.5	458.7	87.2
(ZCSTEST1)	*Sector* gap	115.3	65.0		154.8	21.9		81.0	-19.7	
Students work in small	HS	478.3	488.4	10.1	488.3	496.2	7.9	496.2	501.8	5.6
(ZCSSGRP1) group on	AS	388.8	443.2	54.4	333.7	459.8	126.1	394.2	471.9	77.7
investigation	*Sector* gap	89.5	45.2		154.6	36.4		102.0	29.9	
Effect of general school	HS	486.9	505.1	18.2	489.3	493.3	4.0	495.1	465.6	-29.5
resources	AS	416.6	459.4	42.8	344.2	462.9	118.7	370.5	470.8	100.3
(ZRESOUR1)	*Sector* gap	70.3	45.7		145.1	30.4		124.6	-5.2	
Limit to teaching due to	HS	501.7	521.8	20.1	487.0	497.0	10.0	471.0	470.1	-0.9
disruptive student	AS	442.5	455.8	13.3	350.5	457.0	106.5	340.2	458.2	118.0
behavior (ZTSGLTO6)	*Sector* gap	59.2	66.0		136.5	40.0		130.8	11.9	

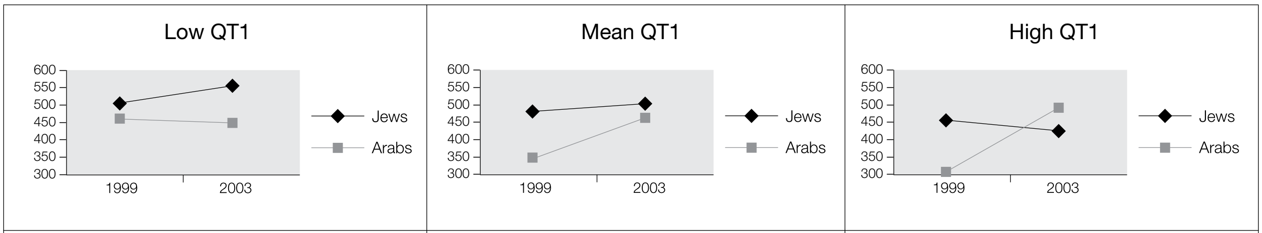
Notes: HS = Hebrew-speaking schools; AS = Arabic-speaking schools.
Year gain = 2003–1999; *Sector* gap = HS–AS.

Figure 1: Predicted Mathematics Scores by Sector, Year, and by Students Working Out Problems on their Own (WPO1)



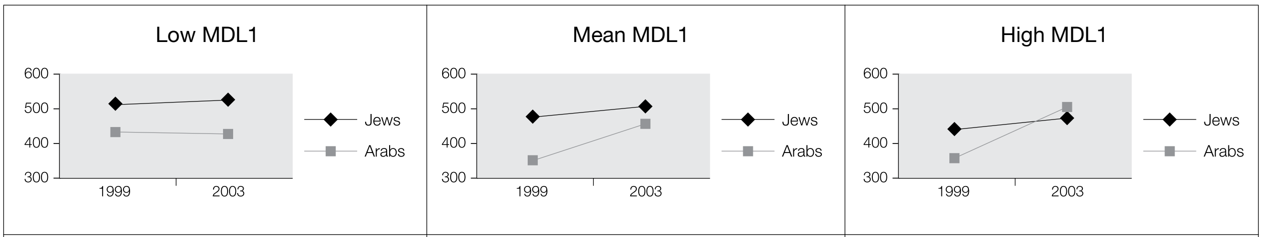
	Low WPO1			Mean WPO1			High WPO1	
	1999	2003		1999	2003		1999	2003
Jews	517.63	399.02	Jews	483.61	496.96	Jews	466.90	545.06
Arabs	370.50	393.46	Arabs	330.74	477.04	Arabs	343.08	518.08

Figure 2: Predicted Mathematics Scores by Sector, Year, and by Students Having a Test or a Quiz (QT1)



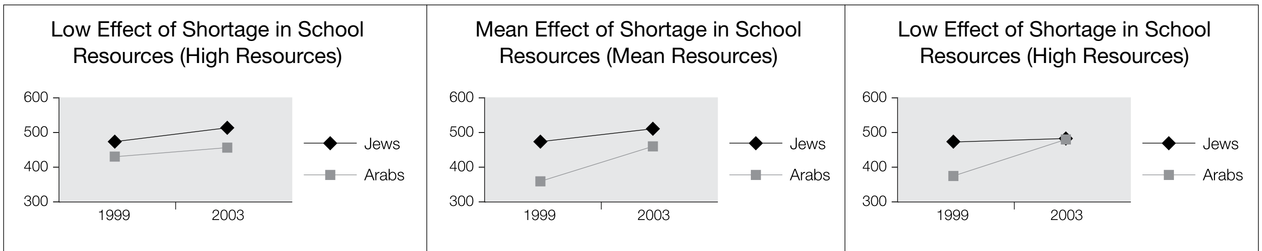
	Low QT1			Mean QT1			High QT1	
	1999	2003		1999	2003		1999	2003
Jews	503.53	554.66	Jews	483.95	504.08	Jews	456.88	434.18
Arabs	456.69	450.29	Arabs	349.55	467.81	Arabs	308.38	492.03

Figure 3: Predicted Mathematics Scores by Sector, Year, and by Students Relating What is Learnt to Daily Life (MDL1)



	Low MDL1			Mean MDL1			High MDL1	
	1999	2003		1999	2003		1999	2003
Jews	513.57	521.52	Jews	481.38	504.77	Jews	438.22	482.32
Arabs	429.41	424.45	Arabs	353.71	459.07	Arabs	357.49	505.50

Figure 4: Predicted Mathematics Scores by Sector, Year, and by Effect of Shortage in General School Resources (ZRESOUR1)


Low Effect of Shortage in School Resources (High Resources)

Mean Effect of Shortage in School Resources (Mean Resources)

Low Effect of Shortage in School Resources (High Resources)

Low Resources

	1999	2003
Jews	482.13	510.79
Arabs	419.90	457.65

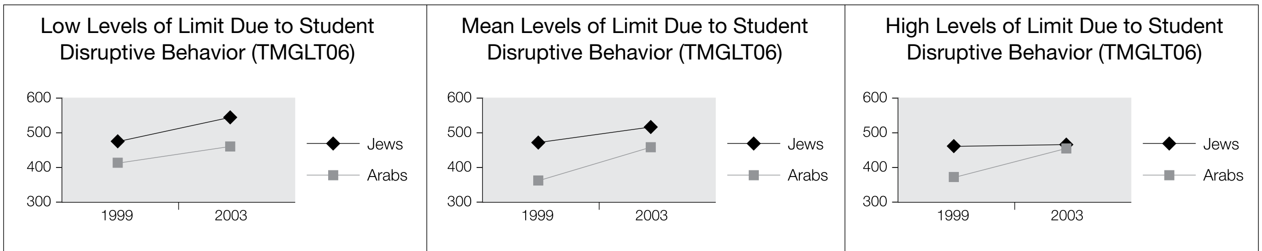
Mean Resources

	1999	2003
Jews	482.27	504.12
Arabs	361.59	464.71

High Resources

	1999	2003
Jews	482.61	488.53
Arabs	375.70	481.22

Figure 5: Predicted Mathematics Scores by Sector, Year, and by Limit to Teaching due to Disruptive Student Behavior (ZTMGLT06)


Low Levels of Limit Due to Student Disruptive Behavior (TMGLT06)

Mean Levels of Limit Due to Student Disruptive Behavior (TMGLT06)

High Levels of Limit Due to Student Disruptive Behavior (TMGLT06)

Low TMGLT06

	1999	2003
Jews	489.19	537.14
Arabs	410.07	462.03

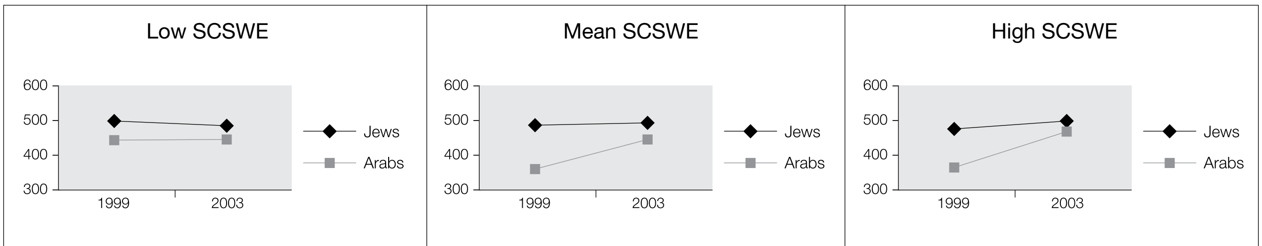
Mean TMGLT06

	1999	2003
Jews	480.19	509.10
Arabs	362.05	459.81

High TMGLT06

	1999	2003
Jews	468.31	472.10
Arabs	371.27	456.88

Figure 6: Predicted Science Scores by Sector, Year, and by Students Providing Explanations (ZTSCSWE)


Low SCSWE

Mean SCSWE

High SCSWE

Low SCSWE

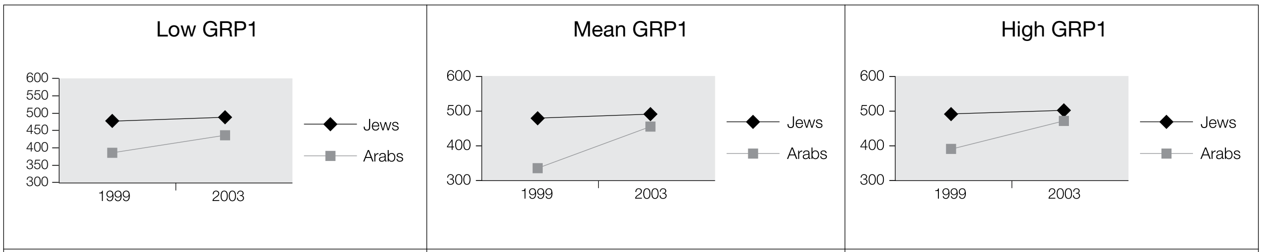
	1999	2003
Jews	499.92	490.97
Arabs	439.16	442.15

Mean SCSWE

	1999	2003
Jews	490.54	495.51
Arabs	358.77	457.30

	1999	2003
Jews	482.33	499.48
Arabs	367.01	470.57

Figure 7: Predicted Science Scores by Sector, Year, and by Students Working in Small Groups on Experiments and Investigations in Class (GRP1)

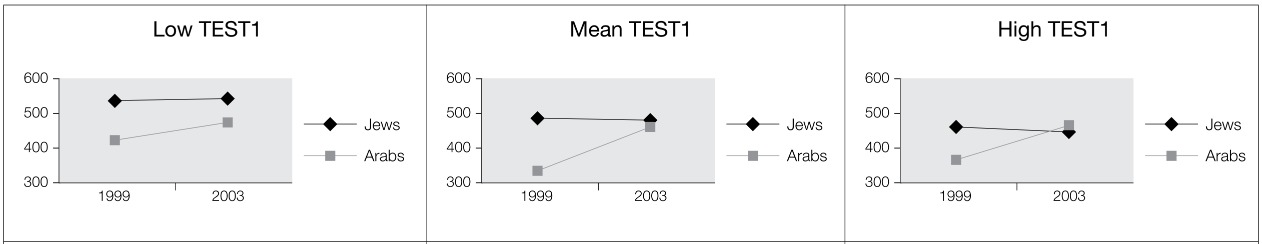


Low GRP1	1999	2003
Jews	478.32	488.43
Arabs	388.78	443.24

Mean GRP1	1999	2003
Jews	488.64	496.18
Arabs	333.70	459.81

High GRP1	1999	2003
Jews	496.16	501.83
Arabs	394.25	471.88

Figure 8: Predicted Science Scores by Sector, Year, and by Students Having Tests or Quizzes in Class (TEST1)

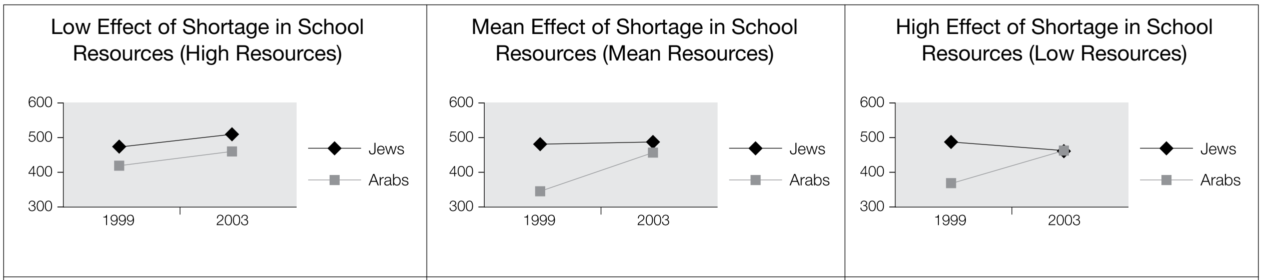


Low TEST1	1999	2003
Jews	537.07	542.71
Arabs	421.81	477.72

Mean TEST1	1999	2003
Jews	492.15	487.61
Arabs	337.28	465.71

High TEST1	1999	2003
Jews	452.53	439.02
Arabs	371.53	458.73

Figure 9: Predicted Science Scores by Sector, Year, and by Effect of Shortage in General School Resources (ZRESOUR1)

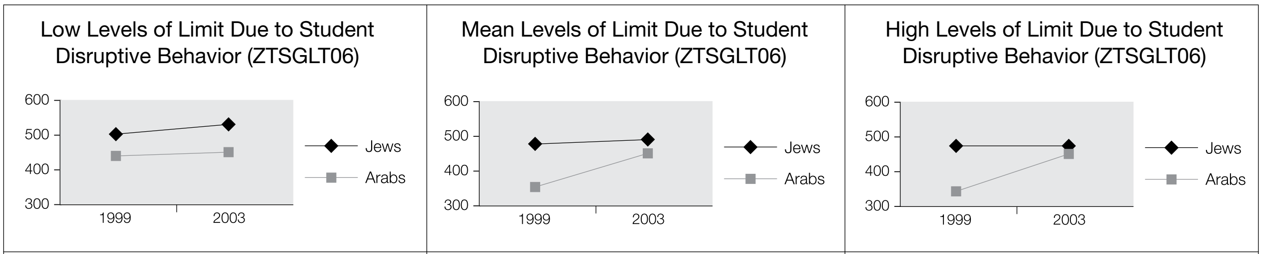


Low	1999	2003
Jews	486.91	505.13
Arabs	416.56	459.45

Mean	1999	2003
Jews	489.35	493.28
Arabs	344.21	462.86

High	1999	2003
Jews	495.06	465.58
Arabs	370.52	470.82

Figure 10: Predicted Science Scores by Sector, Year, and by Limit to Teaching due to Disruptive Student Behavior (ZTSGLT06)



Low ZTSGLTO6	1999	2003	Mean ZTSGLTO6	1999	2003	High ZTSGLTO6	1999	2003
Jews	501.71	521.81	Jews	486.96	496.97	Jews	471.03	470.13
Arabs	442.46	455.78	Arabs	350.46	456.96	Arabs	340.17	458.24

Discussion and conclusions

This paper aimed to capture the dynamics that occurred in Israeli schools from 1999 to 2003, which eventually led to the narrowing of a persistent gap in achievement between students in Hebrew-speaking and Arabic-speaking schools. The trigger for this study was the relatively higher achievement gains in mathematics and science of students in Arabic-speaking schools detected in the TIMSS 2003 study in Israel. In view of continuing social and economic disparities between the two ethnic populations, and given the TIMSS 2003 study, on the one hand, and special efforts made by the Ministry of Education during the 1990s to improve schooling conditions in Arabic-speaking schools, on the other, our attention focused mainly on variables operating at the school/class level.

In contrast to studies that measure effectiveness from data derived from a one-time survey, this study aimed to consider the effect of temporal changes in the frequency and/or the effectiveness of a selected set of school variables describing instruction, school resources, and school learning climate that were obtained at two points in time and were found to interact significantly with *sector*—ethnic affiliation of school—and *year* of study.

The findings of this study revealed differential ethnicity-related changes in the frequency and effectiveness of instructional variables operating mostly at the class level and of variables describing material conditions and learning climate operating at the school level.

Changes occurring in the frequency and effectiveness of certain instructional modes often reflect changing pedagogical fashions worldwide. In contrast, changes occurring in school-level variables that describe resources and learning climate tend to reflect policy interventions and deliberate national efforts invested in implementing them.

The changes that occurred in the frequency and effectiveness of self-managed problem-solving in mathematics classes in both sectors, and the increased emphasis given to relating study material to daily life in Arabic-speaking schools, reflect a new trend in mathematics instruction and echoes a debate in this field regarding the appropriateness of the "conceptual" mode of mathematics teaching versus the "computational" mode of mathematics teaching (Desimone, Smith, Baker, & Ueno, 2005). Scholars who advocate conceptual teaching emphasize real-world problem-solving: working with problems that have no obvious solutions, discussing alternative hypotheses, and using investigations to solve problems (Hiebert et al., 1996). Scholars in favor of a computational instructional practice focus on routine drill and practice (Li, 1999). Conceptual teaching has been regarded as more appropriate for high-performing students, while computational instruction might be seen as more appropriate for low-performing students. However, recent studies suggest that low-achieving students can master more demanding intellectual problems while simultaneously learning basic skills (Lo Cicero, De la Cruz, & Fuson, 1999; Mayer, 1999), and that blending demanding academic work in computational instruction is, in fact, advantageous to underachievers (Knapp, 1995).

Findings in this study concerned independent problem-solving and relating what is learnt to daily life, which both fall into the category of conceptual teaching, together with other findings obtained in TIMSS 2003 in Israel concerning the effectiveness of computational activities (practicing four basic arithmetic operations, working on fractions and decimals, interpreting graphs and tables, and writing equations), which were more effective in Arabic-speaking schools (Zuzovsky, 2005). All of these go toward supporting the claim that this blend of practices will improve attainment of students in Arabic-speaking schools, which have been known as low performing.

In science, differential changes occurred in the frequency and effectiveness of several practices that can be identified with an inquiry mode of science instruction. Students conducting experiments on their own, working in small groups on investigations, and providing explanations to observed phenomena are all practices in line with this instructional approach. These modes of instruction became, over time, more prevalent and more effective in the Arabic-speaking schools while they grew less so in the Hebrew-speaking ones. As in mathematics teaching, there is debate in science teaching as to whether inquiry-oriented instruction—now regarded as mainstream pedagogy—which encourages students to ask questions, find out answers on their own, and experiment on their own, is congruent with some cultural and home values that non-mainstream students bring from home (Au, 1980; Philips, 1983). On the other side of this debate are scholars who consider that replacing the mainstream pedagogy will deprive non-mainstream students of opportunities to learn academic skills such as inquiry skills and scientific knowledge (Fradd & Lee, 1999; Ladson-Billings, 1994, 1995; Lee, 2003; Lee & Luyks, 2005; Lubienski, 2003). These scholars advocate explicit instruction to non-mainstream students in ways that reflect the dominant culture's rules and norms for classroom participation and discourse. In line with the second suggestion, our study indicates a growth in the popularity of inquiry-oriented instruction in Arabic-speaking schools (non-mainstream students and teachers) that, indeed, turned out to be positively associated with achievement in this sector.

To sum up, the higher achievement gains of the Arab sector seem to be a result of adopting mainstream pedagogy in both school subjects: the conceptual approach in mathematics and the inquiry-oriented approach in science. These more demanding modes of instruction did not exclude more traditional, still very effective, modes, such as the computational mode and listening to lectures, which provide students in Arabic-speaking schools with suitable instruction.

The other findings obtained in this study relate to policy interventions carried out in Israel with the aim of reducing inequalities between the two sectors in school resources and infrastructure. These efforts were coupled with a growing demand for accountability, assessment, and monitoring outcomes. The success and impact of these policy interventions could be traced through changes that occurred in the variable that described the effect of shortage in school resources and the effect of testing.

The decrease in the index describing the harmful effect of shortage in school resources on instruction, which occurred in both sectors, but more so in Arabic-speaking schools, points to the success of the five-year plan that aimed to improve the conditions of Arabic-speaking schools. However, it is worth noting that improved conditions do not guarantee their effective use. The findings of this study show that under maximal effect of shortage (low resources), the year gains of Arabic-speaking schools from 1999 to 2003 were found to be higher than those of Hebrew-speaking schools, causing the achievement gap in favor of Hebrew-speaking schools to narrow. Under minimal effect of shortage (high resources), the year gains in the two sectors were almost the same, and the sector gap remained.

Over the last few years, a national assessment project has operated in Grades 4, 6, and 8 in Israel. Each year, these classes in half of the schools in Israel are tested in the major school subjects, including mathematics and science. This testing, in addition to regular teacher tests, is a burden on many students and teachers. While the frequency of testing, mostly decided at a national level, is similar in the two sectors and did not change from 1999 to 2003, it became negatively associated with achievement in Hebrew-speaking schools while less negatively, and even positively, associated with achievement in Arabic-speaking schools. Frequent assessment is another example of policy with a differential effect that should be considered when planning accountability policy nationwide.

The narrowing school achievement gap between the Jewish sector and the Arab sector in Israel is, in the first place, a result of the effects teachers have in their classes due to adopting and adapting suitable modes of instruction at the class level. At the national level, policy interventions, even if successfully implemented, do not always add to the higher achievement gains of students in the Arab sector.

Appendix

School-level variables

BSBGBOOK: Number of books in student's home (1—*none or few* to 5—*more than 200*)
BSBGHFSG: Highest level of education the student aspires to finish (1—*secondary* to 5—*beyond first academic degree*)
ACADEM_F or ACADEM_M: *Either parent with academic education* (1) or *lesser education* (0)
BGPLHO1: Number of people living in the student's home (2 to 8 or 8 or more)

School/class-level variables

sector: 0—*Arabic-speaking schools*; 1—*Hebrew-speaking schools*
year: 0—1999; 1—2003

Student body composition variables

CBGBOOK, CGHFSG, ACADEM-F, ACADEM-M, BGPLHO1
Variables describing the student body composition are standardized class aggregates or means of the above-described student variables.

Variables describing mathematics instruction

Class average of student responses regarding how often certain instructional activities take place (1—*never* to 4—*every lesson or almost every lesson*)
CMHWSG1: Working together in small groups
CMHMDL1: Relating what is learnt to daily life
CMHROH1: Reviewing homework
CMHWPO1: Working out problems on their own
CMHHQT1: Having a quiz or test

Variables describing science instruction

CSDEMO1: Watching the teacher demonstrate an experiment
CSEXPER1: Conducting an experiment
CSSGRP1: Working in small groups on experiments
CSTEST1: Having a quiz or a test

Another instructional variable was derived from teacher questionnaires (on a scale from 1—*never* to 4—*every or almost every lesson*). It described how often students were asked to provide explanations in their class—TSCSWE.

Indices describing shortage in resources that affect school capacity to provide instruction

Indices built on the principal's responses to a set of items describing shortage in "general" school resources (instructional materials, budget, buildings, heating/cooling, space, equipment for handicapped)—RESOUR1; shortage in mathematics school resources (computers, computer software, calculators, library materials, audiovisual resources)—RESOUR2; shortage in science school resources (laboratory equipment, computers, software, calculators, library material, audiovisual resources)—RESOUR3. The scale of these variables has four categories ranging from 1—*not affected due to shortage* to 4—*greatly affected.*

School-level variables describing learning climate

Only two variables were chosen as indicators for this issue: principal's responses on the severity of late arrival at school (scale 1—*not a problem* to 3—*serious problem*)—LSPO1.

Teacher responses regarding the extent disruptive students' behaviors limit their teaching—CTMGLTO6 for mathematics teachers or CTSGLTO6 for science teachers (both on a scale of 1—*not at all* to 4—*a lot*).

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, London, New Delhi: Sage.

Aitkin, M. (1988). *Multilevel models for educational systems*. Paper presented at the Educational Testing Service, Princeton, New Jersey.

Aitkin, M., & Zuzovsky, R. (1994). Multilevel interaction models and their use in the analysis of large-scale effectiveness studies. *School Effectiveness and School Improvement, 5*(1), 45–73.

Al Haj, M. (1995). *Education, empowerment and control: The case of the Arabs in Israel*. Albany, NY: State University of New York Press.

Au, K. H. (1980). Participation structures in a reading lesson with Hawaiian children. Analysis of a culturally appropriate instructional event. *Anthropology and Education Quarterly, 11*, 91–115.

Aviram, T., Cafir, R., & Ben Simon, A. (1996). *National mashov [feedback system] in 8th grade mathematics*. Ministry of Education and Culture, Chief Scientists Bureau and the Center for Testing and Evaluation, Jerusalem (in Hebrew).

Bashi, Y., Kahan S., & Davis, G. (1981). *Learning achievement in the Arabic elementary school in Israel*. Jerusalem: School of Education, The Hebrew University of Jerusalem (in Hebrew).

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Cafir, R., Aviram, T., & Ben Simon, A. (1999). *National mashov [feedback system] in 6th graders' science and technology*. Jerusalem: Ministry of Education and Culture, Chief Scientists Bureau and the Center for Testing and Evaluation (in Hebrew).

Desimone, L. M., Smith, T., Baker, D., & Ueno, K. (2005). Assessing barriers to reform of US mathematics instruction from an international perspective. *American Educational Research Journal, 42*, 501–536.

Dovrat Committee (Israel) National Task Force for the Advancement of Education in Israel. (2005). *National plan for education*. Jerusalem: Government of Israel.

Fradd, S. H., & Lee, O. (1999). Teachers' roles in promoting science inquiry with students from diverse language backgrounds. *Educational Researcher, 28*, 4–20, 42.

Gaziel, H., Elazar, D., & Marom, Z. (1993). *The educational policy in Israel at a crossroad*. Jerusalem: Institute for the Study of Educational Systems (ISES) (in Hebrew).

Hiebert, J., Carpenter, T. P., Fennema, E., Fuson, K., Human, P., Murray, H., et al. (1996). Problem-solving as a basis for reform in curriculum and instruction: The case of mathematics. *Educational Researcher, 25*(4), 12–21.

Kahan, S., & Yelnik, J. (2000). *Discrimination in budget allocation in the non-Jewish sector: Quantitative estimates and the consequences of its extinction*. Jerusalem: Hebrew University (in Hebrew).

Karmarski, B., & Mevarech, Z. (2004). *Reading literacy, mathematics and science*. PISA study 2002: Scientific report. Ramat Gan: School of Education, Bar Ilan University.

Knapp, M. (1995). *Teaching for meaning in high-poverty classrooms*. New York: Teachers College Press.

Knesset Research and Information Center. (2004). *Background report on school outcomes in the Arab sector presented to the Child Rights Committee in the Knesset*. Israel: Author.

Ladson-Billings, G. (1994). The dream-keepers: *Successful teachers of African-American children*. San Francisco, CA: Jossey-Bass.

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal, 32*, 465–491.

Lavi, V. (1997). *Differences in resources and achievement in the Arab education in Israel*. Jerusalem: Florsheimer Institute for Policy Studies.

Lee, O. (2003). Equity for culturally and linguistically diverse students in science education: A research agenda. *Teachers College Record, 105*, 465–489.

Lee, O., & Luyks, A. (2005). Dilemmas in scaling up innovations in elementary science instruction with non-mainstream students. *American Educational Research Journal, 42*, 411–438.

Li, S. (1999). Does practice make perfect? *For the Learning of Mathematics, 19*(3), 33–35.

Lo Cicero, A., De La Cruz, Y., & Fuson, K. (1999). Teaching and learning creatively: Using children's narratives. *Teaching Children Mathematics, 5*, 544–547.

Lubienski, S. (2003). Celebrating diversity and denying disparities: A critical assessment. *Educational Researcher, 32*, 30–38.

Mari, S. (1978). *Arab education in Israel*. Syracuse, NY: Syracuse University Press.

Mari, S., & Dahir, N. (1978). *Facts and trends in the development of Arab education in Israel*. Haifa: Institute for the Arab Education Studies, Haifa University (in Hebrew).

Mayer, D. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis, 21*, 29–46.

Mazawi, A. E. (1996). *The structure of equality in educational opportunities within the Arab school system in the state of Israel*. Doctoral thesis submitted to the Senate of Tel Aviv University.

Peled, E. (1976). *Education in Israel in the 80s*. Jerusalem: State of Israel, Ministry of Education and Culture (in Hebrew).

Philips, S. (1983). The invisible culture. *Communication in classroom and community on the Warm Springs Indian Reservation*. New York: Longman.

Raudenbush, S. (1989). The analysis of longitudinal, multilevel data. *International Journal of Educational Research, 13*(7), 721–740.

Raudenbush, S., Bryk, A., Cheong, Y. E., & Congdon, R. (2000). HLM5. *Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Shavit, Y. (1990). Arab and Jewish minorities in Israeli Education. *American Sociological Review, 55*, 115–126.

Sprinzak, D., Bar, E., Levi-Mazloum, D., & Piterman, D. (2003). *Facts and figures*. State of Israel, Ministry of Education, Economic and Budgeting Administration, Jerusalem (in Hebrew).

Zarzure, S. (1995). The Arab education: Existing state and a look toward the future. In E. Peled (Ed.), *Fifty years of Israeli education: Book C* (pp. 1061–1083). Jerusalem: Ministry of Education, Culture and Sport and Ministry of Defense (in Hebrew).

Zuzovsky, R. (2001). *Learning outcomes and the educational context of mathematics and science teaching in Israel.* Findings from the Third International Mathematics and Science Study-TIMSS—1999. Tel Aviv: Ramot–Tel Aviv University.

Zuzovsky, R. (2005). *Achievement of 8th graders in mathematics and science and the educational context of teaching these subjects in Israeli schools*. Tel Aviv: Ramot–Tel Aviv University.

The paradoxical relationship between students' achievement and their self-perceptions: A cross-national analysis based on three waves of TIMSS data

Ce Shen
Boston College
Chestnut Hill, Massachusetts, USA

Hak Ping Tam
National Taiwan Normal University
Taipei, Taiwan

Abstract

This paper investigates the relationship between eighth-grade students' achievement and self-perceptions in mathematics and science by analyzing the three waves of the Trends in International Mathematics and Science Study (TIMSS) data. A total of three measures on self-perception were used, namely, how much students like the two subjects, their self-perceived competence in the subjects, and their perceived easiness of the subjects. For within-country data, with individual student as the unit of analysis, there is generally a positive correlation between students' achievement and their self-perception. But when the three self-perception measures are aggregated at the country level, the relationship is reversed. In other words, there is a negative correlation between self-perceptions and achievement on a between-country analysis with country as the unit of analysis. This pattern is consistent in both mathematics and science across all three waves of data, even though the sample sizes (number of countries) and the participating countries vary from wave to wave. One possible explanation for this finding is that high-performing countries have higher academic standards; their students have higher pressure to get into top-choice academic institutions by excelling in public examinations. Accordingly, students from these countries have better academic performances in science and mathematics on the average, but lower preference for these subjects.

Introduction

There is a wide range of difference in students' mathematics and science performance across countries in each wave of the Trends in International Mathematics and Science Study (TIMSS 1995, 1999, and 2003). As an illustration, for the eighth-grade students' results in TIMSS 2003, Singapore's average scores were as high as 605 and 578 for mathematics and science respectively, while South Africa's were 264 and 244 (Martin, Mullis, Gonzalez, & Chrostowski, 2004; Mullis, Martin, Gonzalez, & Chrostowski, 2004). Nonetheless, every school system is different with respect to its unique sociocultural and economic context, and this is why the task of explaining why there is so much cross-national variation in students' achievement in the two subjects must be conducted with caution. Challenging as the task may be, this study examined the relationship between students' achievement scores and their self-perception of the two subjects and explored some plausible explanations behind the vast cross-national variation in students' performance.

According to cognitive psychologists and motivation theorists (e.g., Bandura, 1994), students with positive attitudes toward learning and positive self-perceptions toward their competence level can lead to motivation, thereby enhancing students' academic achievement. Many empirical studies have tested these assumptions and generally support this hypothesized continuous feedback loop between people's self-evaluation, or self-efficacy beliefs, intrinsic interest, motivation, and accomplishment (Brown, Lent, & Larkin, 1989; Locke & Latham, 1990; Multon, Brown, & Lent, 1991; Schunk, 1989, 1991; Zimmerman & Bandura, 1994; Zimmerman, Bandura, & Martinez-Pons, 1992). However, these studies and motivation theories are basically rooted in western culture and social conditions. Moreover, self-conceptions can vary from culture to culture. In order to avoid culturally biased results, it is wise to consider the relationship between students' academic achievement and their self-perceptions cross-nationally by means of inspecting cross-national data.

Now that the TIMSS data are available, the hypothesized relationship can be tested out on a larger number of countries with different sociocultural and economic backgrounds. Seemingly, analysis at the country level can go beyond such psychological theories as motivation and self-efficacy. This is because many psychological theories operate at the individual level, not at the country level. We hoped that by analyzing aggregated data at the country level, we could garner insights with respect to explaining cross-national variation of students' academic achievement that may not be apparent by analyzing the data at the individual level alone.

Hypotheses

In this study, we tested two groups of null hypotheses. The first group was based on within-country data (with the individual student as the unit of analysis):

1. There is no correlation between students' mathematics and science achievement scores and the extent of how much they like the two subjects.
2. There is no correlation between students' mathematics and science achievement scores and their self-evaluation of their competence in these two subjects.
3. There is no correlation between students' mathematics and science achievement scores and their perceived easiness of the two subjects.
4. There is no correlation among the three measures of self-perception, namely, how much students like the two subjects, their self-evaluation of their competence in the two subjects, and their perceived easiness of the two subjects.

The second group of null hypotheses was based on between-country data (with country as the unit of analysis):

5. There is no correlation between students' mean mathematics and science achievement scores and their average self-perception about how much they like the two subjects.
6. There is no correlation between students' mean mathematics and science achievement scores and their average self-evaluation of their level of competence in these two areas.
7. There is no correlation between students' mean mathematics and science achievement scores and their average perceived easiness of the two subjects.
8. There is no correlation among students' average self-perceptions about how much they like the two subjects, their average self-evaluation of their competence in the two subjects, and their average perceived easiness of the two subjects.

Countries/school systems included in this study

Forty countries/school systems participated in the TIMSS 1995 study at the eighth-grade level. In the second wave, only 38 countries/school systems participated in the TIMSS 1999 study at the eighth-grade level. Of these 38 countries, 26 participated in TIMSS 1995. In the third wave, 46 countries/school systems participated in the TIMSS 2003 study at the eighth-grade level, including 35 countries/school systems that participated in one or both of the TIMSS 1995 and 1999 studies.

A noticeable difference between TIMSS 1995 (Beaton, Mullis, Martin, Gonzalez, Kelly, & Smith, 1996) and TIMSS 1999 and 2003 was that quite a few of the developed European countries in the 1995 study chose not to participate in the two later studies. These countries included Austria, Denmark, France, Germany, Greece, Iceland, Ireland, Portugal, and Switzerland. At the same time, a good many of the new participants joined in the later studies. Many of these are developing countries and regions, or in transition. In TIMSS 1999, the new participants included Chile, Chinese Taipei, Finland, Indonesia, Jordan, Macedonia, Malaysia, Moldova, Morocco, Philippines, Tunisia, and Turkey. In TIMSS 2003, the new participants included Armenia, Bahrain, Botswana, Chile, Egypt, Estonia, Ghana, Lebanon, Palestine, Saudi Arabia, Scotland, and Serbia and Montenegro.

Data, measurement, and methods

Using data from the three waves of TIMSS (1995, 1999, and 2003), this study investigated the relationship between Grade 8 students' mathematics and science achievement and three measures of their self-perceptions on these two subjects. For the TIMSS 1995 and 1999 data, the first measure was the response to the statement, "I like mathematics (or science)," which was used as an indicator of self-perceived attitude toward the two subjects. Responses were based on a Likert scale, the four points of which were 1 =

dislike a lot, 2 = dislike, 3 = like, and 4 = like a lot. As for the TIMSS 2003 data, the corresponding question was slightly changed to "I enjoy learning mathematics (or science)." The second measure was the response to the statement, "I usually do well in mathematics (or science)," which was used as an indicator of self-efficacy or self-perceived competence in mathematics and science. The third measure of self-perception was, for both the TIMSS 1995 and 1999 data, the response to the statement, "Mathematics (or science) is an easy subject," and was used as a proxy of self-perceived rigor of the subjects. As for the TIMSS 2003 data, this measure was modified to the statement, "I learn things quickly in mathematics (or science)." So far as the coding is concerned, the latter two measures of self-perception for the TIMSS 1995 and 1999 data were based on a four-point Likert scale and were both coded as 1 = strongly disagree, 2 = disagree, 3 = agree, and 4 = strongly agree. For the TIMSS 2003 data, all three measures of self-perceptions used the same four-point Likert scale, which, after recoding, was 1 = disagree a lot, 2 = disagree a little, 3 = agree a little, and 4 = agree a lot. We believe, nevertheless, that even though the questions for two of the three measures of self-perceptions and the response categories were somewhat modified throughout the three waves of the study, the underlying theoretical constructs remained basically the same.

It should be pointed out that, in some countries, Grade 8 science is taught as an integrated subject, whereas in others, it is taught separately as several science subjects, including physics, chemistry, biology, earth science, and environmental science. Thus, in the TIMSS study, two versions of students' science background questionnaires were prepared. While one version asked questions with respect to science being taught as an integrated subject, the other asked questions with respect to science being taught as several separate areas. Both versions contained the three items on self-perception in science as mentioned earlier. Students only responded to the version of questionnaire that matched the way science was taught in their schools. For the version with science being taught as an integrated subject, a mean for each student was computed across as many of the science areas as were taught in the school that he or she attended. In this way, a single variable with a value in the range of 1 to 4 was used for each measure of self-perception, regardless of which version of the questionnaires the students filled out.

We are well aware of the limitations of using the three measures as indicators of the three concepts—self-perceived attitudes toward the two subjects, self-perceived competence or self-efficacy in the two subjects, and self-perceived rigor of the subjects. Scales based on multiple items can provide more reliable measures of the concepts, but the TIMSS student background questionnaire in 1995 unfortunately did not ask a series of questions with which we could have developed scales to measure the three concepts. For TIMSS 2003, the International Study Center at Boston College did develop several indexes measuring students' confidence in their ability to learn mathematics and science, and students' value on mathematics and science (Martin et al., 2004; Mullis et al., 2004). However, in order to maintain consistency across the three waves of the study, we simply used a single item as an indicator for each concept as described above.

We also recognize that many factors may influence how students respond to the statements mentioned above, including their academic goals and aspirations, their parents' and/or teachers' expectations, academic standards, the rigor of curriculum, etc. Yet by using aggregate data to measure such concepts as self-perceived attitude toward the two subjects, self-efficacy, and self-perceived rigor of mathematics and science for each country, we move the unit of analysis from the student level to the country level. What is special about the TIMSS study is that it encompasses countries with very different cultural backgrounds, as well as different social, historical, and political backgrounds. Besides, teaching and learning are cultural activities (Kawanaka, Stigler, & Hiebert, 1999). Thus, under these circumstances, the meaning of the measures will be further affected by the different social and cultural contexts of the participating TIMSS countries. Moreover, examination of the data collected in the TIMSS curriculum analysis reveals substantial differences cross-nationally (Schmidt, McKnight, Valverde, Houang, & Wiley, 1997; Schmidt, Raizen, Britton, Bianchi, & Wolfe, 1997). All these cross-national differences need to be taken into account when examining the relationship between students' achievement and the three self-perception measures used in the present study. For example, "strongly agree" does not necessarily mean exactly the same thing in different languages and cultural contexts. Therefore, caution must be taken when drawing conclusions from cross-national comparisons based on these items. Nonetheless, these constraints should not preclude us from utilizing the data from this unprecedented large-scale study to test the possible relationship between students' achievement and the three self-perception concepts, provided that caution is exercised in terms of over-generalizing any findings.

The methodology of this study is relatively straightforward. For each wave of the TIMSS data, we performed two sets of correlation analyses. The first set of analyses, to test the first group of hypotheses, was the correlation analysis for within-country data at the individual student level. We examined, separately for each of the participating countries, the correlation between students' mathematics and science test scores and the three measures of their self-perception: how much they liked or enjoyed the two subjects, their self-perceived competence in the subjects, and their perceived easiness of the subjects. To test the second set of hypotheses, we performed correlation analyses at the country level (between-country analyses), with country as the unit of analysis. We analyzed the aggregate data for both test scores (country-level achievement scores) and the averages of self-perception measures at the country level.

Results of the analyses

The results of the analyses are reported in the order of the eight hypotheses stated earlier. We anticipated that because of the large sample size for each participating country, the correlation coefficients would almost certainly be significant. Therefore, the significance level is not reported for within-country data analysis. Readers are advised to look at the magnitude of the correlation coefficients as measures of the corresponding effect sizes. As regards the correlation coefficients for between-country data analyses, the significance level and sample size for each coefficient is reported simultaneously because the sample size is not large under this setting.

To save space, we include here only the Pearson correlation matrix of students' achievement scores on mathematics and science and the three measures of self-perception of TIMSS 2003 data (see Table 1). Those readers who are interested in the within-country correlation matrix for TIMSS 1995 and 1999 data should refer to Shen (2002) and Shen and Pedulla (2000). Even though the number of countries or school systems varies from wave to wave, the general pattern still holds.

Columns 1 and 2 of Table 1 report, for each of the 46 school systems, the correlation coefficients between students' achievement scores in the two subjects and their responses to how much they enjoyed the two subjects. To reiterate, for TIMSS 1995 and 1999, the corresponding question inquired how much students "liked" the two subjects rather than enjoyed them. As shown in the table, there is, within each country, a positive relationship between students' actual score and their enjoyment of the two subjects, with only one exception—Indonesia. For most countries, the correlation coefficients fall between .10 and .40, fluctuating from country to country and from mathematics to science. This indicates that within each participating country, students who reported enjoying or liking mathematics and science tended to have higher achievement in these areas than students who reported less enjoyment or liking of the two subjects. Even though the strength of relationship is not particularly strong, this result supports the conventional motivation theory discussed earlier. Thus, hypothesis 1 is rejected, and it is concluded that there is some evidence in support of the alternative hypothesis. It should be pointed out that since the sample design of TIMSS is a two-stage stratified cluster sample design, a jackknife procedure was used in this paper to take into account the clustering effect when the correlation coefficients and their standard errors were computed for each country.

Columns 3 and 4 of Table 1 present the Pearson correlation coefficients for students' achievement scores (for both mathematics and science) and their self-perceived competence in the two subjects. Again, there is, as shown, a positive relationship for all countries except for Indonesia. On average, the magnitudes of the coefficients are greater than those shown in Columns 1 and 2. For several school systems, the correlation coefficients are as high as .50 (Chinese Taipei, Korea, and Norway). Hence the strength of the relationship between the two variables range from low to medium effect sizes. These statistics indicate that, within each participating country, students who reported doing well in mathematics and science tended to have higher achievement in these areas than students who reported doing less well. This result

Table 1: Correlations between Achievement Scores of Mathematics and Science and Three Measures of Self-perception for Grade 8 Students in 46 School Systems Based on TIMSS 2003 Data (in alphabetical order)

Country (school system)	"I enjoy math"	"I enjoy science"	"I do well in math"	"I do well in science"	"I learn math quickly"	"I learn science quickly"
Armenia	0.166	0.106	0.173	0.145	0.210	0.169
Australia	0.222	0.192	0.395	0.274	0.391	0.249
Bahrain	0.156	0.055	0.259	0.121	0.316	0.161
Belgium (Flemish)	0.180	0.085	0.126	0.150	0.196	0.177
Botswana	0.203	0.296	0.113	0.101	0.055	0.122
Bulgaria	0.180	0.089	0.319	0.163	0.265	0.154
Chile	0.056	0.018	0.263	0.102	0.276	0.127
Chinese Taipei	0.462	0.274	0.513	0.333	0.452	0.274
Cyprus	0.304	0.145	0.468	0.211	0.420	0.208
Egypt	0.114	0.203	0.146	0.136	0.146	0.167
England	0.098	0.195	0.263	0.297	0.241	0.272
Estonia	0.175	0.053	0.440	0.198	0.425	0.168
Ghana	0.219	0.331	0.100	0.214	0.085	0.227
Hong Kong SAR	0.315	0.262	0.305	0.196	0.310	0.190
Hungary	0.250	0.094	0.452	0.218	0.475	0.217
Indonesia	0.030	-0.071	-0.122	-0.229	-0.055	-0.104
Iran	0.154	0.065	0.305	0.203	0.298	0.167
Israel	0.055	0.119	0.300	0.284	0.302	0.282
Italy	0.319	0.127	0.435	0.259	0.429	0.212
Japan	0.310	0.257	0.470	0.403	0.385	0.291
Jordan	0.150	0.089	0.225	0.115	0.258	0.193
Korea, Rep. of	0.397	0.294	0.565	0.438	0.478	0.345
Latvia	0.237	0.099	0.440	0.232	0.403	0.182
Lebanon	0.255	0.153	0.300	0.177	0.310	0.198
Lithuania	0.235	0.086	0.410	0.183	0.350	0.189
Macedonia	0.072	0.059	0.206	0.128	0.204	0.130
Malaysia	0.276	0.214	0.400	0.207	0.278	0.164
Moldova	0.192	0.097	0.238	0.167	0.260	0.156
Morocco	0.088	0.057	0.182	0.112	0.136	0.110
Netherlands	0.042	0.045	0.227	0.155	0.263	0.168
New Zealand	0.115	0.145	0.388	0.283	0.372	0.276
Norway	0.272	0.170	0.505	0.288	0.474	0.216
Palestine	0.196	0.161	0.289	0.261	0.294	0.244
Philippines	0.161	0.211	0.076	0.050	0.067	0.116
Romania	0.245	0.106	0.408	0.193	0.346	0.150
Russian Federation	0.266	0.098	0.445	0.270	0.414	0.217
Saudi Arabia	0.057	0.104	0.214	0.144	0.193	0.176
Scotland	0.058	0.226	0.329	0.402	0.279	0.389
Serbia & Montenegro	0.232	0.030	0.455	0.147	0.504	0.171
Singapore	0.275	0.221	0.333	0.208	0.342	0.221
Slovak Republic	0.207	0.067	0.382	0.195	0.443	0.198
Slovenia	0.152	0.072	0.398	0.216	0.430	0.198
South Africa	0.009	0.053	0.060	0.013	0.058	0.056
Sweden	0.164	0.152	0.446	0.204	0.405	0.238
Tunisia	0.215	0.092	0.250	0.132	0.294	0.189
United States	0.129	0.144	0.376	0.266	0.328	0.241

supports the self-efficacy theory discussed earlier and the conclusion reached by many prior research studies (Schunk, 1989; Zimmerman et al., 1992). Thus, hypothesis 2 is rejected, and it is concluded that there is evidence in support of the alternative hypothesis.

Columns 5 and 6 of Table 1 present the Pearson correlation coefficients for students' achievement scores (for both mathematics and science) and how quickly they learned the two subjects. For the TIMSS 1995 and 1999 data, the corresponding question was how easy they perceived the two subjects to be. The pattern and the magnitude of the Pearson's correlation coefficients are similar to those in Columns 1 to 4. We can therefore say that, for almost all countries, students who thought they learned quickly usually performed better in TIMSS than those who thought otherwise. Again, hypothesis 3 is rejected, and it is concluded that there is evidence in support of the alternative hypothesis.

For the TIMSS 2003 data with 46 participating school systems, the only country with a negative correlation between students' achievement and the three self-perception measures is Indonesia. For within-country data, throughout the three waves there is generally a positive relationship between students' achievement and the three measures of their self-perception. The general pattern of a positive association between students' achievement scores and their self-perception is consistent with conventional wisdom and supports existing psychological and motivation theories.

By and large, the relationship among the three measures of self-perception for TIMSS participating school systems/countries is stronger and more consistent than the relationship between students' achievement scores and their self-perception measures. Due to the similarity of the general pattern throughout the waves, and in order to save space, only four countries' correlation coefficients among the three measures of self-perception from TIMSS 2003 are reported in Table 2. The countries selected are Chile, Japan, Morocco, and the United States. They represent a wide spectrum of difference in performance levels and in cultural and geographical characteristics as well. While Japan is a high-performing country, the United States is at the middle range, and Chile and Morocco are relatively low-performing countries, all with different cultural and geographical backgrounds.

As shown in the table, the findings across the four countries are consistent. Despite the diverse achievement levels and variation in cultural and geographical factors, there is a clear positive relationship among the three measures of self-perception within each country, indicating that students who enjoyed or liked the subjects also perceived themselves as doing well, and thought that they learned the two subjects quickly or perceived the two subjects as easy. Thus, hypothesis 4 is rejected, and it is concluded that there is evidence in support of the alternative hypothesis.

Findings from the within-country data analysis support the conventional wisdom about the relationship of students' academic achievement and their attitudes toward the subjects, their self-perceived competence, and their perceived rigor of the two subjects.

The next phase of analysis concerns the aggregate data level, that is, country or school system level. Here, we investigated if the pattern found from the within-country analysis would still hold for the cross-national analyses. For all three waves of TIMSS data, the first step was to compute the mean mathematics and science achievement scores together with the means of the three measures of self-perception for each country. The correlation coefficients for the relevant pairs of variables were then computed at the country level. As mentioned earlier, each student responded to one of the two versions of the background questionnaire. For those students where science was taught as several separate subjects, a mean score for each student was computed across as many science subject areas as there were data.

Tables 3 and 4 present the correlation analysis results based on the aggregate data from TIMSS 1995 for each country's achievement scores in the two areas and the three measures of self-perception at the country level. It is interesting to note that the relationships are negative between the self-perception measures and the achievement scores in mathematics and science. Besides, the magnitudes of the correlation coefficients are moderately strong. The negative correlation stands in sharp contrast to the pattern of positive correlations found within the majority of countries as discussed above. The general pattern is that those countries where most students said they did not like mathematics and science, thought they usually did not do well in the two subjects, and perceived the subjects as difficult were usually high-performing countries, and vice versa.

Table 2: Correlations of Three Measures of Self-perception for Selected Countries (TIMSS 2003)

Chile (*N* = 6,286, using listwise deletion)				(*N* = 6,269, using listwise deletion)			
	1	2	3		1	2	3
1. "I enjoy learning math"	1.000			1. "I enjoy learning science"	1.000		
2. "I do well in math"	0.461	1.000		2. "I do well in science"	0.476	1.000	
3. "I learn math quickly"	0.482	0.566	1.000	3. "I learn science quickly"	0.506	0.538	1.000
Japan (*N* = 4,699, using listwise deletion)				(*N* = 4,781, using listwise deletion)			
	1	2	3		1	2	3
1. "I enjoy learning math"	1.000			1. "I enjoy learning science"	1.000		
2. "I do well in math"	0.415	1.000		2. "I do well in science"	0.448	1.000	
3. "I learn math quickly"	0.448	0.541	1.000	3. "I learn science quickly"	0.506	0.560	1.000
Morocco (*N* = 2,448, using listwise deletion)				*(N = 2,527, using listwise deletion)*			
	1	2	3		1	2	3
1. "I enjoy learning math"	1.000			1. "I enjoy learning science"	1.000		
2. "I do well in math"	0.400	1.000		2. "I do well in science"	0.319	1.000	
3. "I learn math quickly"	0.448	0.498	1.000	3. "I learn science quickly"	0.358	0.441	1.000
United States (*N* =8,592, using listwise deletion)				(*N* = 8,556, using listwise deletion)			
	1	2	3		1	2	3
1. "I enjoy learning math"	1.000			1. "I enjoy learning science"	1.000		
2. "I do well in math"	0.514	1.000		2. "I do well in science"	0.563	1.000	
3. "I learn math quickly"	0.518	0.661	1.000	3. "I learn science quickly"	0.581	0.655	1.000

We also note that the three self-perception measures correlate positively with one another. The TIMSS 1995 study also included data from Grades 3, 4, and 7. Although the number of participating countries was just under 40, the general pattern remained the same. To save space, we have not included them here.

Tables 5 and 6 present the correlation analysis results based on the aggregate data from TIMSS 1999 for each country's achievement in the two areas and the same three measures of self-perception at the country level. Thirty-eight countries (school systems) participated in the TIMSS 1999 study (Grade 8 students in most countries). Because no data were reported for the Netherlands on the extent to which their students liked the subjects, the sample size for some correlation coefficients dropped to 37. By inspecting the coefficients of Tables 5 and 6, a very similar pattern is found that corresponds to those shown in Tables 3 and 4 from the TIMSS 1995 data. The magnitudes of the correlation coefficients are slightly larger for the TIMSS 1999 data, which probably relates to the greater variation of the achievement scores that resulted from some European countries not participating in and more developing countries joining the 1999 study.

The TIMSS 2003 study had the largest number of participating countries in the history of the cross-national educational study. Tables 7 and 8 present the correlation analysis results based on the TIMSS 2003 aggregate data for each country's achievement in the two areas and the three similar measures of self-perception at the country level (Grade 8 students in most countries). As can be seen in these tables, the sizes of many coefficients are fairly strong.

To facilitate a visual interpretation of the data, we further provided six scatterplots to illustrate the relationship between the achievement scores and the three self-perception measures in the two subjects for the 46 participating school systems.

Figure 1 is the scatterplot of mathematics achievement scores versus the eighth grade students' responses to "I enjoy learning mathematics" for the TIMSS 2003 participating countries. The Pearson correlation coefficient amounted to a high r = -.708, p <.001, N = 46. The figure shows that the few top mathematics-performing school systems (upper-left hand corner of the figure), such as Chinese Taipei, Hong

Table 3: Correlations between International Mathematics Achievement Scores and Three Measures of Self-perception for Grade 8 Students at the Country Level (N = 40 Countries/School Systems; TIMSS 1995 Data)

	1	2	3	4
1. Mathematics score	1.00			
2. "I like math"	-.44**	1.00		
3. "I usually do well in math"	-.56**	.45**	1.00	
4. "Math is an easy subject"	-.63**	.59**	.53**	1.00

Note: ** *p* < 0.01.

Table 4: Correlations between International Science Achievement Scores and Three Measures of Self-perception for Grade 8 students at the Country Level (N = 40 Countries/School Systems; TIMSS 1995 Data)

	1	2	3	4
1. Science score	1.00			
2. "I like science"	-.41**	1.00		
3. "I usually do well in science"	-.37*	.37*	1.00	
4. "Science is an easy subject"	-.62**	.56**	.61**	1.00

Note: * *p* < 0.05; ** *p* < 0.01.

Table 5: Correlations between International Mathematics Achievement Scores and Three Measures of Self-perception for Grade 8 Students at the Country Level (TIMSS 1999 Data)

	1	2	3	4
1. Mathematics score	1.00			
2. "I like math"	-0.68 **	1.00		
	37			
3. "I usually do well in math"	-0.60 **	0.61 **	1.00	
	38	37		
4. "Math is an easy subject"	-0.72 **	0.87 **	0.65 **	1.00
	38	37	38	

Notes: ** *p* < 0.01.
The number below the correlation coefficient is the number of school systems (countries).

Table 6: Correlations between International Science Achievement Scores and Three Measures of Self-perception for Grade 8 Students at the Country Level (TIMSS 1999 Data)

	1	2	3	4
1. Science score	1.00			
2. "I like science"	-0.56 **	1.00		
	37			
3. "I usually do well in science"	-0.44 **	0.61 **	1.00	
	38	37		
4. "Science is an easy subject"	-0.74 **	0.73 **	0.71 **	1.00
	38	37	38	

Notes: ** *p* < 0.01.
The number below the correlation coefficient is number of school systems (countries).

Table 7: Correlations between International Mathematics Achievement Scores and Three Measures of Self-perception at the Country Level (N = 46 Countries/School Systems; TIMSS 2003 Data)

	1	2	3	4
1. Mathematics scoreq	1.00			
2. "I enjoy math"	-0.71 **	1.00		
3. "I do well in math"	-0.64**	0.53**	1.00	
4. "I learn math quickly"	-0.70**	0.67**	0.87**	1.00

Note: ** p < 0.01.

Table 8: Correlations between International Science Achievement Scores and Three Measures of Self-perception at the Country Level (N = 46 Countries/School Systems; TIMSS 2003 Data)

	1	2	3	4
1. Science score	1.00			
2. "I enjoy science"	-0.71 **	1.00		
3. "I do well in science"	-0.65**	0.53**	1.00	
4. "I learn science quickly"	-0.74**	0.77**	0.94**	1.00

Note: ** p < 0.01.

Kong SAR, Japan, Korea, and the Netherlands, have a relatively low level of enjoyment of the subject, while low-performing countries, such as Botswana, Egypt, Ghana, Morocco, and South Africa, generally have a high level of enjoyment in learning mathematics.

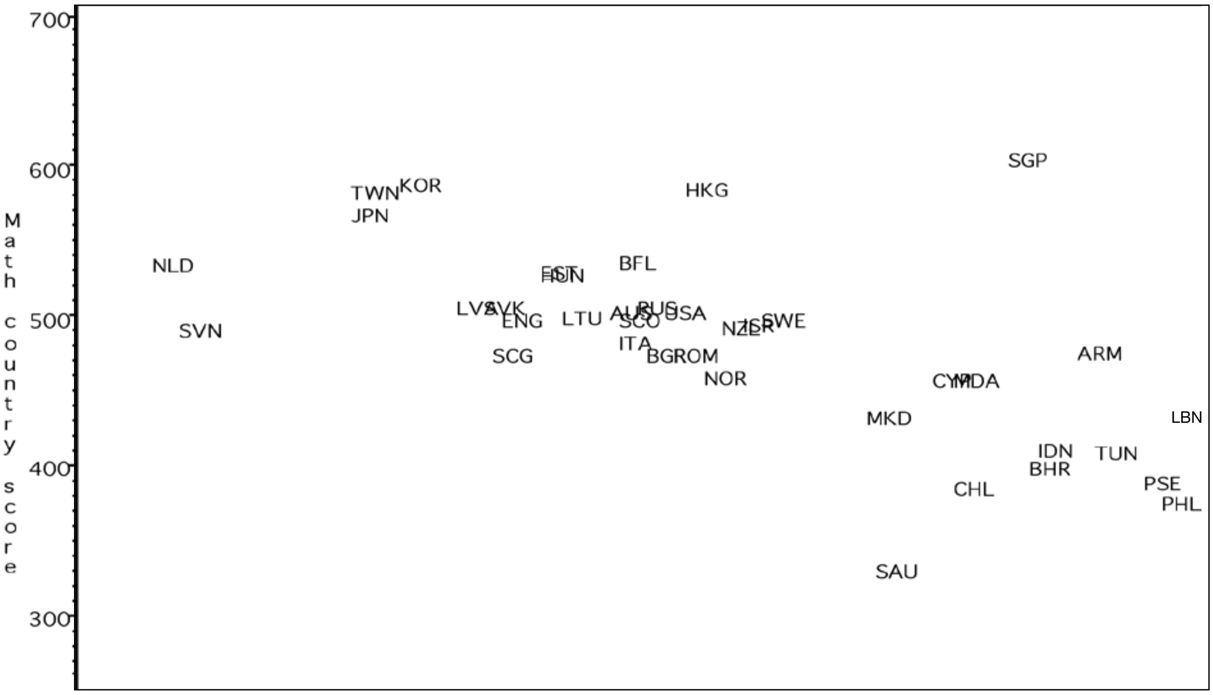
Figure 2 is the scatterplot of science achievement scores versus the eighth-grade students' responses to the statement "I enjoy learning science" for the TIMSS 2003 participating countries. The Pearson correlation coefficient also amounted to a high $r = -.706, p <.001, N = 46$. Again, students in the top-performing countries, which included Chinese Taipei, Hungary, Korea, Netherlands, and Slovenia and Montenegro, generally indicated a low level of enjoyment in learning science, while students in the low-performing countries, including Botswana, Ghana, the Philippines, and South Africa, had the highest level of enjoyment in learning science.

Figure 3 is the scatterplot of mathematics achievement versus "I usually do well in mathematics" for countries that participated in the TIMSS 2003 Grade 8 study. The corresponding Pearson correlation coefficient was $r = -.643, p <.001, N = 46$. Students in the top-performing school systems, such as Chinese Taipei, Korea, and Japan, reported the lowest level of self-efficacy, while students in the bottom-performing-countries, such as Ghana, Saudi Arabia, and South Africa, reported a relatively higher level of self-efficacy. In Figure 4, the scatterplot of science achievement versus "I usually do well in science" also demonstrates a similar pattern. The corresponding Pearson correlation coefficient was $r = - 0.648, p <.001, N = 46$.

Figure 5 is the scatterplot of mathematics achievement versus "I learn things quickly in mathematics." The corresponding Pearson correlation coefficient amounted to $r = -.696, p <.001, N = 46$. As mentioned earlier, we used this measure as a proxy of the perceived rigor of the program. As reflected in the figure, Chinese Taipei, Hong Kong SAR, Japan, and Korea are high-performing school systems but their students were those most likely to feel that they did not learn things quickly in mathematics. In contrast, the students from a number of low-performing countries (lower-right hand corner of the figure) were those most likely to think they learned things quickly in mathematics. Figure 6 has a very similar pattern to that of Figure 5. The corresponding Pearson correlation coefficient is a high $r = - 0.737, p <.001, N = 46$.

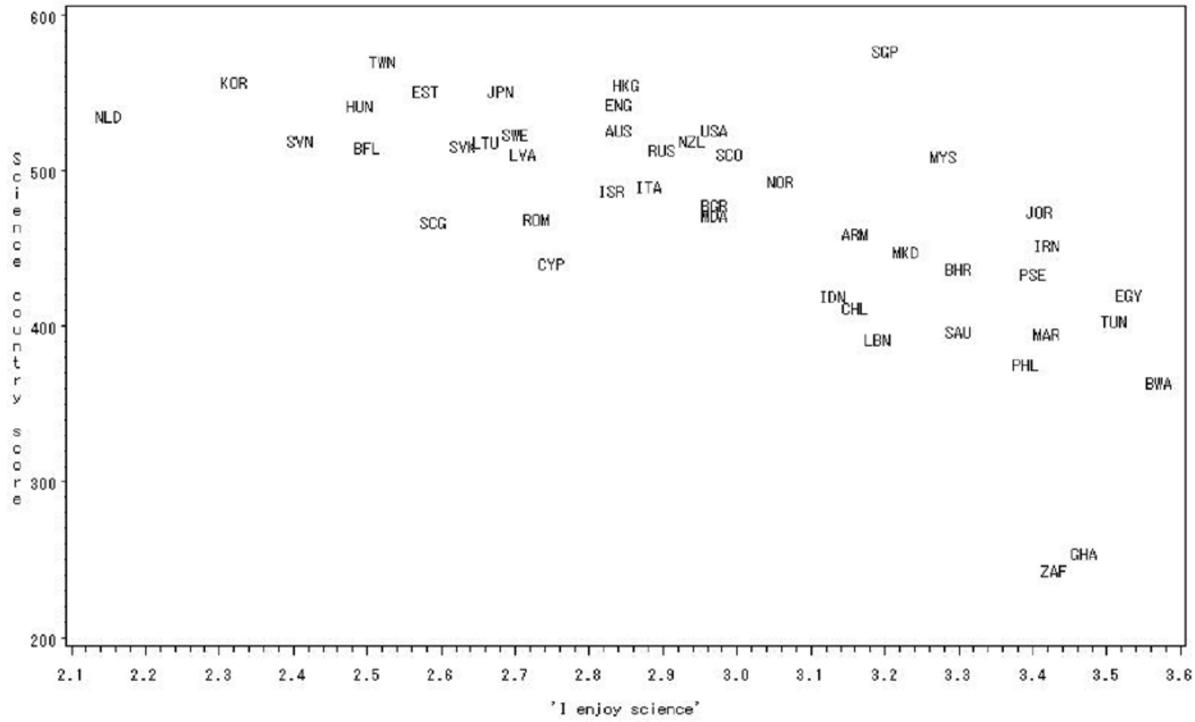
Notice that the correlation coefficients among the three aggregate measures of self-perception in Tables 7 and 8 are all positive and are moderate to strong in terms of magnitude. Based on the summary results from Tables 3 through 8 and the six scatterplots, we reject the null hypotheses 5 to 8 and conclude that there is some evidence in support of the respective alternative hypotheses.

Figure 1: Scatterplot of Mathematics Achievement versus "I Enjoy Mathematics" for Grade 8 TIMSS 2003 Participating Countries



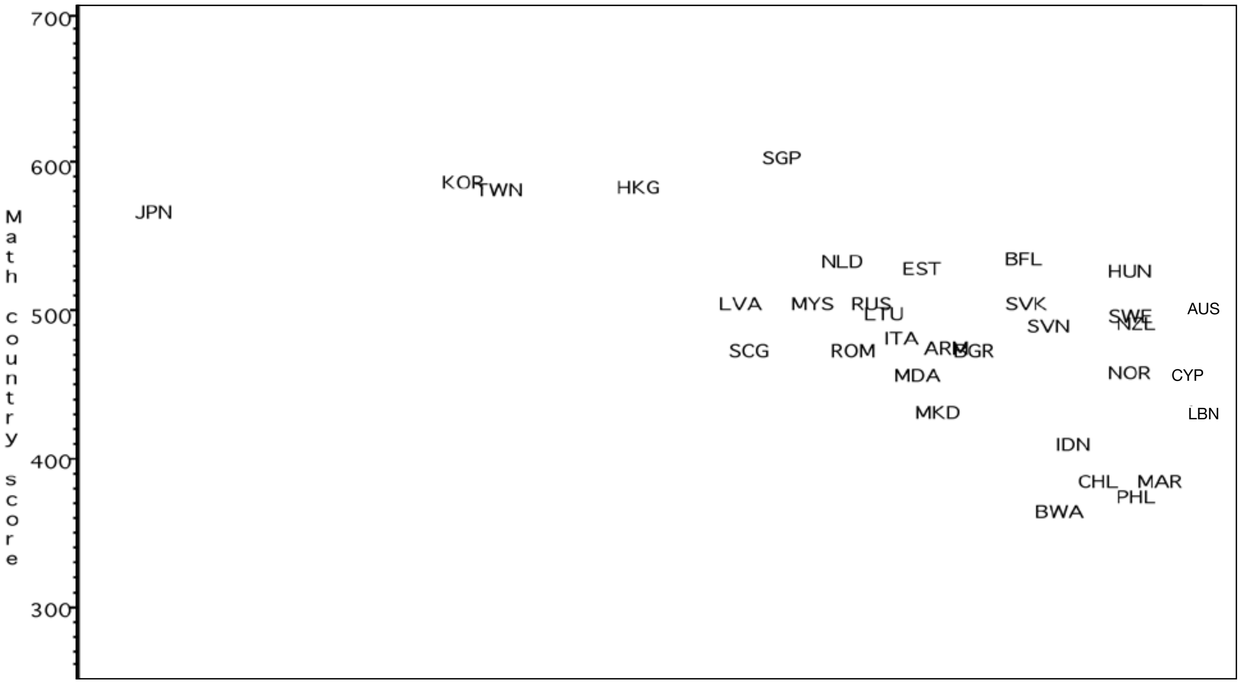
Note: Pearson's correlation = -0.708 (*p* < 0.001, *N* = 46).

Figure 2: Scatterplot of Science Achievement versus "I Enjoy Science" for Grade 8 TIMSS 2003 Participating Countries



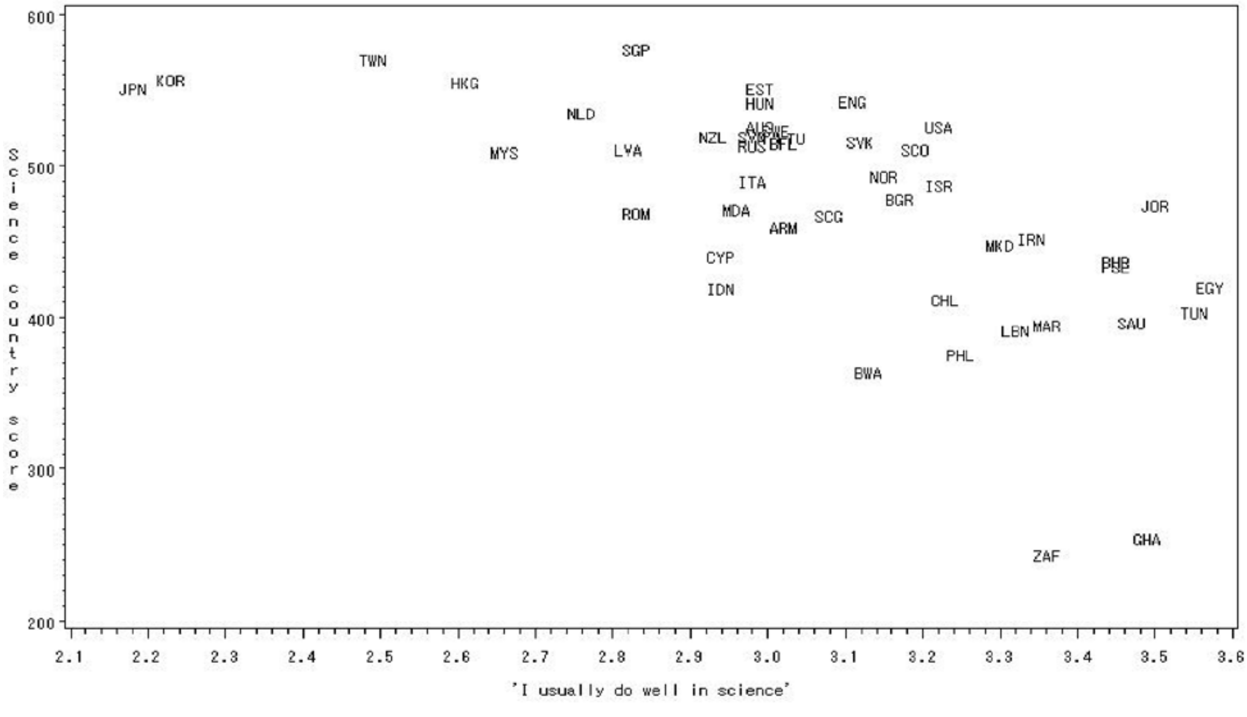
Note: Pearson's correlation = - 0.706 (*p* < 0.001, *N* = 46).

Figure 3: Scatterplot of Mathematics Achievement versus "I Usually Do Well in Mathematics" for Grade 8 TIMSS 2003 Participating Countries



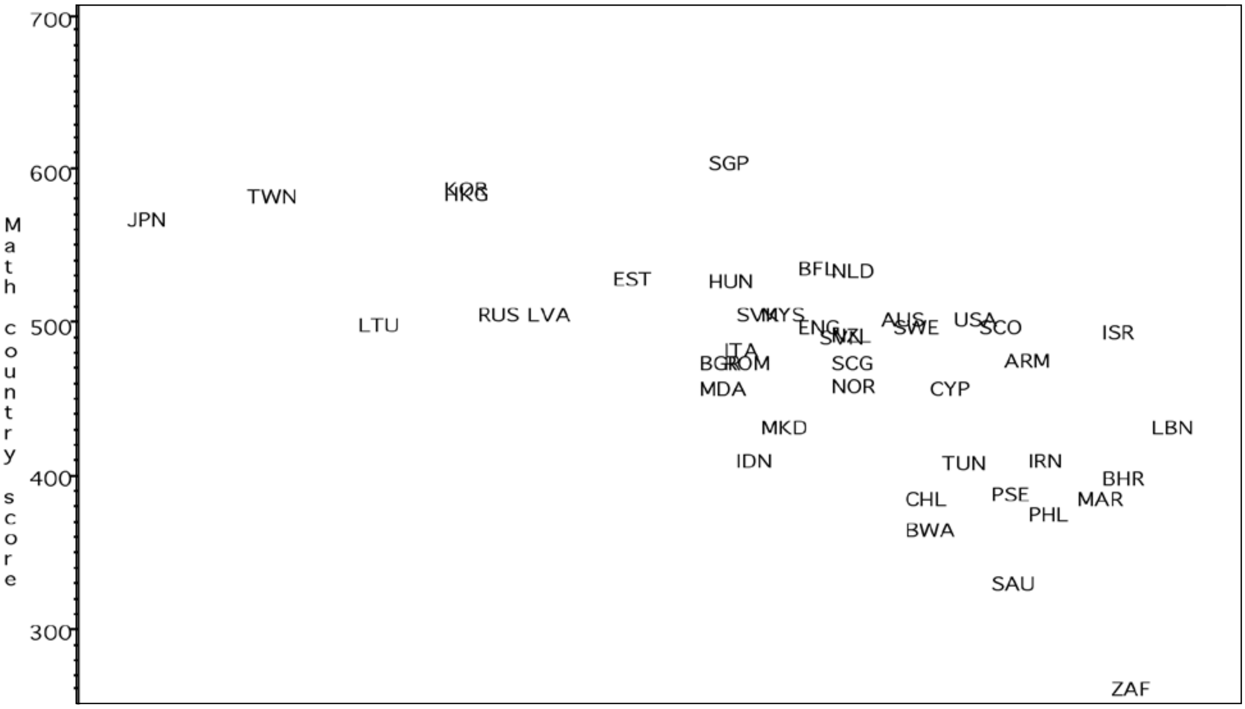
Note: Pearson's correlation = - 0.643 ($p < 0.001$, $N = 46$).

Figure 4: Scatterplot of Science Achievement versus "I Usually Do Well in Science" for Grade 8 TIMSS 2003 Participating Countries



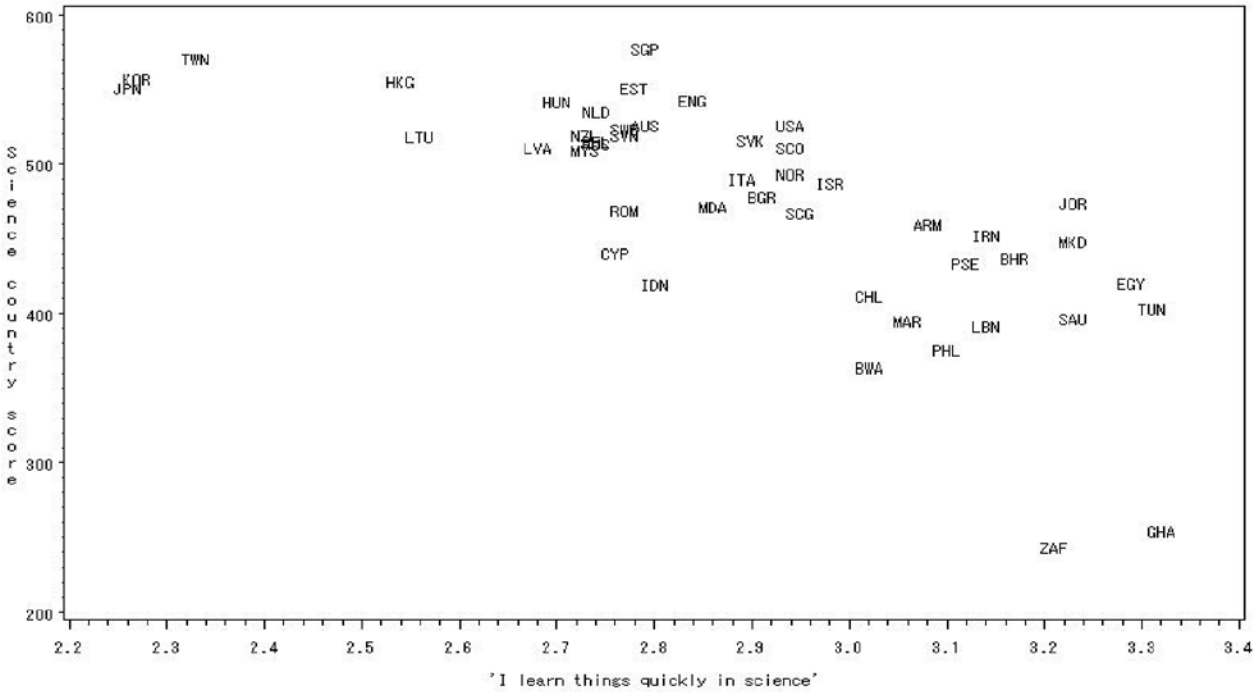
Note: Pearson's correlation = - 0.648 ($p < 0.001$, $N = 46$).

Figure 5: Scatterplot of Mathematics Achievement versus "I Learn Things Quickly in Math" for Grade 8 TIMSS 2003 Participating Countries



Note: Pearson's correlation = - 0.696 ($p < 0.001$, $N = 46$).

Figure 6: Scatterplot of Science Achievement versus "I Learn Things Quickly in Science" for Grade 8 TIMSS 2003 Participating Countries



Note: Pearson's correlation = - 0.737 ($p < 0.001$, $N = 46$).

In summary, for between-country analyses with country as the unit of analysis, there is a negative relationship between each self-perception measure and each achievement score. These findings were consistent for both mathematics and science across all three waves of TIMSS data, even though the sample sizes (number of countries) and the participating countries varied from wave to wave.

Discussion and conclusion

The existing motivation and self-efficacy theories suggest that there is a positive feedback loop among students' academic achievement, their self-evaluation, and their intrinsic interest in the subjects; the results found in the within-country analyses support this. This is, however, in sharp contrast to the consistent finding that negative association between students' achievement and self-perception exists at the country level across the three waves of TIMSS data. The two opposite patterns jointly form an interesting and paradoxical phenomenon, for which there are no ready theories and easy explanations. Of course, one should not interpret the findings from the study as encouraging students to develop negative attitudes toward mathematics and science and to decrease their self-perceived competence in order to raise their achievement. One cannot interpret causal implication out of correlational information; that would be committing an ecological fallacy. The negative relationship is found at the country level, not at the individual level. Besides, one should be careful with the interpretation at the country level. By the same token, the negative correlations found in the between-country analyses do not contradict the existing motivation and self-efficacy theories. These theories, as mentioned earlier, operate at the individual level, not at the country or culture level. The aggregate measures of students' self-perceptions represent overall information and are different from characteristics at the individual level. They reflect a specific country's educational, social, and cultural contexts, which contribute toward shaping the attitudes, values, and beliefs of some individuals in that country.

As mentioned at the beginning of the paper, it is widely assumed that a positive self-regard is an important motivating force that helps to enhance people's achievement. However, some researchers argue, with reference to cross-cultural studies, that the need for self-regard is culturally diverse and that the perception of oneself and regard for oneself differ across cultures. For example, Heine, Lehman, Markus, and Kitayama (1999) observed that anthropological, sociological, and psychological analyses revealed many elements of Japanese culture that are incongruent with such motivations. Instead, a self-critical focus is more characteristic of the Japanese, and that the need for positive self-regard is rooted more significantly in the North American culture. The results from this study also suggest that students in East Asian countries share something in common in terms of self-perceptions. This common ground may perhaps be attributable to their sharing basically the same Confucian root. Similarly, other research has found that East Asian people are, due to cultural reasons, more likely than people from other cultures to "reduce" themselves in relation to other people (see, for example, Stigler, Smith, & Mao, 1985; Uttal, Lummis, & Stevenson, 1988). However, given this special background of students from East Asian countries, we found that removing them from our analysis (not presented in this paper) changed only slightly the magnitude of the correlation coefficients, but did *not* change the overall pattern.

The fact that consistent negative correlations were found at one level but positive correlations were found at the other level among the same three measures of self-perception is reason enough to justify the search for a more coherent and holistic explanation. Shen and Pedulla (2000) put forward a plausible explanation for the negative correlations between students' achievement in mathematics and science and their sense of self-efficacy together with their perceived easiness of the two subjects. They suggested that low-performing countries might have relatively lower academic demands and expectations, whereas high-performing countries might have higher academic demands and expectations. In particular, the aggregate measure of students' perceived easiness of mathematics and science may reflect the corresponding strength of the curriculum of that country. On the other hand, countries with a demanding curriculum and high academic standards in mathematics and science may turn out students with high academic achievement levels.

Since the comprehensive analyses by Schmidt, McKnight, Valverde, Houang, and Wiley (1997) and Schmidt, Raizen, Britton, Bianchi, and Wolfe (1997) on the curricula from countries participating in the TIMSS study did not provide a list of rankings in accordance to the strength of the curricula, it is as yet impossible to verify the explanation suggested by Shen and Pedulla (2000). However, their proposed explanation is consistent with the findings from a number of small-scale comparative studies in the past (cf. Becker, Sawada, & Shimizu, 1999; Stevenson & Stigler, 1992; Stigler et al., 1985). It is also consistent with the findings from the TIMSS videotape study (Kawanaka et al., 1999), which examined eighth-grade mathematics classrooms in Germany, Japan, and the United States. The convergent findings from the three waves of TIMSS data furnish further evidence in favor of the academic strength explanation.

We suggest that the negative correlations are, to a certain extent, indicative of the overall relationship between the rigor of the academic standards and expectations on the achievement of students in mathematics and science. Students from such high-performing countries as Japan and South Korea usually indicate a relatively low level of enjoyment or liking of the two subjects and a lower level of self-evaluation, and they perceive the two subjects as hard and not easily or quickly learned. Conversely, students from low-performing countries, such as South Africa and Morocco, tend to indicate that they enjoy learning the two subjects, do fairly well in them, and consider the subjects as easy and ones in which they can learn things quickly. For some middle school students, high academic expectations or standards may stimulate their intrinsic interest in learning, but, for many others, demanding standards and rigorous curriculum may lead to resentment toward the two subjects. In countries where the expectation is low, the students, unlike their high-performing foreign counterparts, might have less motivation and set lower goals to improve their performance since they perceive their performance to be fairly acceptable already. If they believe that they are doing well and that mathematics and science are easy for them, they would see no need to study harder in these areas or to invest greater effort in them.

We believe that the policy implication from this study and similar previous studies points to the benefit of gradually raising the academic standards and expectation in countries, including the United States, where performance is, relatively speaking, mediocre or dissatisfactory. However, a country's specific historical, sociocultural, and economic environment affects and even constrains its academic standards and curriculum. Therefore, we do not imagine the achievement problem can be solved by simply copying a rigorous curriculum. An understanding of the prevalent beliefs and attitudes with respect to education in a specific society and culture is deemed necessary in order for such a policy to be beneficial and effective.

When we examine the six scatterplots, we find that although all the negative correlations are statistically significant, there may be trends other than the linear one. Although the several East Asian school systems did well in TIMSS, and their aggregate level of self-perceptions were relatively low, there was also variation within this group. For example, students in Japan and South Korea typically had the lowest aggregate level of enjoyment of the two subjects and the lowest level of self-evaluation. Furthermore, they both perceived the two subjects as being difficult. In comparison, the average level of self-perceptions by the top-performing Singaporean students was more positive than those of the Japanese and South Korean students. Singapore's school system might therefore be a better model for other school systems to follow than those of Japan and South Korea. An investigation of how middle school students in Singapore develop such a relatively positive attitude toward the two subjects and confidence in their ability despite the rigor of the curricula is beyond the scope of this study, but it is worth undertaking. Last, but not least, we should point out that the variation as evident in this study suggests that there is limitation in using academic standards alone to explain the negative correlation between achievement and self-perception at the country level. For a full explanation of this paradoxical relationship, we will need to examine other cultural and social factors as well. For example, in recent years, many countries around the world have taken on various reform efforts and policy changes in education. Hence it will be interesting to see if the patterns reported here will again show up in the TIMSS 2007 study. Should there be any changes; one can then investigate to see if such changes can be attributed to the reform efforts undertaken.

References

Bandura, A. (1994). Self-efficacy. *Encyclopedia of Human Behavior, 4*, 71–81.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study* (*TIMSS*). Boston: Center for the Study of Testing, Evaluation and Educational Policy, Boston College.

Becker, J. P., Sawada, T., & Shimizu, Y. (1999). Some findings of the US-Japan cross-cultural research on students' problem-solving behaviors. In G. Kaiser, E. Luna, & I. Huntley (Eds.), *International comparisons in mathematics education* (pp. 121–150). London: RoutledgeFalmer.

Brown, S. D., Lent, R. W., & Larkin, K. C. (1989). Self-efficacy as a moderator of scholastic aptitude: Academic performance relationship. *Journal of Vocational Behavior, 35*, 64–75.

Heine, S. J., Lehman, D. R., Markus, H. R., & Kitayama, S. (1999). Is there a universal need for positive self-regard? *Psychological Review, 106*(4), 766–794.

Kawanaka, T., Stigler, J. W., & Hiebert, J. (1999). Studying mathematics classrooms in Germany, Japan and the United States: Lessons from the TIMSS videotape study. In G. Kaiser, E. Luna, & I. Huntley (Eds.), *International comparisons in mathematics education* (pp. 86–103). London: RoutledgeFalmer.

Locke, E. A., & Latham, G. P. (1990). *A theory of goal-setting and task performance*. Englewood, NJ: Prentice-Hall.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international science report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Boston: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Boston: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology, 38*(1), 30–38.

Schmidt, W. H., McKnight, C. C., Valverde, G. A., Houang, R. T., & Wiley, D. E. (1997). *Many visions, many aims* (Vol. 1). Boston: Kluwer Academic Publishers.

Schmidt, W. H., Raizen, S. A., Britton, E. D., Bianchi, L. J., & Wolfe, R. G. (1997). *Many visions, many aims* (Vol. 2). Boston: Kluwer Academic Publishers.

Schunk, D. H. (1989). Social cognitive theory and self-regulated learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theory, research, and practice* (pp. 83–110). New York: Springer Verlag.

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist, 26*, 207–231.

Shen, C. (2002). Revisiting the relationship between students' achievement and their self-perceptions: A cross-national analysis based on TIMSS 1999 data. *Assessment in Education: Principles, Policy and Practice, 9*(2) 161–184.

Shen, C., & Pedulla, J. J. (2000). The relationship between students' achievement and self-perception of competence and rigor of math and science: A cross-national analysis. *Assessment in Education: Principles, Policy and Practice, 7*(2) 237–253.

Stevenson, H. W., & Stigler, J. W. (1992). *The learning gap*. New York: Simon & Schuster.

Stigler, J. W., Smith, S., & Mao, L. W. (1985). The self-perception of competence by Chinese children. *Child Development, 56*, 1259–1270.

Uttal, D. H., Lummis, M., & Stevenson, H. W. (1988). Low and high mathematics achievement in Japanese, Chinese, and American elementary-school children. *Developmental Psychology, 24*, 335–342.

Zimmerman, B. J., & Bandura, A. (1994). Impact of self-regulatory influences on writing course attainment. *American Educational Research Journal, 31*, 845–862.

Zimmerman, B. J., Bandura, A., & Martinez-Pons, M. (1992). Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal, 29*, 663–676.

A residual analysis of effective schools and effective teaching in mathematics

Constantinos Papanastasiou
University of Cyprus
Nicosia, Cyprus

Abstract

The Trends in Mathematics and Science Study (TIMSS) is the largest and most ambitious study undertaken by the International Association for the Evaluation of Educational Achievement (IEA). TIMSS provides a tool for investigating student achievement and school effectiveness, taking into account the varying influences of instructional contexts and practices and home environment. Schools vary widely in terms of the average achievement of their students in mathematics. Thus, it is of great interest for policymakers worldwide to identify factors that distinguish higher performing schools from lower performing schools. The aim of the analysis was to find indicators related to schools that differentiate between these two groups of schools. For this study, a more effective school was one where the school achievement score was higher than the score that would be predicted from the student characteristics. Data were obtained from 3,116 students, a number that represented 31.8% of the entire population (9,786). Analysis of the differences between the predicted and achieved scores led to identification of schools that performed better than would be expected given the home circumstances of their students. From this analysis, six factors were found to account for school differences that relate to mathematics achievement. The factor that accounted for the greatest differences between the more effective and less effective schools was passive learning, while the second factor was active learning. The third related to self-perception, and the fourth factor was student attitudes toward mathematics. The remaining two factors were family incentives and class climate.

Introduction

Mathematical skills are critical to the economic progress of a technologically based society, which is why many countries question what their school-age populations know and can do in mathematics. More specifically, they want to know what concepts students understand, how well they can apply their knowledge to problem-solving situations, and whether they can communicate their understanding. Of even greater importance is their desire to know what they can do to improve students' understanding of mathematical concepts, their ability to solve problems, and their attitudes toward learning (Beaton, Mullis, Martin, Gonzales, Kelly, & Smith, 1996). Mathematics achievement is a significant factor in decisions concerning placement, promotion, and selection in almost all education systems (Nasser & Birenbaum, 2005), and its importance is confirmed by the number of countries that participate in international mathematics studies like those conducted by the International Association for the Evaluation of Educational Achievement (IEA) and the Organisation for Economic Co-operation and Development (OECD). The findings of these international studies, as well as national surveys, are valuable tools for educators and policymakers (Grobler, Grobler & Esterhuyse, 2001; Nasser & Birenbaum, 2005; Secada, 1992).

The Trends in International Mathematics and Science Study (TIMSS) is one of the most ambitious series of studies undertaken by the IEA. TIMSS provides a tool to investigate both student achievement and school effectiveness, taking into account the varying influences of instructional contexts, practices, and home environment. The study's global focus and its comparative perspective give educators valuable insight into what is possible beyond the confines of their national borders. Data from TIMSS make it possible to examine differences in current levels of performance in relation to a wide variety of variables associated with the classroom, school, and national contexts within which education takes place. Because the IEA studies present objective information on student performance from different countries and cultures, international and national policymakers and educators are provided with an important data source. Data from IEA studies provide solid evidence for the feasibility and efficacy of educational policies, curriculum, and teaching practices (Mullis et al., 2000). Most international studies on education focus on students' academic outcomes, and the main reason for this is societal demand for academic achievement (Gadeyne, Ghesquiere, & Onghena, 2006). Such studies also give direction to policymakers who want to identify the characteristics of schools so they can more effectively plan improvement strategies (Brown, Duffield, & Riddell, 1995).

School effectiveness research has flourished since 1979, and has attracted considerable political support in several countries (Luyten, Visscher, & Witziers, 2005). Studies on school effectiveness identify school characteristics that optimize particular learning outcomes, school improvement factors, and processes that establish these effectiveness-enhancing factors (Scheerens & Demeuse, 2005). This kind of research aims to tease out the factors that contribute to effective education and especially those that schools can implement (Creemers & Reezigt, 2005). Research on school effectiveness pinpoints those characteristics or factors that are important for effectiveness at different levels of the system (i.e., student, learning, teaching, and school), and identifies school achievement in relation to basic cognitive skills. School effectiveness research also highlights the characteristics of schools and classrooms that are associated with differences in school effectiveness. If we know the particular characteristics of an effective school, especially those relating to the sphere of features that could be changed, then we are in a position to improve underperforming schools by encouraging them to adopt those characteristics (Luyten et al., 2005). Another objective of school effectiveness research is to increase the potential that schools have to improve education and especially educational achievement. In other words, school effectiveness research aims to find out what works in education and why (Creemers & Reezigt, 2005).

Many researchers have focused on the fact that the composition of the student body has a substantial impact on achievement over and beyond the effects associated with students' individual abilities and social class. Other researchers support the claim that schools with low social-class intakes have certain disadvantages associated with their context (Baumert, Stanat, & Watermann, 2005; Opdenakker & Van Damme, 2005; Van de Grift & Houtveen, 2006; Wilms, 1992). Other studies argue that the factors that most influence performance are the teaching and learning process and the creation of a learning environment. They argue that schools with high achievement are characterized by clear, well-organized teaching that motivates students and connects to their background knowledge and that keeps students actively involved in the learning process and their lessons—lessons that are efficiently organized and well structured. Recent research reveals that the main problem of underperforming schools is that their students are not given sufficient opportunity to attain the minimum objectives of the curriculum. Van de Grift and Houtveen (2006), for example, point to mathematics textbooks that are not suitable for attaining the basic objectives of the curriculum, insufficient time allotted for learning and teaching, and teaching that is poor and does not stimulate students. These two researchers also found that student performance in underperforming schools improved when the teaching was improved, the class was better organized, and the students were kept actively involved.

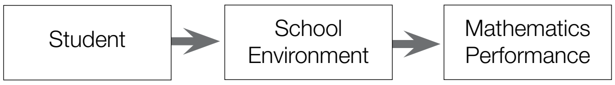
A study by Stoll and Wikeley (1998) indicates that school improvement efforts in recent years have increasingly focused on effectiveness issues such as the teaching and learning processes and student outcomes. This focus on school improvement has led to more research into the factors that make a school effective (MacBeath & Mortimore, 2001; Reynolds & Stoll, 1996). For school improvement to be successful, certain characteristics of the school atmosphere must be favorable. For example, a school and its students must have common goals and the school must feel responsible for its students' success. Other requirements are mutual respect and support and a positive attitude toward learning (Creemers & Reezigt, 2005). Behavioral theorists agree that schools will not change if the staff within the schools—the teaching staff especially—do not change. Three mechanisms that can bring about change are evaluation, feedback, and reinforcement. These mechanisms explain and can therefore be used to improve effective instruction in the classroom (Creemers, 1994).

Schools vary in terms of their students' average achievement in mathematics. In general, the student intakes of schools produce differences in outcome that are not caused by school processes. For this reason, it is necessary, before comparing schools, to correct for student intake. Factors considered relevant in this respect are the socioeconomic status and the educational background of the students' families. School performance is usually expressed in terms of average student achievement by school. These measures ideally include adjustments for such student characteristics as entry-level achievement and socioeconomic status in order to determine the value added by a school. Their main goal is to identify the factors that lead to the best results (Luyten et al., 2005). The student populations of schools can differ considerably in the proportion of students from homes with particular characteristics. The extent to which and how the home situation affects educational achievement has received much attention (Papanastasiou, 2000, 2002; Schreiber, 2002). Moreover, the extent to which schools vary in effectiveness and the school factors that seem to promote effectiveness are academically interesting.

Figure 1 presents a simple model that clearly illustrates how mathematics performance is influenced indirectly by intake input and directly by the school environment. The educational background of the family, the size of a student's home library, and the socioeconomic status of the family are three factors that could be included in the school intakes category. These factors are assumed to have a direct impact on processes within the school as well as on the general performance of the school.

With its basis in school effectiveness research, this present study investigated achievement in schools in relation to the factors that enhance school performance. The main research question is: *Why do students at some schools learn much less than would be expected on the basis of their family background?* The aim of the study was to find out whether a set of indicators from the student TIMSS questionnaire for Grade 8 of the lower secondary school was responsible for differences in academic achievement. In other words, we tried to find out if specific characteristics are associated with students' academic achievement.

Figure 1: Influence of Student Intake and School Environment on Mathematics Performance



Method

The study focused on the TIMSS 1999 sample, the population of which included all students enrolled in Grade 8 in the 1998/99 school year. The participating students completed questionnaires on home and school experiences related to learning mathematics, and school administrators and teachers answered questionnaires regarding instructional practices (Beaton et al., 1996). In Cyprus, all 61 gymnasia participated in this project (the entire population of schools), with two Grade 8 classes from each school. Within each class, all students were tested, and achievement tests and other data were obtained. Data were obtained from 3,116 students, which represented 31.8% of the entire population (9,786). However, among those students, the responses of only those who had completed all the questions were used. This led to listwise deletion of some subjects from the data set. The average age of students tested was 13.8 years.

The study analyzed data from the student questionnaire and the mathematics tests to find certain school indicators that differentiate between more effective and less effective schools. For this study, a more effective school was designated as one where the school achievement score was higher than the mean score predicted from the student characteristics (Postlethwaite & Ross, 1992). In the same way, a less effective school was one for which the school mean in mathematics was lower than the mean expected. Based on the differences between the predicted scores and the actual scores, the residuals that distinguish the more effective from the less effective schools were identified. Analysis of the differences between the predicted and the achieved scores in terms of school quality led to identification of schools that performed better than would be expected given the home circumstances of their students. In total, seven steps were followed during identification of the factors distinguishing the more from the less effective schools (Postlethwaite & Ross, 1992).

Step 1: This first step involved identifying the measures related to home characteristics, given these are thought to affect student achievement. Three factors were identified from the TIMSS student questionnaire: the economic status of the family, the educational background of the family, and the size of the home library. For the first factor, 13 measures related to the economic status of the family, six of which concerned items or facilities in the students' homes: central heating, washing machine, air-conditioning, more than two cars, computer, and more than four bedrooms. The second factor related to the size of the library at home. More specifically, students were asked about the number of books in their homes, excluding magazines and school books. The third factor was the highest education level of their parents.

Step 2: A regression analysis was run in which the dependent variable was mathematics achievement, and the independent variables were the three above-mentioned factors (parents' educational background, the size of the home library, and economic status). The students placed above the regression line were those students with mathematics scores higher than would be expected. The students placed below the regression line were those who achieved lower scores than would be expected.

Step 3: In the third step, the residuals scores were calculated. By residuals, we mean the differences between the actual scores and the predicted scores. Students with positive residuals were those students whose achievement was higher than would be expected, and students with negative residuals were those whose achievement was lower than would be expected. The residuals scores of the students were then averaged for all schools. The schools with positive mean residuals were considered the more effective schools and the schools with the negative mean residuals were deemed the less effective schools.

Step 4: In this step, the schools were ranked from the most effective to the least effective school. The schools with average residuals >+0.10 and the schools with average residuals < -0 .10 were then selected. Our purpose was to select the schools at the extremes, as we considered these schools would give us a more reliable account of the factors determining school effectiveness.

Step 5: During this fifth step, we tried to choose the indicators that educational authorities have under their control and that influence student achievement. We used the following criteria to select the indicators for further analysis: (i) we accepted those where correlations of the residuals with all indicators were statistically significant; and (ii) we excluded those variables which were not related to mathematics.

Step 6: These criteria allowed us to identify 26 variables, which we then grouped in seven categories: passive learning, active learning, self-perception, attitudes, family incentives, class climate, and external incentives.

Step 7: For this step, we calculated *z*-scores and used the *t*-test for the final analysis. The reason for calculating the *z*-score was to place all indicators on the same scale to facilitate interpretation of the difference in mean scores between the more and the less effective schools. We then summed up the values of the indicators to produce a composite value, and standardized the values of each of the seven factors to a mean of zero and a standard deviation of one. Finally, we used the *t*-test to calculate the mean differences between the more effective and the less effective schools.

Results

This study aimed to determine the factors that distinguish schools as more effective or as less effective, specifically in mathematics. From Table 1, we can see that the composite measures of economic status of the family and educational background, as well as the variable "home library," correlated with achievement in mathematics. All three correlations were positive. The highest correlation was between educational background and achievement ($r_{\text{education\_math}} = 0.35$), followed by the size of the home library ($r_{\text{library\_math}} = 0.25$) and then economic status ($r_{\text{economic\_math}} = 0.21$). These correlations indicate that the higher the educational background of the family, the larger the size of the family's home library, and the higher the economic status of the family, the more likely it is that the son or daughter will have a higher level of achievement in mathematics.

Table 2 shows the regression equation of the three composites as independent factors, and mathematics achievement as the dependent variable. The regression analysis was based on the hypothesis that mathematics achievement is a function of parents' education, size of the home library, and the economic status of the family. We can see from the equation that the most significant factor in predicting mathematics achievement was the educational background of the parents. For all three independent factors, the contribution of variance to the prediction of mathematics achievement was statistically significant, although not high ($R = 0.377$, $R^2 = 0.142$).

Table 1: Correlations between Mathematics Achievement and Parents' Educational Background, Size of Home Library, and Family's Economic Status

r	Educational background	Size of library	Economic status
Mathematics achievement	0.35*	0.25*	0.21*

Note: * *p* < 0.000.

Table 2: Regression Equation for Predicting Mathematics Achievement

Predict. Math Achiev. = .28(education background) +.12(size of library) + .07(economic status)
R = 0.377 R^2 = 0.142

Figure 2 presents the position of the 61 schools based on their achievement and their average residuals. This graph allows us to compare schools that were more effective than we might have supposed from their students' mathematics achievement. For example, although school 25 had about the same average achievement ($\overline{X}$= 449) as school 57 ($\overline{X}$= 451), school 25 was more effective than 57. School 25 had a positive average residual (+11) and school 57 had a negative average residual (-21). Furthermore, although school 25 had a lower mathematics achievement ($\overline{X}$ = 449) compared with school 55 ($\overline{X}$= 496), it was a more effective school. The average residual for school 55 was negative (-6).

Figure 3 presents the schools after the exclusion of schools with small average positive or negative residuals. More specifically, the schools with average residuals > + 0.10 and the schools with average residuals < -0.10

Figure 2: Position of Schools Based on the Average Mathematics Achievement and on the Average Residuals of their Students

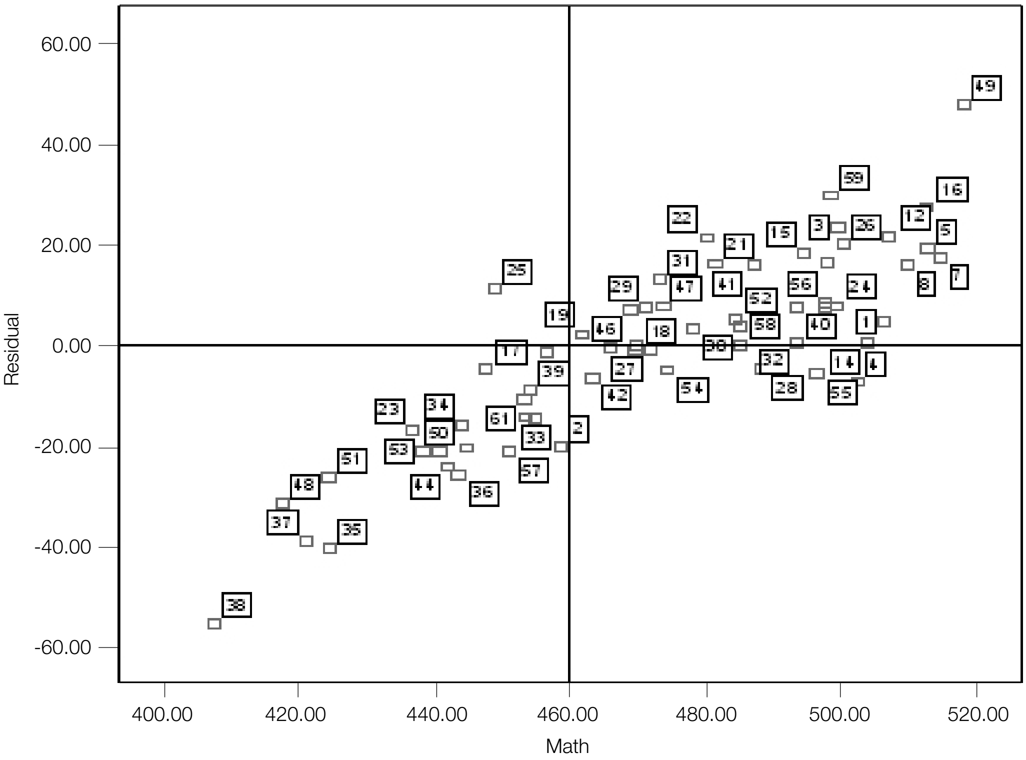
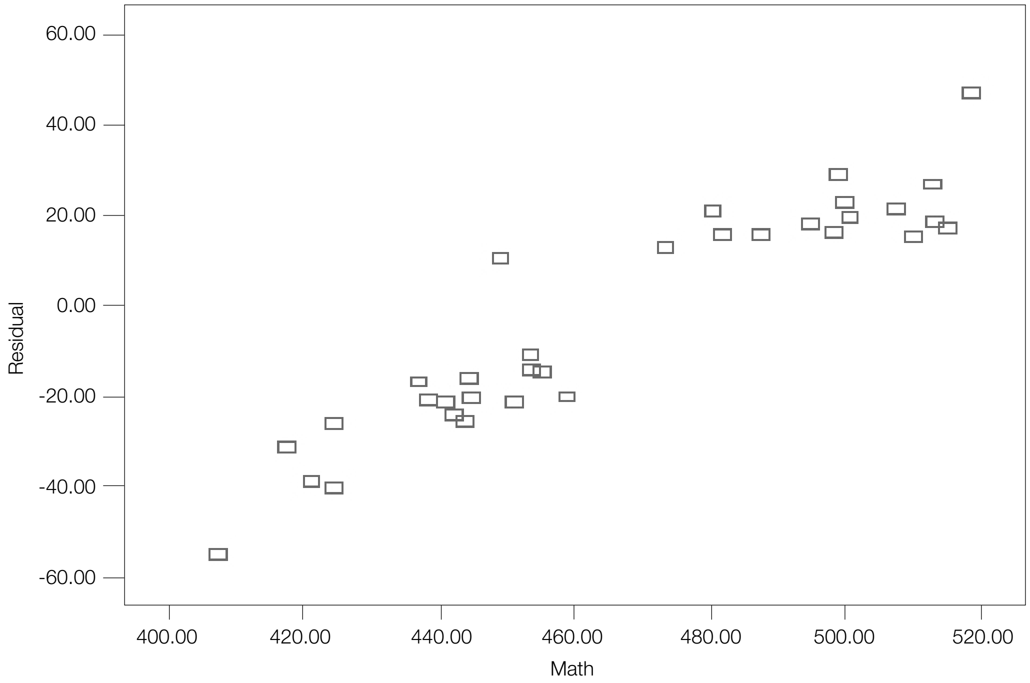


Figure 3: The Remaining Schools for Further Statistical Analysis



were excluded. In total, 33 schools out of 61 were retained, 16 of which had positive residuals and 17 of which had negative residuals.

Table 3 presents the seven categories of indicators that were selected for further analysis, and the corresponding indicators and *r* values between indicators and residuals. The significance of *r* was the criterion for selecting indicators. The seven composites were used to distinguish the more from the less effective schools. Table 4 presents the factors, the *t*-test values, the significance level, and the mean differences for the two groups of schools—the more effective and the less effective.

Our analysis revealed six factors explaining school differences in mathematics achievement. The most influential factor was *passive learning*. By passive learning, we mean that the teacher shows students how to solve mathematics problems, students copy notes from the board, the teacher uses the board during teaching, and the teacher uses an overhead projector during teaching. The second factor was *active learning*. Active learning is the opposite side of passive learning. In active learning, students work on mathematics projects, they use things from everyday life in solving mathematics problems, they work together in pairs or small groups, and they try to solve examples related to a new topic. When we look at these two factors, the most important among the seven, we see they are two sides of the same coin. For both factors—active and passive learning—the probability level was negligible, which led us to believe that these two factors are what really make the difference between the two groups.

The third factor that distinguished the two groups of schools was *self-perception*. Self-perception has been defined as individuals' beliefs regarding their performance capabilities in a particular context, or on a specific task or domain (Bandura, 1997). The beliefs that were included in this factor were the students' beliefs that mathematics is difficult, mathematics is not one of their strengths, mathematics is not an easy subject, and that they are not talented in mathematics. This factor also had a strong influence on differentiating the two groups.

The next relevant factor that distinguishes was the attitudes of students toward mathematics. Positive attitudes were signified by the students saying they enjoyed learning mathematics, they liked mathematics, and they would like a job involving mathematics. Students with positive attitudes were a characteristic of the group of effective schools.

The remaining two factors distinguishing the more effective and the less effective schools were

Table 3: Composites, Indicators, and r Values with Residuals

Factors	Indicators	r^*
1. Self-perception	1. I would like mathematics much more if it were not so difficult	.293
	2. Mathematics is more difficult for me than for many of my classmates	.473
	3. I am just not talented in mathematics	.399
	4. When I do not understand a new topic in mathematics initially, I know that I will never understand it	.254
	5. Mathematics is not one of my strengths	.369
	6. Mathematics is not an easy subject	.168
2. Attitudes	7. I enjoy learning mathematics	.181
	8. Mathematics is not boring	.241
	9. I would like a job that involved using mathematics	.180
	10. I like mathematics	.295
3. External incentives	*I need to do well in mathematics:*	
	11. To get the job I want	.052
	12. To please my parents	.233
	13. To get into the school of my choice	-.106
	14. To please myself	-.049
4. Passive learning	15. The teacher shows us how to do mathematics problems	.066
	16. We copy notes from the board	.109
	17. The teacher uses the board	.157
	18. The teacher uses an overhead projector	.225
5. Active learning	19. We work on mathematics projects	.312
	20. We use things from everyday life in solving math problems	.069
	21. We work together in pairs or small groups	.106
	22. We try to solve examples related to new topic	.253
6. Family incentives	23. Mother thinks it is important to do well in mathematics	.161
	24. I think it is important to do well in mathematics	-.174
7. Class climate	25. In my mathematics class, students are orderly and quiet	.107
	26. In my mathematics class, students do exactly as the teacher says	.105

Note: $^*p < 0.05$.

Table 4: Rank Order of Factors Distinguishing the More from the Less Effective Schools

Factors	*t*-test	*p*	Mean differences of *z*-scores
Passive learning	7.70	0.00	0.35
Active learning	7.41	0.00	0.33
Self-perception	4.95	0.00	0.23
Attitudes	2.61	0.01	0.12
Family incentives	2.00	0.045	0.09
Class climate	1.96	0.05	0.09
External incentives	1.31	0.19	0.06

family incentives and *school climate*, with the levels of significance at 0.045 and 0.05, respectively. These levels reveal the differences between the more and less effective school as marginal. The only factor that did not show a statistically significant difference between the more and the less effective schools was *external incentives*. Here, the *t*-test value was small (t = 1.31) and the probability level high (p = 0.19).

Conclusion

For Cyprus, participating in the IEA studies is of fundamental importance. Findings from these studies allow educational authorities to make cross-national comparisons of achievement, while the quality of the data enables in-depth analyses of the national results in an international context (Gonzales & Miles, 2001). This present article discussed important points of Cyprus's mathematics education. The conceptual framework of the study described in this article was based on instructional practices applied in mathematics teaching, seen from the students' perspectives, together with some background factors. The purpose of this study was to find the school indicators that differentiate more effective from less effective schools. For this reason, the analysis was based on the residuals, which present the differences between the actual mathematics scores and the predicted mathematics scores.

Six factors were found to influence the more from the less effective schools: passive learning, active learning, self-perception, attitudes, family incentives, and class climate. In our analysis, we tried to provide insight into the characteristics of classrooms that are associated with differences in school effectiveness. Such knowledge is often regarded as a potential foundation for school improvement interventions. If we know the features of effective schools, we can improve the lower performing schools by encouraging them to adopt the characteristics of effective schools.

The results of this research corroborate findings of other studies on school effectiveness. Differences were found between more effective and less effective schools, with the more effective schools exhibiting these characteristics:

- Teaching is clear, well organized and keeps students actively involved;
- Class climate is safe and orderly;
- Students are stimulated by (receive incentives from) their families;
- Students have positive attitudes toward mathematics; and
- Students hold positive beliefs regarding their performance capabilities in mathematics.

The contribution of this study is significant in that it was conducted in a country where all Grade 8 students follow the same mathematics curriculum. Our analysis revealed, however, two distinctly different learning environments. The findings that passive and active learning, self-perception, attitudes, and class climate have substantial effects on differentiating schools, at least in terms of mathematics achievement, carry major implications for mathematics education because all these variables are amenable to change through instruction. If less effective schools are to be more effective, they need to take account of all these educational interventions, and to take into consideration all of the factors underlying mathematics achievement.

Researchers suggest that students' self-perceptions/expectations are a major determinant of goal setting, and confirm that self-perceptions/beliefs can predict students' mathematics performance (Bandura, 1997; Pajares & Graham, 1999). The positive relationship between attitudes and mathematics achievement is well documented (MacLean, 1995). The general relationship between attitudes and achievement is based on the concept that the more positive an attitude a student has toward a subject, the more likely it is that he or she will reach a high level of performance. Ma (1997) observed significant positive relationships between students who stated that mathematics was important and that they enjoyed the subject and their achievement in mathematics. Researchers have also found that parental stimulation is another factor characterizing effective schools across many countries (Guzel & Berberoglu, 2005), a finding confirmed in our study.

The results of this analysis on school effectiveness contribute to a fuller understanding of the complicated issue of school improvement. However, the area of educational effectiveness still demands further theoretical and empirical research. Important issues that require further research are outcomes, inputs, and the learning process, and ideas on how we can promote an active learning environment in the classroom and in schools.

References

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: Freeman.

Baumert, J., Stanat, P., & Watermann, R. (2005, August). *Social and intellectual composition of schools' effects on student outcomes.* Paper presented at the EARLI conference in Nicosia, Cyprus.

Beaton, A. E., Mullis, I. V.S., Martin, M. O., Gonzales, E. O., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years.* Chestnut Hill, MA: IEA.

Brown S., Duffield J., & Riddell, S. (1995). School choice, social class and distinctions: The realization of social advantage in education. *Journal of Education Policy, 11*(1), 89–112.

Creemers, B. P. M. (1994). *The effective classroom.* London: Cassell.

Creemers, B. P. M., & Reezigt, G. J. (2005). Linking school effectiveness and school improvement: The background and outline of the project. *School Effectiveness and School Improvement, 16*(4), 359–371.

Gadeyne, E., Ghesquiere, P., & Onghena, P. (2006). Psychological educational effectiveness criteria and their relation to teaching in primary education. *School Effectiveness and School Improvement, 17*(1), 63–85.

Gonzales, E. J., & Miles, J. A. (Eds.). (2001). *TIMSS 1999 user guide for the international database: IEA's repeat of the Third International Mathematics and Science Study at the eighth grade.* Chestnut Hill, MA: Boston College.

Grobler, A., Grobler, A. A., & Esterhuyse, F. G. (2001). Some predictors of mathematics achievement among black secondary school learners. *South African Journal of Psychology, 31*(4), 48–54.

Guzel, C., & Berberoglu, G. (2005). An analysis of the Programme for International Student Assessment 2000 (PISA 2000) mathematical literacy data for Brazilian, Japanese, and Norwegian students. *Studies in Educational Evaluation, 31*, 283–314.

Luyten, H., Visscher, A., & Witziers, B. (2005). School effectiveness research: From a review of the criticism to recommendations for further development. *School Effectiveness and School Improvement, 16*(3), 249–379.

Ma, X. (1997). Reciprocal relationships between attitude toward mathematics and achievement in mathematics. *Journal of Educational Research, 90*(4), 221–229.

MacBeath, J., & Mortimore, P. (2001). *Improving school effectiveness.* Buckingham, UK: Open University Press.

MacLean, B. D. (1995). *Educational traditions compared: Content, teaching and learning in industrialized countries.* London: David Fulton.

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: Boston College.

Nasser, F., & Birenbaum, M. (2005). Modeling mathematics achievement of Jewish and Arab eight grades in Israel: The effects of learner-related variables. *Educational Research and Evaluation, 11*(3), 277–302.

Opdenakker, M. C., & Van Damme, J. (2005, August). *Are school structures and/or teaching processes responsible for the group composition effect?* Paper presented at the EARLI conference in Nicosia, Cyprus.

Pajares, F., & Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. *Contemporary Educational Psychology, 24*(2), 124–139.

Papanastasiou, C. (2000). Internal and external factors affecting achievement in mathematics: Some findings from TIMSS. *Studies in Educational Evaluation, 26*(1) 1–7.

Papanastasiou, C. (2002). Effects of background and school factors on mathematics achievement. *Educational Research and Evaluation, 8*(1), 55–70.

Postlethwaite, T. N. & Ross, K. N. (1992). *Effective schools in reading: Education for educational planners.* Amsterdam: IEA.

Reynolds, D., & Stoll, L. (1996). Merging school effectiveness and school improvement: The knowledge base. In D. Reynolds, R. Bollen, B. Creemers, D. Hopkins, L. Stoll, & N. Lagerweij (Eds.), *Making good schools: Linking school effectiveness and school improvement* (pp. 94–112). London: Routledge.

Scheerens, J., & Demeuse, M. (2005). The theoretical basis of the effective improvement model (ESI). *School Effectiveness and School Improvement, 16*(4), 373–385.

Schreiber, B. J. (2002). Institutional and student factors and their influence on advanced mathematics achievement. *Journal of Educational Research, 95*(5), 274–286.

Secada, G. W. (1992). The reform of school mathematics in the United States. *International Journal of Educational Research, 17*(5), 399–516.

Stoll, L., & Wikeley, F. (1998). Issues on linking school effectiveness and school improvement. In W. Th. J. G. Hoeben (Ed.), *Effective school improvement: State of the art contribution to a discussion* (pp. 29–58). Groningen: GION Institute for Educational Research, University of Groningen.

Van de Grift, W. J. C. M., & Houtveen, A. A. M. (2006). Underperformance in primary schools. *School Effectiveness and School Improvement, 17*(3), 255–273.

Wilms, J. D. (1992). *Monitoring school performance: A guide for educators.* London: Falmer Press.

Change in mathematical achievement in the light of educational reform in Lithuania

Jolita Dudaitė
Kaunas University of Technology
Kaunas, Lithuania

Abstract

This paper considers change in the mathematics achievement of basic school students from 1995 to 2003 in Lithuania. The analysis draws on data from TIMSS 1995, 1999, and 2003. The TIMSS cycles and the scaling methodology used for calculating the scores provide opportunity for participating countries not only to compare their results with results from other countries, but also to track the changes in their students' achievement across the years. This facility is of particular importance to countries experiencing considerable changes in their education systems. Lithuania is one such country, as it has undergone considerable educational reform since the early 1990s. Participation in the aforementioned three TIMSS cycles provided Lithuania with a reliable means of measuring the impact of reform as it related to the mathematics achievement of students in Grade 8 of the basic school. The analysis described in this paper involved content analysis and classical statistical investigation. (The main statistical software used was SPSS 12.0.)

Introduction

Over the last century, educational reforms in various countries have become a part of the daily routine of educational institutions (Horne, 2001; Kinsler & Gamble, 2001). Many researchers examine the results of these reforms in various countries. Some praise the reforms (Draper, 2002; Gamoran, 1997); others say that the reforms have not had the desired results (Horne, 2001). When considering reforms in mathematics curricula, some researchers point to positive impacts. These include:

- Students finding mathematics more interesting to learn when the subject is connected to real life-and-work contexts (Nicol, Tsai, & Gaskell, 2004);
- Girls' attitudes to and performance in mathematics improving when extra attention is given to teaching girls this subject (Richardson, Hammrich, & Livingston, 2003);
- Students gaining a better understanding of algebra following changes to the content of mathematics lessons (Krebs, 2003);
- Students upping their achievement scores in mathematics following a change from traditional lecture-type teaching methods to active and problem-solving methods (Sawada et al., 2002).

However, other researchers claim that, despite considerable efforts to reform the content of mathematics education, teaching methods, and instructional aids, these efforts often do not have the desired results (Vann, 1993). The desire to reform mathematical programs has come not only from the expectations that schools now have for higher student achievement in mathematics as a result of their respective country's overall program of educational reform (Kelly & Lesh, 2000), but also, and usually because of, students' generally low level of mathematics achievement (Betts & Costrell, 2001; Frykholm, 2004; Hess, 2002). Thus, one of the main goals of the reform has been to improve students' achievement in this subject, and the success of the reform has been, not surprisingly, frequently measured by the changes to students' achievement scores in mathematics (Finnan, Schnepel, & Anderson, 2003).

It is not easy, though, to measure improvement (Sawada et al., 2002), especially over a short time period, as is usually wanted (Grissmer & Flanagan, 2001). The reasons why vary. For example, the high standards expected of reformed mathematics programs force teachers to teach students only in the topics that will be tested and force students to cheat during tests (Hess, 2002). Although increases in student achievement may be observed soon after a reform has been put in place, this improvement tends to be short-lived, with achievement frequently reverting to previous levels—and often in line with the normal distribution curve—after a few years (Vann, 1993).

Many researchers claim that reforms do not produce any positive changes in the students' mathematics achievements, and can even worsen their achievement (Alsup & Sprigler, 2003). They explain this in terms of reforms introduced too quickly and/or the reform tried in only a few schools, over a very short period. In these cases, the intended guidelines of the reform are not sufficiently well grounded. Sometimes the results expected from the reform are immeasurably high (Gordon, 2004) or the goals of the reform so unrealistic, such as "graduates who know mathematics better than graduates of all other countries in the world" (Hill & Harvey, n. d.) that measurement of changes in achievement becomes pointless. Also, it can be fruitless to try to improve students' mathematics results because of incompatibility between the intent or theory of the reform and actual practice in the classroom. Moreover, teachers may be reluctant or not have the competence to embrace the ideas of the reform and to implement them (Finnan et al., 2003; Kyriakides, 1997).

The majority of researchers who say the reforms rarely lead to the expected improvements in achievement consider this is because the reforms focus only on the changes within the classroom and the school education environment. However, schools are not islands (Fullan, 2003), and the learning achievements of students strongly relate to their home socio-educational environments, as well as to other non-school environments (Cohen & Hill, 2000; Green, 1987; Rotberg, 2000; Viadero, 1996). These researchers also suggest that student achievement depends more on these non-school than in-school environments (Barton, 2001; Coleman, cited in Edmonds, 1987). As Barton (2001) points out, although the school itself naturally influences student achievement (see also Fullan, 1998), it is quite unrealistic to expect the school to have the sole influence.

To recap, of the many researchers who have analyzed the influence of educational reform on students' mathematics achievement, some see positive outcomes, but others think that, for various reasons, the reforms produce limited or no results in the long run. They stress that educational reforms tend to be associated with the schooling environment and fail to recognize that their effectiveness (in terms of student achievement) also depend on the students' home socio-educational environments. Because every country has its own specific schooling and social environments, it is worth analyzing specific countries' specific educational reforms and the results of those reforms. Lithuania, a country that has been implementing educational reform for some time, provides a case in point.

After Lithuania claimed independence in 1991, radical changes in society made it necessary to make changes to the education system (Želvys, 1999). These included rewritten study programs and educational standards, new textbooks, and modified teaching priorities and goals. A more modern stance began to inform teaching and learning. By drawing on the experience of other countries, Lithuania is endeavoring to form a national integrated system of schooling. It is also trying to move away, within the basic school, from an academic teaching approach to basic literacy, from an emphasis on reproduction of knowledge to development of skills, and from "dry" theory to more real-life situations. Teaching methods are changing: in addition to using the traditional lecture style of teaching, teachers are being encouraged to use various active teaching methods. In short, the prevailing model of a reproductive education system is being rejected, and an interpretative education system created.

Research questions and methods

The main research focus of this paper is to explore the extent to which changes in Lithuania's educational school environment (in association with changes within political, social, and home spheres) appear to be reflected in students' mathematics achievement. More specifically: *Can we detect changes in mathematical literacy level while an educational reform is taking place? Are students achieving, on average, at a higher level in mathematics than they did at the beginning of the educational reform?*

We can answer such questions not only by analyzing the present situation, but also by comparing it with the situation at the beginning of Lithuania's political independence. The way to do this is to conduct a longitudinal study, during which researchers collect data on students' mathematics achievement and data relating to factors that have a bearing on those achievements. The only research of this type conducted in Lithuania is TIMSS (Trends in International Mathematics and Science Study), organized by the International Association for the Evaluation of Educational Achievement (IEA). Lithuania has participated in the three cycles of this research, conducted in 1995, 1999, and 2003.

Lithuania's continuing participation in TIMSS has thus provided the country with opportunity to evaluate the effectiveness of Lithuania's educational development, to document the changes, and to identify possible general problems in education. In Lithuania, TIMSS has been the only study of educational achievement conducted consistently throughout the time of the educational reform.

The information registered as a result of Lithuania's participation in TIMSS related (for the 1995 cycle) to students who learned from mathematics textbooks, translated from the Russian and Estonian languages, and (for the 1999 and 2003 research cycles) to students who had studied from textbooks, written by Lithuanian authors. In Lithuania, only students in Grade 8 were tested during the three cycles.

TIMSS uses the IRT (Item Response Theory) scaling methodology (in which the mean on the scale of student achievement scores is set at 500 and the standard deviation at 100). This allows each participating country not only to compare the average level of achievement for its students with the average levels of achievement in the other participating countries but also to compare the results for its own students across all three cycles of the study.

The TIMSS results for Lithuanian students and the changes in these results over the three cycles have received minimal analysis. Zabulionis (1997a, 1997b, 2001) and Trakas (1997) undertook some analysis of the 1995 results, and the several publications produced on TIMSS within Lithuania offer only a limited presentation, without analysis, of the results (Čekanavičius, Trakas, & Zabulionis, 1997; Dudaitė, Elijio, Urbienė, & Zabulionis, 2004; Mackevičiūtė & Zabulionis, 2001). Dudaitė (2006) edited a text presenting analyses of Lithuanian students' mathematics results for the period 1995 to 2003.

This present paper presents a further analysis of the changes in Lithuanian students' mathematics achievement in the TIMSS assessment across the three cycles. In 1995, 2,547 Grade 8 students from Lithuania participated in the study; 2,361 students participated in 1999; and 5,737 students in 2003. The main goal of the analysis was to identify changes in Lithuanian Grade 8 students' mathematics achievement results in TIMSS from 1995 to 2003 and to offer possible explanations for changes. The data for this work were drawn from the TIMSS 1995, 1999, and 2003 databases, and the analyses involved content analysis and classical statistical investigations. The main statistical software used was SPSS 12.0.

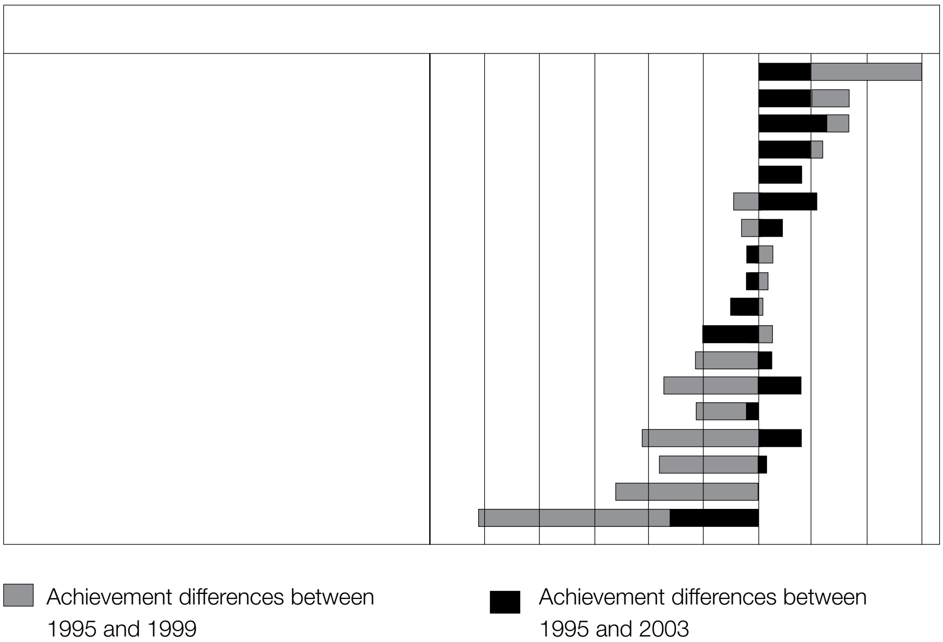
Review of Lithuanian students' achievement on the TIMSS mathematics tests

Analysis of the TIMSS results showed a general improvement across the three cycles in the mathematics achievements of Lithuanian Grade 8 students. The students' average score on the 1999 scale was 10 points higher (*SE* = 6.1) than on the 1995 scale, but this difference was not statistically significant. However, the difference between the students' average scores on the mathematics scale for TIMSS 1999 and TIMSS 2003 was much higher (20 points, *SE* = 5.0) and statistically significant (Mullis, Martin, Gonzalez, & Chrostowski, 2004). We can gain a clearer perspective on the import of this increase by comparing the Lithuanian results with the results of the other countries that participated in all three TIMSS cycles.

From Figure 1, we can see that the increase across the three cycles in Lithuanian students' achievement was higher than the increase for any other country. (The shaded bars on the right-hand side of the figure signify increases in average mathematics achievement and those on the left signify a decrease.) Latvia, which neighbors Lithuania, had a 17-point increase in achievement between 1995 and 1999, but the country had progressed no further by 2003. Russia, another country neighboring Lithuania, saw a decrease in achievement from 1995 to 2003. The highest decrease in mathematics achievement from the first TIMSS assessment in 1995 to the third assessment in 2003 was in Bulgaria (51 scale points).

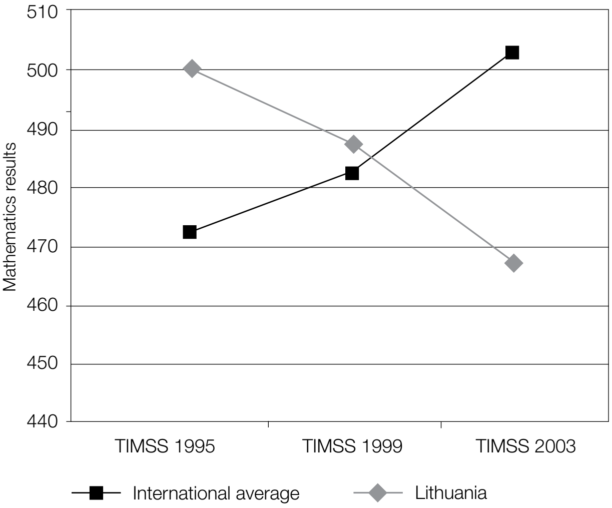
Comparison of Lithuanian students' average mathematics results with the international averages for each cycle (Figure 2) is also informative. Figure 2 shows us that across the three cycles of the TIMSS study, the international average decreased from 500 to 467 scale points, but that the Lithuanian average increased from 472 to 502 scale points.

Figure 1: Comparison of Average Mathematics Achievement of Students in the Countries that Participated in all Three TIMSS Cycles



In 1995, Lithuanian average achievement was significantly below the international average; Lithuania, in fact, appeared at the bottom of the country list. However, in 1999, Lithuanian average achievement was similar to the international average. In 2003, Lithuanian students proved themselves very successfully and outstripped the international average; the difference was marked. It needs to be acknowledged that the international average had strongly decreased by 2003, but this must also be seen in relation to the fact that the countries participating in each TIMSS cycle were not the same. Also, the comparison between the international average for TIMSS 1995 and the Lithuanian results for TIMSS 2003 showed that, by 2003, Lithuanian Grade 8 students had reached the international average of 1995. As such, it is fair to say that Lithuania outstripped the international benchmark not by 35 points but by only about two points. Another consideration is that, in 1995, the countries participating in TIMSS were nearly all from West European and Asian countries, which tended to have the higher achievement results. It was therefore particularly useful for Lithuania to have these countries as a comparison point because of the generally higher results across the participating countries. By 2003, the list of participating countries had greatly expanded and included many developing countries.

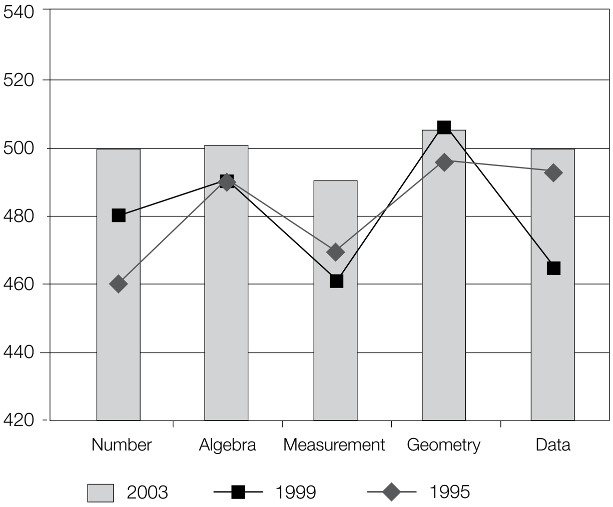
Figure 2: Comparison of the Shift in International Average Achievement and Lithuanian Average Achievement of Grade 8 Students on the Three TIMSS Cycles



Having considered the Lithuanian Grade 8 students' general achievement, let us now take a closer look at particular results. I would like to suggest that each single TIMSS item can be examined as if it were an international mathematics mini-contest. Therefore, it is interesting to observe how many times Lithuanian students won or lost these competitions, and how these results changed over the years. The leaders of these contests undoubtedly were Asian countries: Singapore, Japan, Hong Kong SAR, South Korea, and Taiwan. A comparison of Lithuania with those countries that participated in the TIMSS study all three times shows that Lithuania in TIMSS 1995 was in the top five countries on only 1.9% of items (from 155 items), and in the bottom five countries on 44.6% of items (from 155 items). TIMSS 2003 saw an almost two-fold improvement on the first result (4.1%, from 194 items), and a four-fold decrease on the second (10.3%, from 194 items) (see Figure 3). With this latter result, Lithuania took the lead amongst the countries participating in the study all three times.

Consideration of the Lithuanian students' results for the various mathematics content areas showed that in 1995 the knowledge and the abilities of these students in five content areas were very different (see Figure 4). In 1995, the best-solved items were those relating to geometry (508 scale points), followed by those relating to algebra (488 scale points). The results of the other three mathematics content areas were much worse (measurement, 457 scale points; number, 462; data, 465). By 1999, the differences between the students' results in mathematics content areas had decreased slightly, with the best improvement evident for data (28 scale points). By 2003, the students' levels of achievement in the different mathematics content areas had become very similar. Over the eight-year period, the results showing the least change were those for geometry and algebra, and those showing the most (and significant) improvement were for number and data.

Figure 4: Changes in Lithuanian Students' Results across the Three TIMSS Cycles by Mathematics Content Areas



Note: The advanced benchmark was set at 625 or more points on the scale; the high at 550–624 points; the intermediate at 475–549 points; and the low at 400–474 points.

Analysis of the Lithuanian students' results against the international benchmarks also showed constant improvement across the three TIMSS cycles (see Figure 5). Fewer students were at the low benchmark in 2003 (10%) than in 1995 (19%).

In summary, from 1995 to 2003, the average mathematics achievement of Lithuanian Grade 8 students improved. Let us now consider possible explanations for that improvement.

Figure 3: The Number of Times that Lithuanian Grade 8 Students' Performance on TIMSS Mathematics Items Placed Them in the Five Top- and Five Bottom-performing Countries for Those Items

Lithuania	Share of items in TIMSS math test, on which Lithuania is in the group of five bottom countries	Share of items in TIMSS math test, on which Lithuania is in the group of five top countries
TIMSS 2003	10.3%	4.1%
TIMSS 1999	36.1%	2.4%
TIMSS 1995	44.6%	1.9%

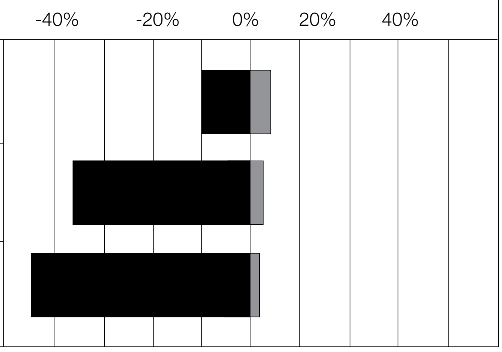
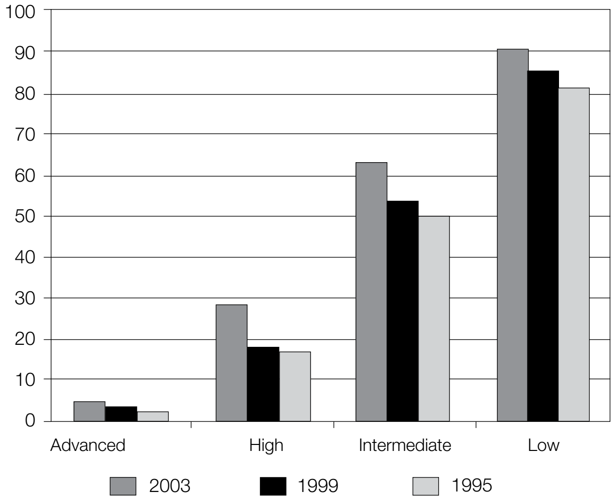


Figure 5: Trends in Percentages of Lithuanian Students at the International Benchmarks for the Three TIMSS Cycles



Explanations for improvement

It is clear that Lithuania's educational reform influenced the improvement in Lithuanian students' mathematics achievement from the time of the first to the third cycle of TIMSS. In particular, it seems fair to say that this improvement was an outcome of the newly established educational standards, the rewritten study programs, and the mathematics textbooks written in the "spirit" of TIMSS. After Lithuania participated in the TIMSS assessment for the first time and achieved very low results, educational reform (including school mathematics) was deflected more toward the style of the TIMSS items. This development signified recognition that one of the main objectives of the educational reform should be transformation from the conveyance of knowledge to the education of competence, from academic-style mathematics to mathematics literacy. Because TIMSS assesses students' mathematics literacy, Lithuania's participation in the study provided a good and appropriate impetus for this change.

We can therefore partially explain Lithuania's low level of achievement in the first TIMSS assessment by recognizing that in 1995 Lithuanian schools did not emphasize or teach mathematics literacy. Lithuanian students were used to a different type of mathematics, and therefore were not able to demonstrate their knowledge in TIMSS 1995. TIMSS 2003, executed after implementation of the educational reform, assessed students educated in contemporary Lithuanian schools. This argument alone is a solid one in explaining the marked improvement in the Lithuanian results in 2003.

Lithuanian mathematics study programs and the TIMSS frameworks

I would now like to look at the differences and the similarities between the TIMSS research frameworks and Lithuanian mathematics study programs as well as the changes within them. In 1995 and in 1999, the structure of the TIMSS research had the following three dimensions (Robitaille, McKnight, Schmidt, Britton, Raizen, & Nicol, 1993):

- *Content:*
  - Numbers
  - Measurement
  - Geometry
  - Proportionality
  - Functions, relations, equations
  - Data, probability, statistics
  - Elementary analysis
  - Validation and structure.
- *Performance Expectations:*
  - Knowing
  - Using routine procedures
  - Investigating and problem-solving
  - Mathematical reasoning
  - Communicating.
- *Perspectives:*
  - Attitudes
  - Careers
  - Participation
  - Increasing interest
  - Habits of mind.

In 2003, the structure of the TIMSS research was somewhat changed; two structural dimensions were left, but they had been slightly amended (Mullis et al., 2004):

- *Content Domains:*
  - Number
  - Algebra
  - Measurement
  - Geometry
  - Data.
- *Cognitive Domains:*
  - Knowing facts and procedures
  - Using concepts
  - Solving routine problems
  - Reasoning.

Analysis of the Lithuanian study programs shows considerable differences between those programs written before the reform and those written during it. The reformed programs contain new themes such as statistical elements, elements of probability theory, combinatorics, elements of economics, elements of computer science, and problem- solving (mathematical reasoning). The detailed themes of algebra, geometry, and number remain almost the same as they were before the reform (Dudaitė, 2000; Lietuvos Respubliko švietimo ir mokslo ministerija, 1997a, 1997b; Lietuvos TSR švietimo ministerija, 1988).

Comparison of the mathematics content of the Lithuanian study programs with the TIMSS mathematics content shows the pre-reform Lithuanian mathematics study content differs most substantially from the TIMSS 1995 frameworks because the former did not contain data representation, probability, and statistics topics, as well as elementary analysis, validation, and structure. However, in relation to other mathematics content themes, analysis reveals no differences between the content of the TIMSS 1995, 1999, 2003 frameworks and the pre-reform and the post-reform Lithuanian study programs for mathematics. Thus, we could assume the low results for Lithuanian students in TIMSS 1995 were because some of the TIMSS questions tested knowledge or skills that the Lithuanian students had not learned. However, when we take into account the number of TIMSS 1995 items that matched the content of Lithuanian pre-reform mathematics study programs, we get a high result—95.7% (Beaton et al., 1996).

If only 4.3% of the TIMSS 1995 items did not match the content of the Lithuanian mathematics study programs, this difference alone could not have provided the reason for the Lithuanian students' low results. It is also important to note that the Lithuanian students' results for the data representation, probability, and statistics domains (data) in TIMSS 1995 were not their lowest domain scores (465 scale points; in comparison: number, 462; measurement, 457). With all this in mind, we must conclude that the improvement in the Lithuanian students' mathematics results across the three TIMSS cycles can be only partially explained by change to the content of the Lithuanian mathematics study programs.

Lithuanian mathematics teaching goals and the TIMSS frameworks

Another point of comparison is mathematics teaching goals before and after the reform, and it is here that the influence of changes on students' achievement in mathematics is particularly evident. In 1988, before the educational reform, the main mathematics teaching goals were formulated as follows (Lietuvos TSR švietimo ministerija, 1988):

- To give *knowledge*
- To form *skills*
- To train *logical thinking*
- To teach students how to use the knowledge in *mathematics-related subjects*
- To prepare students in such a way that they could *continue their studies*.

The teaching goals formulated during the educational reform in 1997 had a different tone (Lietuvos Respublikos švietimo ir mokslo ministerija, 1997b):

- To develop mathematical *communication*
- To teach to solve *standard* mathematical *procedures*
- To teach to solve mathematical problems and to investigate
- To seek for *mathematical reasoning*
- To train positive *attitudes* toward mathematics
- To encourage mathematical, scientific, and technological *careers*
- To promote the *studying* of mathematics
- To form a mathematical, scientific thinking *habit*.

In addition, the reformed study programs of 1997 stated the main purpose of mathematics teaching to be that of guaranteeing *mathematical literacy* for all members of society. One main point, and a very important one, regarding the changes in wording in the programs relates to the appearance of the notion of *mathematical literacy*. Pre-reform, schools taught a more academic style of mathematics. Mathematical literacy was not something to aim for.

Comparison of the goals of mathematics teaching formulated before and during the reform with the content of the TIMSS frameworks (Robitaille et al., 1993) shows equivalency between the 1997-formulated goals and the following structural dimensions of the TIMSS 1995 and 1999 frameworks: "performance expectations" (all parts except the first one, "knowing") and "perspectives" (all parts). The goals of mathematics teaching formulated before the reform were equivalent only for the first two parts of the TIMSS 1995 framework's dimension "performance expectations" (i.e., "knowing" and "using routine procedures"). Thus, the Lithuanian mathematics teaching goals articulated during the time of the reform (in 1997) are in essence equivalent to the TIMSS research format, but the same cannot be said of the goals set down before the reform. This means the Lithuanian students participating in TIMSS 1999 and 2003 had received mathematics education that accorded with the TIMSS "spirit," a factor that explains, to a good degree, the significant improvement in the Lithuanian students' mathematics results by 2003.

Lithuanian mathematics textbooks and the TIMSS frameworks

While the Lithuanian study programs and educational standards set down the goals of mathematics teaching, mathematics content areas, and detailed topics, they do not indicate how much time schools should spend on each topic. However, it is possible to establish approximate times through analysis of the mathematics textbooks.

The students who participated in TIMSS 1995 studied, during their Grades 5 and 6 years, the following textbooks, which were translated into Lithuanian from the Estonian language:

- Nurkas, E., & Telgma, A. (1990). *Matematika: Vadovėlis V klasei [Mathematics: Textbook for Grade 5]*. Kaunas: Šviesa
- Nurkas, E., & Telgma, A. (1991). *Matematika: Vadovėlis VI klasei [Mathematics: Textbook for Grade 6]*. Kaunas: Šviesa.

In Grades 7 and 8, these students studied from textbooks from the Russian language:

- Teliakovskis, S. (1991). *Algebra: Vadovėlis VII klasei [Algebra: Textbook for Grade 7]*. Kaunas: Šviesa
- Teliakovskis, S. (1990). *Algebra: Vadovėlis VIII–IX klasei [Algebra: Textbook for Grades 8–9]*. Kaunas: Šviesa
- Atanasianas L., et al. (1991). *Geometrija: Vadovėlis VII–IX klasei [Geometry: Textbook for Grades 7–9]*. Kaunas: Šviesa.

According to the reformed mathematics study programs and educational standards, the students who participated in TIMSS 1999 and 2003 studied from textbooks written by Lithuanian authors:

- Stričkienė, M., & Cibulskaitė, N. (1996). *Matematika 5*. Vilnius: TEV
- Stričkienė, M., & Cibulskaitė, N. (1996). *Matematika 6*. Vilnius: TEV
- Cibulskaitė, N. et al. (1998). *Matematika 7*. Vilnius: TEV
- Cibulskaitė, N. et al. (1998). *Matematika 8*. Vilnius: TEV.

Analysis of the Lithuanian mathematics textbooks reveals that the topics of algebra and geometry receive less attention than they did in the earlier texts but that number and measurement receive more (Zybartas, 1999). The new texts also include new topics: statistics, probability theory, combinatorics, and mathematical reasoning. These features explain why the results for Lithuanian students in algebra and geometry changed little over the three TIMSS cycles, and why the areas of greatest improvement were number and data (statistics and probability).

Students' socio-educational home factors

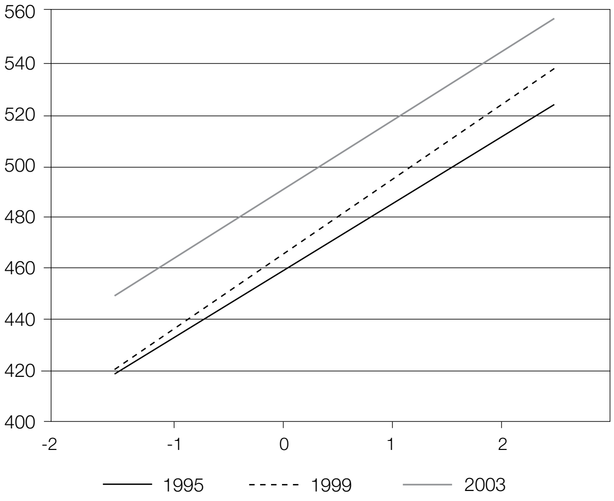
We cannot consider changes in Lithuanian students' mathematics results without considering societal factors, such as changes in the students' economic and educational home environments. The results of many studies show a strong relationship between students' home socio-educational environment and their mathematics achievement.

Let us therefore form a home socio-educational environment factor. Because the indicators that need to be taken into account must be in all three TIMSS cycles, the possible indicators are these: mother's and father's highest educational qualification, number of books at home, owning an encyclopedia, a dictionary, and a calculator, and having a work-table at home. Now let us take these possible indicators and use them to form a home socio-educational environment factor SES (Cronbach alpha: TIMSS 2003, 0.631; TIMSS 1999, 0.557; TIMSS 1995, 0.383). Regression analysis showed a strong relationship between Lithuanian students' mathematics results and their SES (see Figure 6). From this figure, we can see that students from the same home socio-educational environments gained more mathematics points with each TIMSS cycle.

It is interesting to observe the extent to which the students' actual home socio-educational environment changed over the eight-year period. We can do this by forming an index with the previously used indicators:

highest parental education, number of books at home, owning a calculator, an encyclopedia, and a dictionary, and having a work-table. Here, highest education of parents is categorized as follows: lower than or equivalent to ISCED 3; equivalent to ISCED 4; and equivalent to or higher than ISCED 5. Figure 7 shows students' home socio-educational environment worsened over the eight years. Consequently, despite the strong relationship between the Lithuanian students' mathematics results and their home backgrounds, SES does not explain the improvement in the students' performance across the three TIMSS cycles.

Figure 6: Relationship between Lithuanian Students' Home Socio-educational Environment and Their Mathematics Results across the Three TIMSS Cycles



TIMSS	B	B_1	Sig.
1995	470.677	31.525	0.000
1999	481.157	35.597	0.000
2003	504.675	31.253	0.000

Students' attitudes toward mathematics

Another possible explanation for the large improvement in Lithuanian students' mathematics achievement relates to students' attitudes toward mathematics as a subject. The correlation between student achievement and attitudes toward mathematics (measured by the statement, "I like mathematics X much") in TIMSS 1995 was 0.230, in TIMSS 1999, 0.288, and in TIMSS 2003, 0.239. Figure 8 shows, for Lithuanian students, a clear improvement in attitude toward mathematics between 1995 and 1999, but a worsening of attitude between 1999 and 2003, despite the improvement in achievement over this latter period. The improvement in attitude in the former period was probably due to the introduction of the new mathematics textbooks in 1996, which differed substantially from the previous textbooks in terms of design and variety and number

Figure 7: Changes in Lithuanian Students' Home Socio-educational Environment

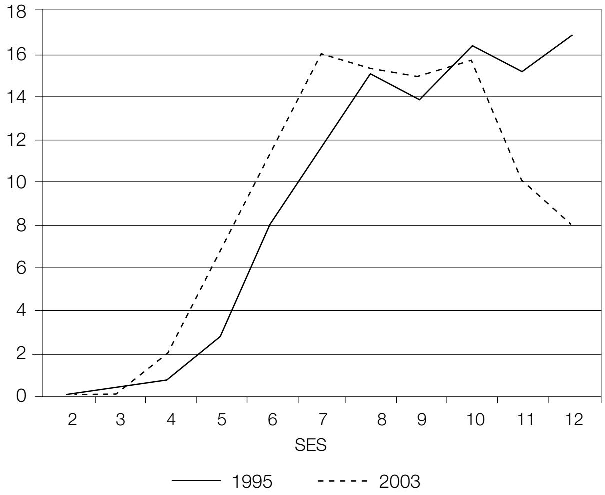
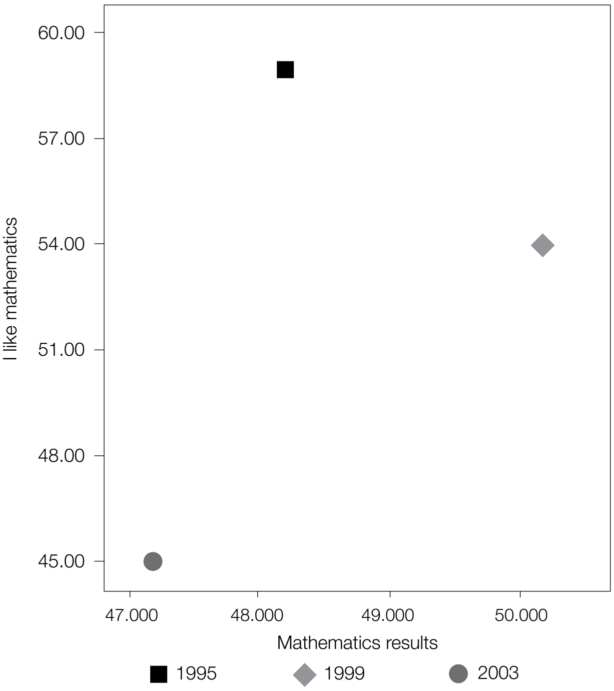


Figure 8: Relationship between Lithuanian Students' Attitudes toward Mathematics and Their Mathematics Results across the Three TIMSS Cycles



of exercises. However, it seems that once the novelty of the new books wore off, mathematics again became a less interesting subject for many students. Nonetheless, it is likely that the higher test results of 2003 reflected the stronger interest in mathematics in 1999.

Format of test questions

Another possible reason for the Lithuanian students' improved achievement was greater familiarity with the multiple-choice format of the test items. Most of the Lithuanian students who participated in TIMSS 1995 had not encountered this format because it was rarely, if ever, used in the pre-reform mathematics textbooks. Analysis of omitted (not solved) TIMSS items illustrates this. Table 1 shows the average percentages of students who omitted items without solving them in the TIMSS 1995 and 2003 assessments. Students omitted the multiple-choice answer format items twice as often in 1995 as in 2003, but there was no significant difference between the two dates in the percentages of students omitting open-response items.

We get a similar result with the trend items of TIMSS 1995 and 2003 (see Table 2). Once again, it can be seen that the students omitted the multiple-choice items twice as often in 1995 as they did in 2003. A comparison showed little difference in the difficulty of the items.

To verify that the Lithuanian students presumably were more proficient in 2003 at answering questions in the multiple-choice format, we can look at the extent to which students in 1995 and students in 2003 omitted the most difficult and the easiest items. As Table 3 shows, Lithuanian students in 2003 were twice as likely in 1995 to omit these questions as they were in 2003. Thus, by 2003, many Lithuanian students were able to apply the strategies needed to answer test items in multiple-choice format.

Conclusions

The key findings of this paper are as follows:

1. From 1995 to 2003, the average mathematics achievement of Lithuanian Grade 8 students improved significantly.
2. Of the countries that participated in all three TIMSS cycles, Lithuania improved the most in terms of its Grade 8 students' mathematics achievement.
3. In 2003, Lithuanian Grade 8 students' achievement across the different mathematics content domains was more homogenous than it was in 1995.
4. From 1995 to 2003, the mathematics content domains in which Lithuanian Grade 8 students improved most were data and number. The least amount of change was evident for algebra and geometry.
5. In relation to solving particular test items, the performance of Lithuanian Grade 8 students placed Lithuania amongst the five top countries twice as often in 2003 as in 1995. The students'

Table 1: Percentage of Multiple-choice (MC) and Open-response (OR) Test Items Omitted by Lithuanian Students in TIMSS 1995 and 2003

TIMSS	MC omitted (%)	OR omitted (%)
1995	7.50 (*SE* = 0.55)	20.52 (*SE* = 1.84)
2003	3.60 (*SE* = 0.24)	23.38 (*SE* = 1.60)

Table 2: Percentage of Multiple-choice Trend Items Omitted by Lithuanian Students in TIMSS 1995 and 2003

TIMSS	MC omitted (%)	Item difficulty (average)
1995	6.83 (*SE* = 0.84)	57.12 (*SE* = 3.48)
2003	3.18 (*SE* = 0.37)	61.45 (*SE* = 3.09)

Table 3: Differences between Lithuanian Students' Omission of 10 Most Difficult and 10 Easiest TIMSS 1995 and 2003 Items (Multiple-choice Only)

TIMSS	10 most difficult items		10 easiest items	
	Omitted (%)	Item difficulty	Omitted (%)	Item difficulty
1995	12.31 (*SE* = 3.59)	24.03 (*SE* = 1.96)	1.48 (*SE* = 0.49)	86.55 (*SE* = 1.16)
2003	5.77 (*SE* = 0.73)	24.36 (*SE* = 2.18)	0.81 (*SE* = 0.10)	85.65 (*SE* = 1.38)

performance placed Lithuania in the list of the five bottom countries four times less often in 2003 than in 1995.

6. The number of Lithuanian Grade 8 students below the low international benchmark for mathematics achievement decreased between 1995 and 2003.
7. The improvement in Lithuanian students' mathematics achievement over the period 1995 to 2003 is best explained by educational reform encompassing revised mathematics study programs, educational standards, and mathematics textbooks written in the TIMSS "spirit."
8. The improvement in Lithuanian students' mathematics achievements over the period of 1995 to 2003 cannot be explained by changes in students' home socio-educational environments.
9. The improvement in Lithuanian students' mathematics achievement from 1995 to 2003 can be partly explained by a change in students' attitudes toward mathematics.
10. The improvement of Lithuanian students' mathematics achievements from 1995 to 2003 can be partly explained by the fact that students learned strategies needed to answer test items in multiple-choice format.

References

Alsup, J. K., & Sprigler, M. J. (2003). A comparison of traditional and reforming mathematics curricula in an eighth-grade classroom. *Education, 123*(4), 689.

Barton, P. E. (2001). *Facing the hard facts in education reforms.* Princeton, NJ: Educational Testing Service.

Beaton, A. E., Mullis, I. V., et al. (1996). *Mathematics achievement in middle school years: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Betts, J., R., & Costrell, R. M. (2001). Incentives and equity under standards-based reform. In D. Ravitch (Ed.), *Brookings papers on education policy* (pp. 9-74). Washington, DC: Brookings Institution Press.

Čekanavičiuš V., Trakas G., & Zabulionis A. (1997). *Trečioji tarptautine matematikos ir gamtos mokslų studija. 7–8 kl. moksleivių tyrimo statistinė ataskaita.* Vilnius: Margi rastai.

Cohen, D. K., & Hill, H. C. (2000). Instructional policy and classroom performance: The mathematics reform in California. *Teachers College Record, 102*(2), 296–345.

Draper, R. J. (2002). School mathematics reform, constructivism and literacy: A case for literacy instruction in the reform-oriented math classroom. *Journal of Adolescent & Adult Literacy, 45*(6), 520–529.

Dudaitė, J. (2000). *Pagrindinės mokyklos matematikos egzaminų kaita Lietuvoje nuo 1919 iki 1999 metų.* Unpublished Master's thesis, Vilnius University.

Dudaitė, J. (Ed.). (2006). *Tarptautinis matematikos ir gamtos mokslu tyrimas TIMSS 2003. Rezultatų analizė. Matematika,* VIII klasė. Vilnius: NEC.

Dudaitė, J., Elijio, A., Urbienė, Ž., & Zabulionis, A. (2004). *Tarptautinis matematikos ir gamtos mokslų tyrimas TIMSS 2003. Ataskaita.* Vilnius: NEC.

Edmonds, R. (1987). A discussion of the literature and issues related to effective schooling. In R. Carlson, V. Robert, & E. R. Ducharme (Eds.), *School improvement—theory and practice. A book of readings* (pp. 7–46). Lanham, MD: University Press of America.

Finnan, C., Schnepel, K. C., & Anderson, L. W. (2003). Powerful learning environments: The critical link between school and classroom cultures. *Journal of Education for Students Placed at Risk, 8*(4), 391–418.

Frykholm, J. (2004). Teachers' tolerance for discomfort: Implications for curricular reform in mathematics. *Journal of Curriculum and Supervision, 19*(2), 125–149.

Fullan, M. (1998). *Pokyčių jėgos. Skverbimasis į reformos gelmes.* Vilnius: Tyto alba.

Fullan, M. (2003). *Change forces with a vengeance*. London and New York: Routledge Falmer.

Gamoran, A. (1997). Curriculum change as a reform strategy: Lessons from the United States and Scotland. *Teachers College Record, 98*(4), 608–620.

Gordon, A. B. (2004). Symbolic politics and institutional boundaries in curriculum reform: The case of National Sectarian University. *The Journal of Higher Education, 75*(5), 572–593.

Green, R. L. (1987). Tips on educational testing: What teachers and parents should know. In R. V. Carlson, & E. R. Ducharme (Eds.), *School improvement-theory and practice: A book of readings* (pp. 475–488). Lanham, MD: University Press of America.

Grissmer, D., & and Flanagan, A. (2001). Searching for indirect evidence for the effects of statewide reforms. In D. Ravitch (Ed.), *Brookings papers on education policy* (pp. 181–230). Washington, DC: Brookings Institution Press.

Hess, F. M. (2002). Reform, resistance ... retreat? The predictable politics of accountability in Virginia. In D. Ravitch (Ed.), *Brookings papers on education policy* (pp. 69–122). Washington, DC: Brookings Institution Press.

Hill, P. T., & Harvey, J. [n. d.]. *A fresh assessment: Why reform initiatives fail.* Retrieved from http://www.brookings.edu/press/books/chapter_1/makingschoolreformwork.pdf

Horne, M. (2001). Transformation not reform. *Education Review, 15*, 87–91.

Kelly, A. E., & Lesh, R. A. (2000). *Handbook of research design in mathematics and science education.* Mahwah, NJ: Lawrence Erlbaum Associates.

Kinsler, K., & Gamble, M. (2001). *Reforming schools.* London and New York: Continuum.

Krebs, A. S. (2003). Middle grades students' algebraic understanding in a reform curriculum. *School Science & Mathematics, 103*(5), 233.

Kyriakides L. (1997). Primary teachers' perceptions of policy for curriculum reform in mathematics. *Educational Research and Evaluation, 3*(3), 214–242.

Lietuvos Respublikos švietimo ir mokslo ministerija. (1997a). *Bendrojo išsilavinimo standartai. I–X klasės. Projektas. 2.* Vilnius: Lietuvos Respublikos švietimo ir mokslo ministerija.

Lietuvos Respublikos švietimo ir mokslo ministerija. (1997b). *Lietuvos bendrojo lavinimo mokyklos bendrosios programos. I–X klasės.* Vilnius: Lietuvos Respublikos švietimo ir mokslo ministerija.

Lietuvos TSR švietimo ministerija. (1988). *Vidurinės bendrojo lavinimo mokyklos programos. Matematika V–XII kl.* Kaunas: Lietuvos TSR švietimo ministerija.

Mačkeviciūtė, A., & Zabulionis, A. (2001). *Trečioji tarptautinė matematikos ir gamtos mokslų studija–1999.* Matematika. Vilnius: NEC.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report.* Chestnut Hill, MA: Boston College.

Nicol, C., Tsai, L.-L., & Gaskell, J. (2004). Students and applied academics: Learner agency in a changing curriculum. *Canadian Journal of Science, Mathematics and Technology Education, 4*(2), 209–221.

Richardson, G., Hammrich, P. L., & Livingston, B. (2003). Improving elementary school girls' attitudes, perceptions, and achievements in science and mathematics: Hindsights and new visions of the Sisters in Science Program as an equity reform model. *Journal of Women and Minorities in Science and Engineering, 9*, 333–348.

Robitaille, D. F., McKnight, C. C., Schmidt, W. H., Britton, E. D., Raizen, S. A., & Nicol, C. (1993). *Curriculum frameworks for mathematics and science* (TIMSS monograph No.1). Vancouver: Pacific Educational Press.

Rotberg, I. C. (2000). What are we doing to our schools? *Education Week, 20*(9), 44.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science & Mathematics, 102*(6), 245–253.

Trakas, G. (1997). *Testo informacija ir jos taikymai.* Retrieved from www.leidykla.vu.lt

Vann, A. S. (1993). Let's get the curriculum reform train off the bell-curve track. *Education Digest, 59*(1), 32–33.

Viadero, D. (1996). Teen culture seen impeding school reform. *Education Week, 15*(37), 1.

Zabulionis, A. (1997a). A first approach to identifying factors affecting achievement. In P. Vari (Ed.), *Are we similar in math and science? A study of Grade 8 in nine central and eastern European countries* (pp. 147–168). Amsterdam: International Association for the Evaluation of Educational Achievement.

Zabulionis, A. (1997b). Student achievement. In P. Vari (Ed.), *Are we similar in math and science? A study of Grade 8 in nine central and eastern European countries* (pp. 131–147). Amsterdam: International Association for the Evaluation of Educational Achievement.

Zabulionis, A. (2001). Similarity of mathematics and science achievement of various nations. *Education Policy Analysis Archives, 9*(33). Retrieved from http://epaa.asu.edu/epaa/v9n33

Želvys, R. (1999). *Švietimo vadyba ir kaita.* Vilnius: Garnelis.

Zybartas, S. (1999). *Matematikos mokymo lyginamoji analizė Skandinavijos šalių ir Lietuvos švietimo sistemose*, Vilnius Pedagogical University, Unpublished doctoral thesis.

The influence of item-stem format and item-answer format on the results for selected items in TIMSS 2003

Jolita Dudaitė
Kaunas University of Technology
Kaunas, Lithuania

Abstract

This paper considers the influence of the stem format and answer format of survey items on the responses (results) for those items. The items referred to in this analysis are from TIMSS 2003, conducted by the International Association for the Evaluation of Educational Achievement (IEA). The analysis focuses on the national test booklets completed by 806 Grade 8 students (from 150 schools) in Lithuania. In TIMSS 2003, Lithuania added two national booklets for mathematics, in which the TIMSS items were changed in a way that made it possible to examine the effect of the stem and answer format of an item on the results for that item. The analysis showed that Lithuanian Grade 8 students were significantly better at solving multiple-choice answer format items than open-ended answer format items. They also were less likely to omit multiple-choice than open-ended items. No difference was found in these regards between the genders. The analysis also showed that the distracters given in multiple-choice answer format items influenced students' choice of answer. When students were presented with exactly the same items but without the answers, considerably fewer of them gave the answers given in the distracters. The hypothesis that formulating the stems of the TIMSS items more in "Lithuanian style" would help students solve the items was discounted. However, in general, the wording of the stem item did affect the answers to (the results of) the item.

Introduction

When educational research involving achievement tests is conducted, items offering multiple-choice answers and items requiring open-ended answers are used. There is constant debate among researchers on what proportion of a test should include multiple-choice items and what proportion should be open-ended. Each type of item has its benefits and its faults. With multiple-choice, more material can be covered and in a shorter time than is possible with open-ended. Multiple-choice items guarantee an easy, absolutely objective and cheaper way of marking. If these items are well made, they correlate well with the results received from solving other types of items, they allow easy identification of standard mistakes, and they are especially appropriate when the choices of possible answers can be clearly determined. On the negative side, it is difficult to choose appropriate distracters for these items, and it is difficult to examine higher levels of ability (problem-raising, argumentation, etc.). Also, these items can be easily "copied," and some answers may be guesses. With open-ended items (including short-answer, short-solution, structured), there is no need to choose appropriate distracters or to consider that answers might be guesses. In addition, these items are suitable for evaluating various levels of ability. However, solving open-ended items takes a lot of time, which means tests made up of these items may not cover as many topics as tests made up of multiple-choice items. Open-ended questions are more difficult to evaluate than multiple-choice, and involve greater subjectivity and expense during marking. Often, with these items, the most typical mistakes are not identified.

Because these different answer types have their pros and cons in terms of developing and administering tests, it is important to know what impact the two types have on the answers to (results for) each type. To what extent does the answer format influence the difficulty of the item for students and to what extent does it influence students not answering (omitting) that item? Efforts to answer these questions have produced many studies, but the findings of these differ. Answer format appears to have a different effect on different teaching subjects as well as at different grade levels.

According to Elley and Mangubhai (1992), answer format does not have a statistically significant differential influence on the results for items in large-scale tests of reading ability. Hastedt's (2004) analysis of the tests for the international reading ability study PIRLS 2001 shows the opposite—that answer format does affect the results for the given items. Hastedt found that, on average, students are better able to solve multiple-choice items than open-ended items. The difference was statistically significant. Abrahamson (2000) concluded that the different answer formats of items in physics tests had no influence on the results for those items. Nasser (n. d.) claims that in tests of statistical literacy, students are worse at answering multiple-choice items than open-ended items. Gadalla (1999), considering the field of mathematics computation, found no statistically significant differences in the results for the different answer formats of items in tests given to Grades 4, 5, and 6 students. However, Grades 2 and 3 students answered the multiple-choice items statistically significantly better than they answered the open-ended items. Traub (1993) found the different answer formats had no effect on the results for tests relating to the quantitative domain. Zabulionis (1997) analyzed the TIMSS 1995 tests completed by Grades 7 and 8 students from Eastern European countries. He stated that in some of these countries, students were more likely to omit a multiple-choice format item than an open-ended format item, but that in the majority of the countries, answer format had no bearing on whether or not students omitted the item.

Because the results of these various studies differ, it is useful to add to the debate by analyzing how the answer format of the test items, as well as the stem format of the items, influenced the results for these items in the mathematics component of TIMSS 2003. The results discussed in this paper relate to those for the Grade 8 students from Lithuania who participated in TIMSS 2003.

Method

The analysis focused on the TIMSS 2003 Grade 8 national test booklets for Lithuania, answered by 806 students from 150 schools. In TIMSS 2003, Lithuania added two national booklets for mathematics, in which the TIMSS items were changed in a way that allowed us to do the following:

1. Verify the effect of the *answer format* of the item on the item results by determining:
   1.1. Whether the results for the multiple-choice answer format items differed from the results for the open-ended answer format items;
   1.2. How the distracters of multiple-choice answer format items influenced the answers to this type of item.
2. Verify the effect of the *stem format* of the item on the item results by determining:
   2.1. Whether students were better able to solve TIMSS items rephrased in the "Lithuanian style" than in the original style (by "Lithuanian style," we mean the formulations more commonly used in Lithuanian mathematics textbooks);
   2.2. Whether the students' responses to the TIMSS items differed when the phrasing of the item stem was changed in a certain way.

The changes encompassed the following:

1. The TIMSS multiple-choice answer format items were changed to open-ended answer format, and vice versa.
2. The TIMSS items were rephrased in these ways:
   2.1. The stems were rephrased in "Lithuanian style;"
   2.2. The stems were rephrased in other different ways;
   2.3. The fractions in the stems were written in words, and vice versa.
3. A few extra "TIMSS-style" items were created so that the above-mentioned ideas could be verified.
4. A number of TIMSS items remained unchanged so that we could check if the achievements of the students who answered the national booklets were identical to the achievements of the students who answered the TIMSS booklets.

In line with the main goal of the analysis in this paper, that is, verifying the effect of the answer format and the stem format of items on item results, several hypotheses were formulated:

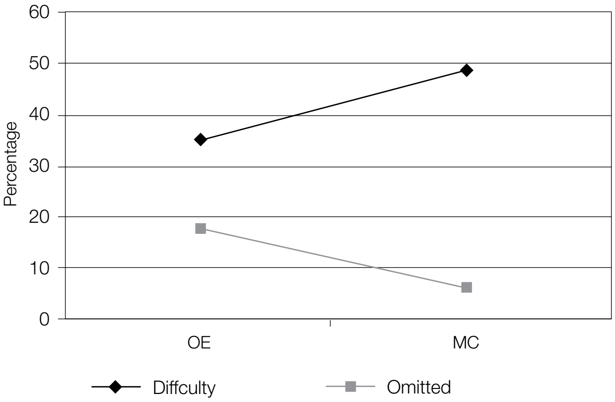
1. Students solve items in multiple-choice answer format better than items in open-ended answer format.
2. Students are more likely to omit items that require open-ended answers than they are to omit items written in multiple-choice answer format.
3. The distracters of the multiple-choice items influence choice of answer (when solving items, students are less likely to get the certain wrong answer to the questions if the questions have an open-ended answer format then they are if they have to select from the options offered under the multiple-choice format);

4. The phrasing of the stem of the item influences students' ability to answer the item correctly.
5. Students are more likely to answer items written in "Lithuanian style" than items written in standard TIMSS style.
6. Students are less likely to correctly answer items where fractions in the item stem are written in words.

Results

Grade 8 students in Lithuania were statistically significantly better at solving multiple-choice answer format items than they were at answering open-ended answer format items, and they were less likely to omit the former than the latter (see Figure 1). Forty-two items were analyzed. As we can see from the figure, the difficulty and omitting curves of the items with the multiple-choice answer format cover the curves of the items with the open-ended answer format. Thus, if required to answer an item presented in a multiple-choice answer format, students, on average, were better able to solve it and less likely to omit it than they were if the same item was presented in an open-response answer format. Figure 2 presents the average difference between a different answer format difficulty and the omitting of an item. There was no difference between the genders in solving the different answer format items, both in terms of the difficulty of the item and the omitting of the item.

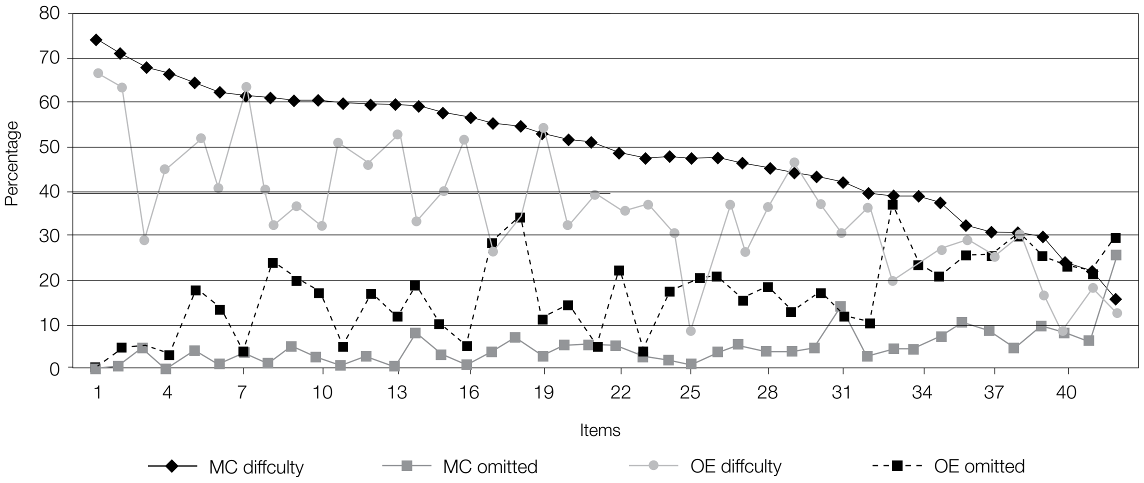
Figure 2: Average Percentages of Lithuanian Grade 8 Students Who Correctly Answered and Who Omitted Test Items Written in Multiple-choice Format and Those Written in Open-ended Format



With items in multiple-choice answer format, the given distracters influenced students' choice of answer. When students were given exactly the same items without the multiple-choice answers, considerably fewer of them provided the answers given in the distracters in the multiple-choice version. This finding suggests that the distracters, which are considered typical mistakes, need to be evaluated with much care.

Figure 3 presents the results for one such item. (Chi square analysis shows that the answers students gave to an item depended on its answer format; $\chi^2 = 481.368$, df = 6, p = 0.000.) We can see from the figure that with the multiple-choice case, most students chose the

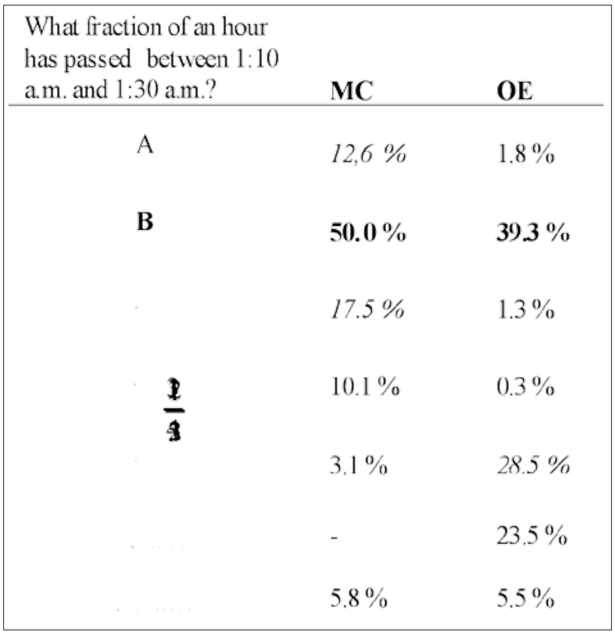
Figure 1: Percentages of Lithuanian Grade 8 Students Who Correctly Answered and Who Omitted Test Items Written in Multiple-choice Format and Those Written in Open-ended Format



wrong answers, C and A. Their choice of distracter A suggests they "forgot" that an hour consists of 60 minutes, not 100. This mistake is a common one in basic school. Those students who choose distracter C were perhaps misled by the "similarity" of 20 minutes and 1/2. With the open-ended case, we can see that almost nobody gave the wrong answers 1/5 and 1/2 (given in distracters A and C in the multiple-choice format case), but that many of the students gave answer 3/4 (which was the distracter E). These results indicate the different typical mistakes that students tend to make when answering the same question presented with different answer formats.

The wording of the stem of an item also influences the answers students give to that item. However, it is difficult to determine what change in the stem might make the item easier or more difficult for students to answer correctly. For example, when the stem was taken away from one item ("calculate the expression and write down the answer in decimals") and only the expression left, which needs to be calculated (add two simple fractions), 20% more students were unable to solve the item than was the case when the item included the worded stem by the expression. But in another identical item (to subtract two simple fractions), 20% more students were able to solve the item after removal of the worded stem.

Figure 3: Difference in Students' Answers for the Same Question Presented in Multiple-choice and Open-ended Answer Formats



What fraction of an hour has passed between 1:10 a.m. and 1:30 a.m.?	MC	OE
A	*12,6 %*	1.8 %
B	**50.0 %**	**39.3 %**
	17.5 %	1.3 %
[illegible]	10.1 %	0.3 %
	3.1 %	*28.5 %*
	-	23.5 %
	5.8 %	5.5 %

Rephrasing some of the TIMSS items in different ways produced various results. For example, students were less able to answer an item correctly when the unknown quantity (x) rather than the number was written. Similarly, students were better able to solve an item when minutes instead of the part of an hour were written.

By making the stems of TIMSS items more Lithuanian-like, we hoped that the more familiar wording of the stem would help more students solve the item. However, while students were better able to solve some of these rephrased items (for an example, see Table 1), they were less able to solve others correctly (see Table 2), and for some items, students were just as likely to answer the Lithuanian-styled item incorrectly as they were the originally worded item. Although rewording the item stems brought the familiarity of Lithuanian style to the students, it is possible that, with some items, rewording produced a longer and more complicated stem, which may have influenced students' ability to answer the item correctly.

Rephrasing the item stem by writing down the fraction in words was the change that most often altered the difficulty of the item. In most cases, students were less able to solve items worded in this way, in some cases students were just as likely to solve the originally worded as the rephrased items, and, in a very few cases, students were better able to solve the item when it was reworded. Figure 4 provides an example in which students were considerably less able to solve the item when the fraction (1/4) was written in words (a quarter) (the answers depend on the item format: $\mathcal{X}^2$ = 102.358, df = 5, p = 0.000). Here, we can see that with the item in the left-hand column, the main mistake was distracter A, which meant that when doing their calculation, students used just two fractions—1/2 and 1/5 (i.e., those written like fractions). These students did not recognize "a quarter" as a number. Almost 10% of the students chose distracter B, again indicating that students did not use "a quarter" when doing their calculation. In the case in the right-hand column, we can see that fewer students chose distracters A and B. Here, the main mistake was distracter C, which clearly shows that the students were using all three fractions to do their calculation.

Table 1: Example of a TIMSS Item that Students Were Better Able to Solve after It Had Been Reworded in "Lithuanian Style"

If 4(x+5)=80, then x=	TIMSS stem	Solve the equation: 4(x+5)=80	Lithuanian stem
True	**50.9%**	True	**59.3%**
False	30.9%	False	27.3%
Omitted	16.6%	Omitted	10.1%

Note: x^2 = 19.521, df = 2, *p* = 0.000.

Table 2: Example of a TIMSS Item that Students Were Less Able to Solve after It Had Been Reworded in "Lithuanian Style"

Which of these is 370·998+370·2?	TIMSS stem	Carry out the number before the parenthesis: 370·998+370·2. Which expression will you get after the calculation?	Lithuanian stem
A 370·1000	**47.5%**	A 370·1000	**35.9%**
B 372·998	4.5%	B 372·998	5.2%
C 740·998	17.1%	C 740·9998	12.0%
D 370·998·2	28.2%	D 370·998·2	37.0%
Omitted	2.4%	Omitted	3.7%

Note: x^2 = 19.521, df = 2, *p* = 0.000. .

Conclusions

1. The difficulty of the open-ended answer format items was higher than that of the multiple-choice answer format items.
2. Students more frequently omitted the open-ended answer format items than the multiple-choice answer format items.
3. The distracters of the multiple-choice items influenced the answers that students gave. When answering open-ended answer format items, students were less likely to get the certain wrong answer than they were when having to choose from the possible answers listed for a multiple-choice question.
4. The wording of an item's stem statement influenced students' ability to solve the item.
5. Students were no more likely to correctly answer items written in "Lithuanian style" than they were to correctly answer items written in original "TIMSS style."
6. Students were less likely to answer items successfully when the item-stem used fractions written in words.

References

Abrahamson, M. (2000). *KSU studies the effects of multiple-choice format on the FCI.* Retrieved from www.bedu.com

Elley, W. R., & Mangubhai, F. (1992). Multiple-choice and open-ended items in reading tests: Same or different? *Studies in Evaluation, 18*, 191–199.

Gadalla, T. M. (1999). *Multiple-choice versus constructed-response tests in the assessment of mathematics computation skills.* (ERIC Document Reproduction Service No. ED431813)

Hastedt, D. (2004). Differences between multiple-choice and constructed response items in PIRLS 2001. In C. Papanastasiou (Ed.), *Proceedings of the IEA International Research Conference 2004: PIRLS* (Vol. 3). Nicosia: University of Cyprus Press.

Nasser, F. (n. d.). *On the relationship between test format, attitudes towards and performance in a statistics test.* Retrieved from http://www.stat.auckland.ac.nz/~ iase/publications/4/618.pdf

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennet & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 29–44). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zabulionis, A. (1997). Student achievement. In P. Vari (Ed.), *Are we similar in math and science? A study of Grade 8 in nine central and eastern European countries* (pp. 99–146). Budapest: International Association for the Evaluation of Educational Achievement.

Differences in science teaching and learning across Australian states

John Ainley and Sue Thomson
Australian Council for Educational Research
Camberwell, Victoria, Australia

Introduction

This paper examines the differences among and within the Australian States in science teaching and learning based on the analysis of data from TIMSS. It focuses on science achievement at Grade 8 in 2002. The paper begins with a consideration of the differences among states in science achievement at Grade 4 and Grade 8 and the way in which patterns changed between 1994 and 2002. It then examines the influence of factors operating at state, school, and student levels on science achievement at Grade 8 in the national picture and the way those influences differ among states. It concludes with a discussion of the factors influencing Grade 8 science achievement.

Context

Australia's national goals for schooling assert that by the time students leave school they should have attained high standards of knowledge, skills, and understanding in eight key learning areas: the arts; English; health and physical education; languages other than English; mathematics; science; studies of society and environment; and technology. The Performance Measurement and Reporting Taskforce (PMRT)[1] administers the National Assessment Programme (NAP). This defines key performance measures and monitors progress toward the achievement of the national goals (MCEETYA, 2005).

The Trends in International Mathematics and Science Studies (TIMSS) is defined as part of the NAP. For each cycle of TIMSS, extensive national reports are produced that detail the pattern of results for Australia (Thomson & Fleming, 2004a, 2004b; Lokan, Hollingsworth, & Hackling, 2006; Lokan, Ford, & Greenwood, 1996, 1997). In addition, the NAP incorporates annual assessments of literacy and numeracy that use the full population of students at Grades 3, 5, and 7. Assessments for civics and citizenship and for ICT literacy are conducted every three years for sample surveys of students in Year 6 and Year 10. For science, there is a sample survey at Grade 6 every three years.

Australia has a federal system of government, with states having the major responsibility for education. There are differences among states in educational organization and curriculum in many fields, including science education, and there is increasing interest in examining the differences among states in fields such as science and mathematics. There are also differences among the states in the age of students at any given grade and in the demographic characteristics of the population. Table 1 contains an indication of some of these variations.

Boosting science learning has become a priority of the federal government. A national review of the quality and status of science education in Australian schools concluded that there was a gap between the ideal and reality, especially in secondary schools and particularly in relation to the teaching of science as scientific literacy (Goodrum, Hackling, & Rennie, 2001). The TIMSS 1999 Video Study reported that Australian lessons were characterized by a core pedagogical approach that involved analyzing data gathered through independent practical activity and focusing on connections between ideas and real-life experiences (Lokan, Ford, & Greenwood, 2006).

There is no common school curriculum in science across the country, although there is a non-mandatory national statement of learning in science that outlines the learning opportunities that should be provided at each stage of schooling from Grade 1 to Grade 10 (Curriculum Corporation, 2006). A national online science assessment resource-bank (SEAR) has been developed for use by schools to support science teaching. Within states, the pattern is that central authorities specify broad curriculum frameworks and schools have considerable autonomy in deciding curriculum detail, textbooks, and teaching

[1] Established by the Ministerial Council for Education, Employment, Training, and Youth Affairs (MCEETYA).

Table 1: Population Characteristics of Australian States

	Average age [a]	IRSED [b]	% LBOTE [c]	% Metro [d]
New South Wales	14.0	1000	15	69
Victoria	14.1	1016	16	72
Queensland	13.4	989	7	62
South Australia	13.8	994	7	65
Western Australia	13.4	996	8	65
Tasmania	14.2	969	1	49
Northern Territory	13.8	903	1	20
Australian Capital Territory	14.1	1076	10	99
Australia	13.9	1000	11	66

Notes: a. Based on data recorded in the TIMSS Australia report for science (Thomson & Fleming, 2004b).

b. Based on the "Socioeconomic Indexes for Areas: Index of Relative Socioeconomic Disadvantage" (IRSED) (ABS, 2004).[11] The national mean for collection districts is 1,000, and the standard deviation is 100.

c. Percentage of Year 9 students for whom the main language spoken at home is other than English. Data are from the longitudinal surveys of Australian Youth (LSAY).

d. Percentage of Year 9 students living in a metropolitan zone according to the MCEETYA three-category classification of geolocation. Data are from the longitudinal surveys of Australian Youth (LSAY).

methodology. Learning materials and tests are prepared by a variety of agents, including the curriculum sections of state education departments, academics, commercial publishers, and teachers' subject associations. As a consequence, what is taught in science varies between states and between schools within states. There are also variations among states and among schools within states in the amount of time allocated to science in the junior secondary grades and specifically in Grade 8.

Differences in measures of science achievement need to be interpreted in relation to curriculum and policy differences at state level, differences in educational practices at school and classroom levels, and differences in student characteristics. This paper presents an analysis of the ways in which various aspects of science teaching impact on student learning.

Data

The international sample design for TIMSS is a two-stage stratified cluster sample design (Martin, Mullis, Gonzales, & Chrostowski, 2004). The first stage consists of a sample of schools, which in Australia is stratified by state and by sector (with disproportionate sampling across strata, followed by weighting). Nationally, non-government schools enroll 33% of students (29% of elementary students and 38% of secondary students). The second stage consists of a random sample of one classroom from the target grade in each sampled school. The numbers of students in TIMSS 2002 for Population 1 and Population 2, along with the number of schools, are shown in Table 2.[2] In the achieved sample for Grade 8, there were 5,335 students from 210 schools; for Grade 4, there were 4,675 students from 204 schools.

Analysis

Results from two types of analysis are reported in this paper. The first is a comparison of means. The comparison of means is based on weighted data so that the distribution of students accurately reflects the population, and uses jackknife replication techniques to properly estimate standard errors from complex samples.[3] Because the estimation of errors is based on rigorous procedures and because our focus is on implications for policy and practice (rather than on establishing laws), we have commented on results significant at the 10% level.

The second type of analysis is based on multi-

[1] The Index of Relative Socioeconomic Disadvantage (IRSED) is one of the Socioeconomic Indexes for Areas (SEIFA) computed from census data for collection districts and published by the Australian Bureau of Statistics (ABS, 2004). For all SEIFA indexes, the mean is 1,000 and the standard deviation is 100. A higher score indicates greater average advantage.

[2] Australia achieved the required participation rate of 85% of sampled schools and 85% of sampled students (or a combined schools and students participation rate of 75%) for Grade 8 but just fell short of the minimum requirements for Grade 4. Sampling weights were calculated by Statistics Canada to ensure that the population at each year level was appropriately represented by the students participating in TIMSS.

[3] For this analysis, we used the program AM developed by the American Institutes for Research.

Table 2: Australia's Designed and Achieved Sample in TIMSS 2002

State	Design school sample	Grade 4				Grade 8			
		N schools	*N* students	Weighted percent	Weighted *N*	*N* schools	*N* students	Weighted percent	Weighted *N*
NSW	40	35	912	90,781	35.3	34	880	84,456	32.8
VIC	35	32	675	62,852	24.4	34	860	65,435	25.4
QLD	35	31	759	43,597	16.9	33	881	48,270	18.8
SA	30	27	600	20,901	8.1	28	703	18,902	7.3
WA	30	27	661	26,123	10.2	26	702	27,616	10.7
TAS	30	25	501	6,444	2.5	26	625	6,424	2.5
NT	15	13	251	2,300	0.9	14	321	1,578	0.6
ACT	15	14	316	4,224	1.6	15	383	4,727	1.8
Total	230	204	4675	25,7222	100.0	210	5,355	257,408	100.0

level regression analysis[4] that examines the patterns of association between science achievement measured on the TIMSS science scale and predictors measured at the level of the student, the school (or classroom), and the state. Two main forms of Hierarchical Linear Modeling (HLM) are reported. The first form was a three-level national analysis with predictors considered at state, school/classroom, and student levels. This analysis provided information about national patterns of influences on science achievement. The second form was a series of replicated within-state two-level models, with predictors considered at the school/classroom and student levels. These analyses established whether the effects of the predictors were similar or not across different contexts.

The analyses refer to the school/classroom level because of the nature of the sampling for TIMSS in Australia. Within-school sampling is based on the random selection of one mathematics class per school. Grade 8 students from that class may either be in the same class for science or be dispersed among different classes for science (with some of those classes containing very few sampled students). Much of the data of interest to these analyses is based on information provided by teachers. We chose to aggregate these data to school level so as to ensure stability in the level-two unit.

The following predictor variables were included at the student level in the final model.[5]

- Gender was coded with males as 0 and females as 1. Fifty-one percent of respondents were female.
- Age was expressed in months. The mean age was 166 months (13 years, 10 months) with a standard deviation of six months.
- Indigenous status was coded with non-Indigenous students as 0 and Indigenous (Aboriginal or Torres Strait Islander) students as 1. Three percent of the sample was Indigenous.
- Language spoken at home was coded so that where a language other than English was the main language, the code was 0 and where English was the main language, the code was 1. Eight percent of students spoke a language other than English at home.
- Parental education was coded so those whose parents who had not reached university level were coded as 0 and those whose parents had participated in university education were coded as 1. Twenty-one percent of the sample had parents who had attained university level.

The following predictor variables were included at the school/classroom level in the final model.[6]

- Participation in professional development on science assessment was recorded as the average proportion of science teachers in the school who had participated in such programs over the past two years. Teacher responses were initially coded as 0 for non-participation and 1 for participation. The mean school level of participation was 0.51.

[4] For this analysis, we used the program Hierarchical Linear Modeling Version 6.03 (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004).

[5] Location coded as metropolitan or non-metropolitan was included in the initial analysis but dropped from the final model because there was no significant effect.

[6] A large number of variables relating to teacher qualifications and other aspects of teaching were included in the initial analyses but were dropped from the final model.

- The extent to which teachers reported that students were required to "write explanations about what was observed and why it happened" when doing science investigations in half their lessons or more was coded 0 for not reporting that and 1 for reporting it. The among-school mean proportion of teachers recording this emphasis was 0.62, with a standard deviation of 0.39.
- The percentage of time spent teaching physics was recorded as the within-school average for responding science teachers. The among-school mean was 21.8%, with a standard deviation of 7.4%.
- Homework emphasis was represented as a variable indicating the proportion of teachers at the school who recorded a high or medium emphasis on homework (assigning it in more than half the lessons) in science. The mean proportion was 0.93.

The following predictor variables were included at the state level in the final model.

- The average time in minutes allocated to science each week for the state was based on data provided on the teacher questionnaire. The national mean was 198 minutes, but the values ranged from 169 to 223 minutes per week.
- The average age in months for students in Grade 8 was recorded for each state.

Results

Differences among states

Table 3 records the mean achievement scores in science for all states of Australia. At Grade 4 level, there were essentially no differences in TIMSS science achievement among the states. The difference in the means for Victoria and New South Wales, the two most populous states, and those most comparable in terms of student age and demographic characteristics, was only two scale points. The only difference that was statistically significant at Grade 4 was between the Australian Capital Territory and Western Australia. At Grade 8, the significant differences in TIMSS science achievement were between New South Wales and Victoria (31 scale points), New South Wales and the Northern Territory (65 scale points), and between the Australian Capital Territory and the Northern Territory (56 scale points).

On the national science assessment at Grade 6, students from the Australian Capital Territory achieved a significantly higher mean score than did students from all the other states and territories except New South Wales and Tasmania. Students from New South Wales achieved a significantly higher mean score than students from all the other states and territories except the Australian Capital Territory, Tasmania, and

Table 3: TIMSS 2002 Science Scores and National Science Assessment Scores for Australian States

	TIMSS 02 Grade 4		TIMSS 02 Grade 8		National Test 03 Grade 6	
	Mean	*SE*	Mean	*SE*	Mean	*SE*
New South Wales	526	10.1	547	9.6	411	4.1
Victoria	528	6.8	516	5.3	399	4.2
Queensland	513	7.7	516	6.0	392	3.8
South Australia	515	8.5	524	10.9	393	4.1
Western Australia	502	7.3	520	6.9	390	4.8
Tasmania	517	11.6	504	11.7	407	6.1
Northern Territory	503	13.8	482	13.7	379	9.9
Australian Capital Territory	547	9.7	538	9.2	430	6.2
Australia	521	4.2	527	3.8	400	1.9

Notes: The international mean for Grade 4 was 489, with a standard deviation of approximately 100.
The international mean for Grade 8 was 474, with a standard deviation of approximately 100.
The national test was calibrated to have a mean of 400 and a standard deviation of 100.

Victoria. There were no significant differences among the performances of students from Victoria, Western Australia, South Australia, Queensland, and the Northern Territory.

Changes between 1994 and 2002

Between 1994 and 2002, there was a small increase in the average scale score in TIMSS science for Grade 8 students in Australia—from 514 to 527. This increase was significant at the 5% level. For Grade 4 students in Australia, the average science scores did not change at all (being 521 on both occasions).

As shown in Table 4, there were differences between states in the extent of the change from 1994 to 2002. In New South Wales, there was an improvement in the average Grade 8 score of 30 scale points, and in Victoria there was an improvement of 19 scale points.[7] Although neither gain was statistically significant at the 5% level, both were significant at the 10% level. There were declines in student science achievement in Western Australia of 25 points at Grade 4 (just failing to reach the 5% level) and 10 points at Grade 8 (but these declines were not statistically significant).

Multi-level regression analysis of Grade 8 science scores in 2002

The results of a three-level regression analysis for the Australian Grade 8 TIMSS science data are recorded in Table 5.[8] In this analysis, the dependent variable was the TIMSS science scale score. The results show all predictor variables that were statistically significant at the 10% level. A number of potential predictors that were examined and found to be not statistically significant at the 10% level are not shown.

State-level predictors

Although state-level variables contributed little to the percentage of variance in student scores (there were only eight units at this level), there were some moderately strong effects evident in relation to the state means. Two state-level factors were related to student science achievement.

Table 4: TIMSS Science Scores in 1994 and 2002 for Australian States

	TIMSS 1994		TIMSS 2002	
	Mean	*SE* (Mean)	Mean	*SE* (Mean)
Grade 4				
New South Wales	522	6.1	526	10.1
Victoria	529	10.7	528	6.8
Queensland	503	7.6	513	7.7
South Australia	519	7.1	514	8.5
Western Australia	527	6.1	502	7.3
Tasmania	523	8.7	517	11.6
Northern Territory	512	11.2	503	13.8
Australian Capital Territory	557	6.0	547	9.6
Australia	514	3.9	527	3.1
Grade 8				
New South Wales	517	8.2	547	9.6
Victoria	497	6.2	516	5.3
Queensland	510	8.4	516	6.0
South Australia	510	5.9	524	10.9
Western Australia	531	6.6	520	6.9
Tasmania	496	10.7	504	11.7
Northern Territory	466	16.8	482	13.7
Australian Capital Territory	529	12.7	537	9.2
Australia	521	3.8	521	4.2

[7] New South Wales and Victoria are the two most populous states and together enroll just under 60% of the student population in Australia.

[8] For this analysis, we used Hierarchical Linear Modeling Version 6 (Raudenbush et al., 2004).

Table 5: Results of a Three-Level Regression Analysis of Grade 8 TIMSS Science Achievement in Australia

	Coefficient	Standard error	*p*-value
Student level			
Gender (male = 0, female =1)	-13.35	1.87	0.00
Age (months)	-1.04	0.19	0.00
Indigenous (non-Indigenous = 0, Indigenous =1)	-39.26	5.07	0.00
Language at home (LOTE = 0, English = 1)	23.97	3.49	0.00
Parental education (non-university = 0, university =1)	23.16	2.22	0.00
School/class level			
Teacher participation in PD on science assessment	-14.14	8.23	0.09
Explanations about observations in science investigations	27.29	9.75	0.01
Percentage of science time on physical science	1.29	0.44	0.00
Teacher emphasis on science homework	67.07	17.12	0.00
State level			
Average time allocated to science (minutes)	0.56	0.16	0.02
Average age in Grade 8 (months)	2.44	0.90	0.04
Initial variance			
Level 1	37%		
Level 2	62%		
Level 3	1%		
Variance explained by the model	15%		
Level 1		6%	
Level 2		27%	
Level 3		100%	

Note: *N* = 4,737 Level 1 units, 205 Level 2 units, and 8 Level 3 units.

- The average time allocated to science for the state was related to science achievement. Each additional 10 minutes of time (the national average was 195 minutes per week) was associated with an increment of just under six scale points, other factors being equal.
- Average age for the state was related to science achievement, with each additional six months being associated with approximately 14 scale points, other factors being equal.

School/class-level predictors

School- or classroom-level influences accounted for 10% of the variance in student science scores. Four of the school/classroom-level predictors were significantly related to science achievement.

- The percentage of science classroom time allocated to physics was related to science achievement (the average was 21% with a standard deviation of 7%). For each additional 20% of time allocated to physics within science, there was a net gain of 26 scale points.
- In instances where most teachers placed a moderate or high emphasis on science homework, there was a gain in science achievement. For each additional 20% of teachers who placed a moderate or high emphasis on homework, the net gain was 13 scale points. The difference between schools where all teachers placed this emphasis on homework and those where none did was 67 scale points.
- Students from schools where a higher proportion of teachers indicated half or more science lessons involved formulating hypotheses performed better than their peers from other schools. The net gain was six scale points for every additional 20% of teachers indicating this emphasis. The difference between schools where all teachers indicated this and those where none did was 27 scale points.
- Students from schools where more teachers participated in professional development focused on assessment performed better than students from schools where teachers did not engage in this form of professional development. For each additional 20% of teachers who participated, the difference was three scale points. The difference between schools where all teachers participated and those where none did was 14 scale points.

Student-level predictors

Student background factors contributed to more than one third of the variance in student science scores. Five student characteristics were related to science achievement at Grade 8.

- Male students performed better than female students (51% of students) on the TIMSS science assessment in Grade 8, with the net difference being 13 scale points.
- Non-Indigenous students performed better than Indigenous students (3% of students), with the net difference being 39 scale points.
- Students for whom English was the main language spoken at home (93% of students) performed better than students for whom a language other than English was the main language at home. The net difference was 24 scale points.
- Students whose parents had experienced a university education (18% of students) performed better than those whose parents had not proceeded to university. The net difference was 23 scale points.
- The younger students in Grade 8 performed better than the older students, with each additional month being associated with a one-scale-point difference in achievement.

Differences among states in influences on science achievement

The multi-level analysis conducted for the Australian sample found that gender, parental education, Indigenous status, language at home, and age related to achievement on the TIMSS science scale for Grade 8. As reported by the Australian Bureau of Statistics (2004), student background differs among the states and territories. The extent to which student background influenced the pattern of differences in TIMSS science achievement among the states and territories depended on:

- The magnitude of the relationship between a characteristic and student science achievement.
- The extent to which the distribution of that characteristic differed among states. We found, for example, differences among the states in the distribution of parental occupation but not of gender.

Regression analysis was used to examine the effect on the means after controlling for specified student background characteristics. The student background characteristics included in the analysis were age, parental education, Indigenous status, gender, and home language. All of the background characteristics were measured at the student level, but the analysis was conducted using two-level HLM to make allowance for the clustered sample design in estimating standard errors. The regression analyses were conducted separately for each state so that the adjustment took account of the effects of the variable in each state. The intercepts from the regression provided an indication of the adjusted means for each state. Results are recorded in Table 6. When interpreting the results, it is necessary to focus on the magnitude of the effects. This is because the significance level depends on the numbers in the sample and, more especially, the numbers of students with a given characteristic in the state.

The results presented in Table 6 can be summarized as follows:

- The influence of gender (males performing better than females) is evident in five of the eight states. It is non-existent in South Australia and Tasmania, not significant in the Australian Capital Territory, and strongest in Western Australia and Queensland.
- The influence of Indigenous status is largest in Western Australia and moderately large in Tasmania and the Northern Territory. The influence is relatively weaker in South Australia. Note that the numbers of Indigenous students are very small in Victoria and the Australian Capital Territory, and therefore the effect is not significant. In Queensland, the effect is not significant because the dispersion is wide.
- Parental education is related to science achievement in all states but is stronger in Tasmania and the Northern Territory and less strong in New South Wales, Western Australia, and Queensland. This finding can be taken as some indication of inequality in the social distribution of science achievement.
- Being from a home where a language other than English is the main language spoken has no effect in South Australia and Tasmania (where there are fewer students in this category) but a stronger effect in Queensland and the two territories. The effect is similar in Victoria and New South Wales, the states that have the highest proportions of students from a non-English speaking background.
- Age within grade has effects in Victoria, Queensland, and the Australian Capital Territory (with older

Table 6: Results of Two-Level Regression Analyses of Grade 8 TIMSS Science Achievement in each Australian State

	Intercept	Gender	Indigenous	Parent education	Home language	Age	% variance explained by model
New South Wales	**516.8**	**-13.6**	-13.7	17.6	25.6	-0.5	6%
Victoria	**494.4**	**-8.4**	-15.9	**31.7**	**21.4**	**-1.4**	5%
Queensland	**481.3**	**-19.3**	-20.7	**21.5**	**43.0**	**-1.7**	10%
South Australia	**510.2**	-0.2	**-19.4**	**28.2**	5.5	-1.2	11%
Western Australia	**510.4**	**-20.4**	**-89.9**	**17.5**	**24.6**	-0.5	8%
Tasmania	**484.3**	0.3	**-44.6**	**45.2**	11.7	0.5	11%
Northern Territory	**439.0**	**-14.2**	**-34.8**	**46.6**	**52.2**	-1.5	11%
ACT	**486.4**	**-13.4**	-49.1	**21.0**	**52.8**	**-1.5**	9%

Note: Coefficients in bold are statistically significant at the 5% level.

students performing a little less well than younger) but not elsewhere.

The intercepts from these analyses can be taken as related to the adjusted means for each state, based on the effects of each factor on achievement for that state. It can be seen from Table 6, when compared with Table 2, that the adjustment process does not alter the relative order of the states greatly (the correlation coefficient between adjusted and unadjusted scores is 0.81).

Discussion

The discussion of the results from the analyses conducted for this paper focuses on the junior secondary years because that is where most of the variation at state level is evident. These analyses of Australian Grade 8 TIMSS science data provide perspectives on three sets of issues. The first concerns the factors that influence science achievement in the junior secondary years in Australia. The second set focuses on the extent to which there are differences among states in the influence of various factors on science achievement. The third concerns the extent to which there are differences among states (as the formal locus of authority for decisions about policy organization and curriculum) in science achievement in secondary schools.

In general, secondary school science achievement appears to be influenced by factors at state, school and classroom, and student levels. Our analysis of the TIMSS data showed that, at state level, science achievement was influenced by the average age of students and the average amount of time allocated to the teaching of science. Although these variables account for a minute percentage of the variance in student scores, the effect sizes are not trivial. In systems where the average age of students was higher, then achievement was higher. The gap between states with the youngest and the oldest Grade 8 students respectively was nine months, which corresponds to an effect of 22 scale points. The average time allocated to science also related to science achievement: the range in allocated time was more than 50 minutes per week, which corresponds to approximately 31 scale points. The average age of students in Grade 8 reflects the age at which students commence school and is only alterable by a major change to the structure of schooling. Allocated time is more readily susceptible to policy changes, and increasing the time allocated to science on a state-wide basis could result in improved science learning.

It is interesting that time was not a significant influence at school level, possibly because variations within states in allocated time are much less (even where time is not prescribed, state patterns of allocating time between learning areas have been evident for several decades). We could hypothesize that allocating more time at state level would allow a greater breadth and depth to what students are expected to learn, and that this would be reflected in what they do learn. Variations among schools within states are less likely to reflect variations in the breadth and depth of the curriculum coverage.

School or classroom influences included in this analysis accounted for approximately 10% of the variance in student science scores. Three pedagogical factors relate to science achievement: the percentage of time allocated to physics within science teaching;[9] the extent to which students are required to provide written explanations about observations made during science investigation; and the emphasis on homework. Taken together, these factors can be seen as manifestations of an emphasis on science learning that requires a relatively high level of cognitive engagement with science content. The finding that teacher participation in assessment-related professional development had a negative influence on student achievement possibly reflects a similar orientation, namely, that professional development focus is less orientated to the deeper understanding of science than are other forms of professional development.

At the national level, the student-level influences on science achievement followed a pattern similar to that found in many Australian studies of science. Male students achieved higher scores than their female counterparts, and by a larger margin than in most other countries. Students whose home language was not English performed less well than those whose home language was English (the net effect was 24 points), and Indigenous students performed less well than non-Indigenous students (the net effect was 39 points). Students whose parents had completed a post-secondary qualification performed better than students whose parents had not, with the net effect being 23 scale points.

There is a tendency to regard the influence of student-background characteristics on achievement as almost immutable. However, one of the benefits of international studies is that they allow us to demonstrate that these effects are not immutable and that the strength of the influence varies between countries that are comparable in other ways. The results presented in this paper indicate variations among the Australian states in the influence of various background characteristics on science achievement. The difference in achievement between males and females was non-existent in South Australia and Tasmania but substantial in Western Australia and Tasmania. The effect of parental education was less in New South Wales and Western Australia than elsewhere. Home language had nearly twice the effect in Queensland as it did in Victoria. Importantly, the achievement gap between Indigenous and non-Indigenous students was much greater in Western Australia than in New South Wales. Differences such as these invite further inquiry into the differences in demographics, social policy, and educational practice that are associated with these disparities.

We observed near the start of this paper the significant differences in TIMSS science achievement at Grade 8 between New South Wales and Victoria.[10] The relatively strong performance of students from New South Wales is also reflected, but less strongly, in results from the national sample survey of science literacy at Grade 6. At Grade 4, there are almost no differences between Victoria and New South Wales. In addition, it is pertinent to note that the relatively high achievement of students from New South Wales emerged over the period from 1994 to 2002. Although both Victoria and New South Wales improved over that time, New South Wales improved to a greater extent.

The differences between these two states do not appear to arise from differences in the age of students (the average age is similar, and the effect of age on achievement is similar). Nor do they arise from gender (although the gender gap is smaller in Victoria than in New South Wales) or language background (the composition of the population in the two states and the effects of language are almost identical). There is a difference in the effect of parental education, with New South Wales exhibiting less of an effect of parental education, a finding that may reflect the smaller percentage of students in non-government schools. However, this difference had very little influence on the state average score.

The difference between New South Wales and Victoria in the average time allocated to science would account for a substantial part of the difference between the two states (possibly 23 of the 31-point difference). The school- and classroom-level influences identified in these analyses do not differ substantially between these states. The remaining differences possibly reside in factors not captured in these data, such as the extent to which there is a strong curriculum framework that shapes teaching in the schools.

[9] Compared to Grade 8 students in other countries, Australian Grade 8 students performed relatively better in life, earth, and environmental sciences and relatively worse in chemistry and physics (Martin et al., 2004).

[10] Significant differences were also evident between New South Wales and the Northern Territory and between the Australian Capital Territory and the Northern Territory.

References

Australian Bureau of Statistics (ABS). (2004). *Information paper: Census of Population and Housing: Socio-economic indexes for areas, Australia* (Catalogue 2039.0). Canberra: Author.

Curriculum Corporation. (2006). *National consistency in curriculum outcomes: Draft statements of learning and professional elaborations for science.* Melbourne: Author.

Goodrum, D., Hackling, M., & Rennie, L. (2001). *The status and quality of teaching and learning of science in Australian schools.* Canberra: Department of Education, Employment, Training and Youth Affairs.

Lokan, J., Ford, P., & Greenwood, L. (1996). *Maths & science on the line: Australian junior secondary students' performance in the Third International Mathematics and Science Study* (TIMSS Australia Monograph No. 1). Melbourne: Australian Council for Educational Research.

Lokan, J., Ford, P., & Greenwood, L. (1997). *Maths & science on the line: Australian middle primary students' performance in the Third International Mathematics and Science Study* (TIMSS Australia Monograph No. 2). Melbourne: Australian Council for Educational Research.

Lokan, J., Hollingsworth, H., & Hackling, M. (2006). *Teaching science in Australia* (TIMSS Australia Monograph No. 8). Melbourne: Australian Council for Educational Research.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international science report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grade.* Chestnut Hill, MA: Boston College.

Ministerial Council for Education, Employment, Training and Youth Affairs (MCEETYA). (2005). *A measurement framework for national key performance measures.* Melbourne: Author.

Raudenbush, S., Bryk, A., Cheong, Y. F., Congdon, R., & du Toit, M. (2004). *HLM6: Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International.

Thomson, S., & Fleming, N. (2004a). *Summing it up: Mathematics achievement in Australian schools in TIMSS 2002* (TIMSS Australia Monograph No. 6). Melbourne: Australian Council for Educational Research.

Thomson, S., & Fleming, N. (2004b). *Examining the evidence: Science achievement in Australian schools in TIMSS 2002* (TIMSS Australia Monograph No. 7). Melbourne: Australian Council for Educational Research.

Exploring student-level explanatory factors in science achievement in South African secondary schools

Sarah Howie, Vanessa Scherman, and Elsie Venter
University of Pretoria, South Africa

Abstract

South Africa's education system is still deep in the throes of reform under its third Minister of Education since 1994. However, it is marked by underachievement of students at each level of the system. Poor communities, in particular, those of rural Africans, bear the brunt of the past inequalities, and these are continually reflected in the national results of the final-year examinations in Grade 12. Equity and access are at the top of the government's priority list, which makes this paper important, as it attempts to analyze the progress made in terms of equity in education. South Africa participated in TIMSS 1995, 1999, and 2003. Secondary analyses of these studies have revealed the large inequities in the education system, with 55% of the variance in the students' mathematics scores explained by differences between schools (Howie, 2002). This variance, in turn, is mostly explained by the historical inequities imposed on communities and schools over the past 40 years. The challenge was to explore the extent of the "gap" in students' scores by comparing the advantaged and disadvantaged communities in this context, namely students in better-resourced, largely urban schools and students in largely under-resourced, black rural schools. The TIMSS-Repeat 1999 data were explored to focus on the extent of the gap in student science achievement between advantaged and disadvantaged communities and specifically on the factors at student level that predicted the outcomes in science in both types of communities. Three categories of students were ultimately identified. These were advantaged, semi-advantaged, and disadvantaged groups. Partial least square analysis was applied to explore the science performance in order to identify factors that predicted science performance across and within these groups. Very few factors were found that consistently predicted performance across and within these groups. However, one dominant factor emerged in these models and that was the students' performance in the English test that provided a measure of students' proficiency in English, the language in which more than 70% of the students wrote the science tests. Students who had a higher score on the English test also performed better in the science test, despite their backgrounds.

Introduction

South Africa's education system is still deep in the throes of reform under its third Minister of Education since 1994. However, it is marked by underachievement of students at each level of the system. Poor communities, in particular, those of rural Africans, bear the brunt of the past inequalities, and these continue to be reflected in the national results of the final-year examinations in Grade 12. Equity and access top the South African government's priority list for the country's education system, which accommodates approximately 12.3 million students (50.5% female). Access to education has improved to the extent that primary education is almost universal. However, only 86% of South African students are enrolled in secondary school, even though education in South Africa is compulsory and supposed to be free for Grades 1 to 9 (Grade 8 is the first grade of most secondary schools). Students are expected to pay fees only for Grades 10 to 12, but educational user fees are widespread across all the grades.

South Africa participated in TIMSS 1995, 1999, and 2003. Secondary analyses of these studies have revealed the large inequities in the education system, with 55% of the variance in students' mathematics scores explained by differences between schools (Howie, 2002). This variance, in turn, is mostly explained by the historical inequities imposed on communities and schools over the 40 years prior to 1994. The challenge is to explore the extent of the "gap" in students' scores by comparing the advantaged and disadvantaged communities in this context. The former includes well-resourced, largely urban schools; the latter includes largely under-resourced, mostly black, rural schools. Previous work conducted for mathematics pointed to crucial student-level factors such as language, socioeconomic status, perceptions of mathematics, attitudes toward mathematics, and self-concept (see Howie, 2002). In this paper, the focus is on performance in science and the predictors of this performance across three groups. The three groups, classified on the basis of their background characteristics, are advantaged, semi-advantaged, and disadvantaged.

The paper aims to ascertain the extent to which the previously mentioned factors (for mathematics) also have an effect on science and to what extent other factors play a role. Specifically, the following research questions are addressed:

1. To what extent does the location of the school have an effect on the performance of students in science?
2. How do students from advantaged and disadvantaged backgrounds compare in their science performance?
3. What other factors have an effect on the performance of South African students in science?

The paper is structured in the following manner. The next section provides an overview of previous research on the topic. Section 3 presents the conceptual framework for this research. Section 4 provides a brief description of the research design and methods. The main findings are presented in Section 5, and this is followed by conclusions and implications for further research.

Previous research conducted on factors related to science performance

The 21st century beckoned a new era with many possibilities. However, it was estimated that at the end of the 20th century about 140 million people in sub-Saharan Africa could not read or write. Amongst the 47.4 million people (South Africa.info, 2006) of South Africa's multicultural society, approximately 4.5 million adults over 20 years of age have received no education (Statistics South Africa, 2001), which can be attributed to the decades of apartheid policies implemented under the nationalist government of South Africa. These separatist policies forced cultural groups apart under the guise of separate development. The education system became divided, with children of each race group attending schools separated on the basis of these racial groupings. Schools for white children received more funding than others, had better facilities, were better equipped, and had better qualified teachers (Alexander, Badenhorst, & Gibbs, 2005). Therefore, today, in addition to the other challenges facing the rest of Africa and other developing worlds, South Africa has to deal with a set of special circumstances.

However, the complexity and peculiarities of schools in general can be magnified by highly disadvantaged settings such as those evident in South Africa and other developing countries. There is a need to explore and disentangle the multiple associations and divergent outcomes derived from the same set of input variables. For example, qualifications and experience of the teacher relate to the quality of learning taking place, and inexperienced and poorly qualified teachers tend to concentrate in certain geographical areas or schools serving students of particular socioeconomic backgrounds (Mabogoane, 2004).

Cross-national research at the student level indicates that many antecedent factors, such as students' home background and their age, religion and gender, in addition to the type of school they attend and the locality of that school, affect student achievement in science, particularly in Trinidad and Tobago (Kutnick & Jules, 1988). Lockheed and Zhao (1993) found, after holding constant students' socioeconomic status (SES), age, and gender, significant differences in achievement among students attending different types of schools in the Philippines. Young (1998) found that the location of schools in Western Australia had a significant effect on the performance of students attending them. Students from rural areas performed less well than their counterparts in urban areas. Research by van der Berg and Burger (2003) found the achievement of students from poor schools in the Western Cape (a province in South Africa) to be worse than that of students from other SES groups and population groups. The students attending the Western Cape schools were predominantly African and to a lesser degree colored. There appears to be a consensus within the literature of racial/ethnic differences in students' science performance (see also Bacharack, Baumeister, & Furr, 2003; Ikpa, 2003; Von Secker, 2004). The literature also shows a relationship between racial/ethnic origin and SES.

Studies comparing the achievement of students from different SES groups in the United States of America show that students from high SES groups tend to outperform students from lower SES groups (see, for example, Von Secker, 2004). The same picture emerges in South Africa, where students from lower SES groups obtain significantly lower scores than those from higher SES groups (National Department of Education, 2005). Here, SES is associated with possessions in the home and the expanded opportunities which the home environment provides to students (Von Secker, 2004; Yang, 2003). Parents' level of education, their occupations, and their aspirations for their children are also linked to SES (Tamir, 1989; Young & Fraser, 1990).

Many inconsistencies have been reported with regard to the effect of parental involvement on science achievement (see, for example, McNeal, 2001). As McNeal (2001) shows, parental involvement is more effective for higher SES students than it is for lower SES students, but the mother's perception of her child's ability does affect the self-efficacy of the child (Bleeker & Jacobs, 2004). Parents' reinforcement of educational expectations and family involvement in educational activities has also been linked to performance in science (Onocha & Okpala, 1987). Parental involvement in educational activities includes taking an interest in school work and helping with homework.

Students who regularly have homework tend to perform better in science in the upper grades of school (Van Voorhis, 2003). Opportunities to learn and motivation have also been linked to science performance (Tamir, 1989). In an analysis of the Longitudinal Study of American Youth, Young, Reynolds and Walberg (1996) found that attitudes toward science affect the performance of pupils in science. A similar result was obtained in a secondary analysis of the Cypriot TIMSS 1995 data conducted by Papanastasiou and Zembylas (2002). Finally, the literature suggests that gender influences science achievement, with boys generally performing better than girls in this subject (Bacharack et al., 2003; Dimitrov, 1999; Von Secker, 2004).

Conceptual framework

The conceptual framework for this study is based on one used by Howie (2002), who adapted it from a model developed by Shavelson, McDonnell, and Oakes (1987). The framework also draws on thinking in relation to the mathematics and science curriculum by the International Association for the Evaluation of Educational Achievement (IEA) (Travers & Westbury, 1989). The framework used in this study (and depicted in Figure 1) includes a number of adaptations to the original frameworks to better suit this secondary analysis of the TIMSS-R (1999) data.

The model shown in Figure 1 presents the education system in terms of inputs, processes, and outputs. The curricula for academic subjects play a central role in an education system. The IEA also considers curricula to be the key feature of any evaluation of educational achievement. This is reflected by their inclusion of curriculum-based, explanatory designs (IEA, 1998, p. 32). The organization differentiates between the intended, the implemented, and the attained curriculum. The central positioning of the three curricula and their links between and among elements within the model illustrate this key role. The model also provides an important theoretical and conceptual basis for analysis of the TIMSS-R (1999) data. Because the data were collected at a number of education levels, namely, school, classroom, and student, the model also serves as a means of exploring, identifying and/or confirming the reasons for differences in student achievement in science.

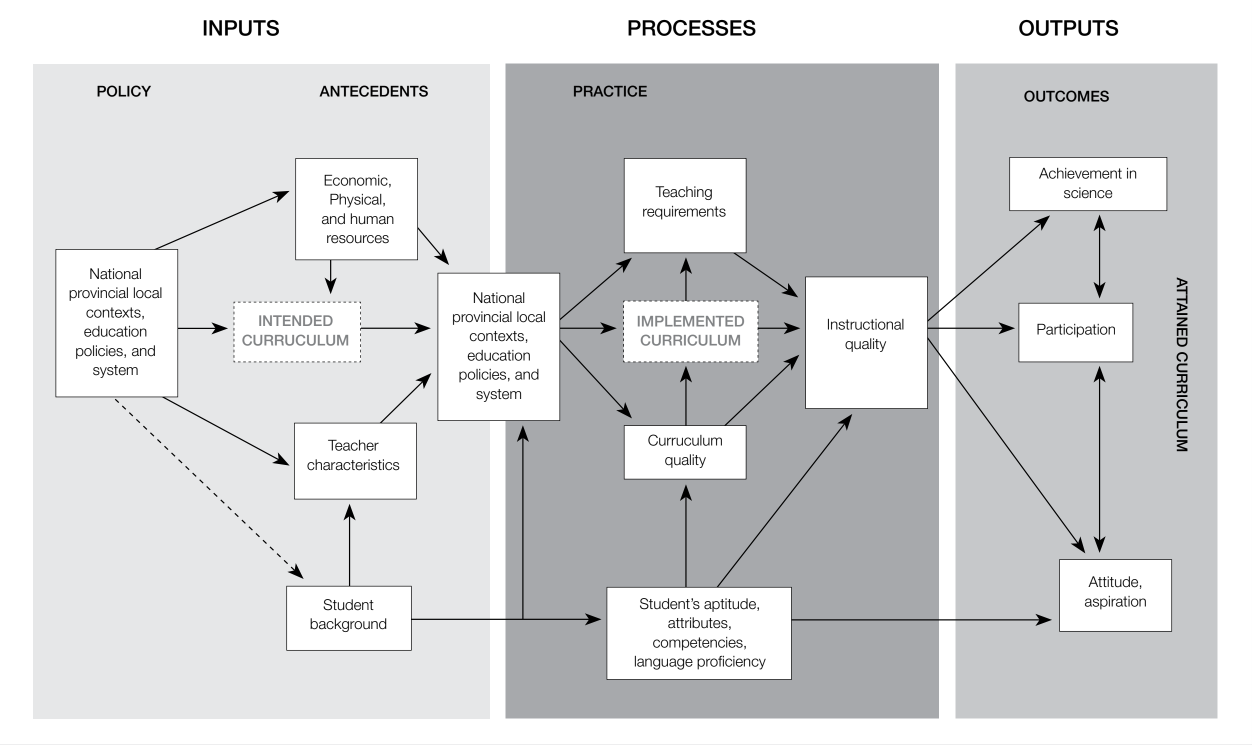
Research design and methods

This study is a secondary analysis of the South African data from the IEA's Third International Mathematics and Science Study-Repeat (TIMSS-99), collected in 1998 (as was the case in all Southern Hemisphere countries). The data were explored to determine if there were significant differences among groups of students classified in terms of the relative advantage of their background and in order to provide possible explanations for the students' performance. The analysis identified factors that had a direct or indirect effect on science achievement of South African students.

Sample

The TIMSS-R (1999) sample for South African students was used for the analysis. A national sample of schools, stratified according to province, type of education (government or private), and medium of instruction (English and Afrikaans). The data were obtained from 8,142 student achievement tests and questionnaires as well as from 189 school principals.

Figure 1: Factors Related to Science Achievement



Note: This model was used to explain mathematics achievement in Howie (2002), who adapted it from a framework developed by Shavelson, (1987).

Instruments

The instruments included the TIMSS-R (1999) science achievement test, the student questionnaire and the principal's questionnaire, although the only information considered for the analysis from the principal's questionnaire was school location.

Data analysis

Descriptive statistics were generated to provide descriptive results and to prepare the data for further analysis. Further preparation entailed preparing a correlation matrix in order to identify variables that were related, conducting reliability analysis to investigate the reliability of certain scales, and building constructs for inclusion in a Partial Least Squares (PLS) analysis. Finally, in order to prepare a number of models for the purpose of comparison, indices were constructed for advantaged, semi-advantaged, and disadvantaged learners based on the socioeconomic status of each student, the possessions in the student's home (including number of books), and the language(s) used in the student's home.

The PLS employed the software PLSPATH to explore and analyze those student-level factors (together with one variable from the school-level data, namely the location) that had an effect on students' achievement in science. For more details on the PLS, see Howie (2002).

Results

In this section, the main findings of the research are presented first in terms of the overall South African sample and thereafter for each of the groups classified in terms of their relative advantage in background.

Predictors of science achievement for South African students

The students performed well below the international average for the science test (as well as for mathematics). They achieved an average score of 249 out of 800 points. There was a considerable difference in the mean ages of the students between the participating countries and the South African sample. The South African students were also the oldest in TIMSS-R (1999), with an average age of 15.5 years. However, after the TIMSS data were cleaned to investigate specific factors, it was found that the average age of the sample analyzed in this paper was actually 16 years.

As mentioned earlier, the South African students also wrote an English language proficiency test. The overall average score was 17 out of 40. Those students whose first language was English (*n* = 533) or Afrikaans (*n* = 1,281) achieved the highest marks on this test (means of 25 (*SE* 0.4) and 21 (*SE* 0.2) respectively out of 40). Children speaking indigenous African languages (*n* =5,496) achieved very low scores, with an average of 15 marks out of 40 (*SE* 6.5) on the test. Overall, the scores on this test showed the students, as a group, had low proficiency in English.

The overall statistics for the univariate analysis conducted for this paper are set out in Table 1. Originally 22 factors were considered for inclusion in the student model. However, after trimming the model, only 18 factors (latent variables) remained, reflecting 19 manifest variables (see Appendix A for details). These factors were included in four models: the overall science model, the advantaged group model, the disadvantaged group model, and the semi-advantaged group model. However, in the final trimming, the location variable had no effect in any of the paths tested, and was therefore removed from the overall model. This was not the case in the other three models, and therefore that variable remained in those models. The overall model is discussed below; the other three models are discussed in the following sections. The same variables were included for all four models and with the same patterns in order to explore the effects of the same variables overall and within different groups on performance in science.

The inner model results for the South Africa sample are presented in Table 2. These results provided the direct effects for the overall model. The full inner model, including direct, indirect, correlation, and total effects, is given in Appendix A, as is the outer model.

Description of students in advantaged, semi-advantaged, and disadvantaged groups

The sample was divided into sub-samples according to the criteria specified below. These were the advantaged group (*n* = 225), the semi-advantaged group (*n* = 3,656), and the disadvantaged group (*n* = 4,151).

The majority of the South Africa sample was African (71%), and the ethnic/racial group with the lowest number of students in it was Indian (3%). None of the African students was in the advantaged category, and

Table 1: Univariate Statistics for Manifest Variables Included in the Student Model (N = 8,142)

Variable	Variable name	Mean	*SD*	Minimum	Maximum
Science score	bsssci01	248.816	132.237	5.00	775.400
English score	totscore	17.007	6.396	.00	40.00
Dictionary	diction	1.253	.434	1.00	1.00
Age	age_1	15.520	1.803	9.42	28.80
Extra lessons	lesson_1	1.212	1.098	.00	4.00
Aspirations	selfed_1	3.899	1.420	1.00	5.00
Books in home	books	2.019	1.158	1.00	5.00
Success attribution	luck	1.987	.990	1.00	4.00
Home language	homelang	1.314	.601	.00	3.00
Race of student	race_1	1.802	1.359	1.00	5.00
Language on radio	ralang_1	1.583	.855	1.00	3.00
Maths is boring	bores_1	2.860	1.040	1.00	4.00
Possessions in home	posses10	14.439	2.479	10.00	20.00
No. of parents at home	parent	1.079	.997	.00	2.00
Self-concept in science	difsci	11.034	2.876	4.00	16.00
Attitude to science	sciimp	10.217	1.977	3.00	12.00
Language of learning	lanlearn	5.104	1.447	2.00	7.00
Location of school	sccom_1	2.796	.898	1.00	4.00
Attendance at school	attend	2.456	.606	.00	6.00
Student activities	studact	1.779	2.789	.00	24.00

Table 2: The Inner Model that Provided the Direct Effects for the Overall Model

Latent Variable	Description of Variable	Direct	Total	Indirect	*R*-squared
ASPIRE	Aspirations for future				.023
AGE	Age of student	-.1507	-.1507	–	
LANLEARN	Language of learning				.244
LANG	Language spoken at home	.4942	.4942	–	
ATTEND	Attendance at school				.016
AGE	Age of student	.0820	.0945	.0125	
ASPIRE	Aspirations for future	-.0832	-.0832	–	
SUCATTRB	Attributions for success				.50
LANG	Language spoken at home	.1732	.1732	–	
SES	Socioeconomic status	–	-.0845	-.0845	–
ATTITUDE	Attitude to science				.024
AGE	Age of student	-.1302	-.1370	-.0068	
ASPIRE	Aspirations for future	–	*.0060*	*.0060*	
ATTEND	Attendance at school	*-.0723*	*-.0723*	–	
SELFCNPT	Self-concept in science				.187
AGE	Age of student	–	*.0363*	.0363	
ASPIRE	Aspirations for future	–	*-.0016*	-.0016	
LANG	Language spoken at home	*.0553*	*-.0553*	–	
SES	Socioeconomic status	–	*.0270*	.0270	
ATTEND	Attendance at school	–	*.0192*	.0192	
SUCATTRB	Attribution for success	-.3191	-.3191	–	
ATTITUDE	Attitude to science	-.2652	-.2652	–	
BOOK	Books in the home				.109
SES	Socioeconomic status	-.3295	-.3295	–	
ENGTEST	English test score				.379
AGE	Age of student	-.1637	-.1781	-.0144	
ASPIRE	Aspirations for future	.0959	.0959	–	
LANG	Language spoken at home	.3880	.3880	–	
SES	Socioeconomic status	-.1915	-.2164	-.0249	
HOME	Number of parents at home	.0206	.0206	–	
BOOK	Number of books in home	*.0757*	*.0757*	–	
SCISCR	Science score				.542
AGE	Age of student	–	*-.0772*	*-.0772*	
ASPIRE	Aspirations for future	–	*.0399*	*.0399*	
LANG	Language spoken at home	.2122	.4336	.2214	
SES	Socioeconomic status	-.1438	-.2360	-.0922	
HOME	Number of parents at home	–	*.0086*	*.0086*	
SCIIMPT	Science is important	*.0271*	*.0271*	–	
LANLEARN	Language of learning	.1122	.1122	–	
ATTEND	Attendance at school	–	*-.0017*	*-.0017*	
SUCATTRB	Attribution for success	–	*.0282*	*.0282*	
ATTITUDE	Attitude to science	–	*.0235*	*.0235*	
SELFCNPT	Self-concept in science	-.0884	-.0884	–	
BOOK	Number of books in home	–	*.0314*	*.0314*	
ENGTEST	English test score	.4152	.4152	–	

Notes: Beta coefficients equal to and above .08 are regarded as significant relationships.
Beta coefficients below .08 are in italics.

just over half of this group was colored (of mixed race). The majority of the African students was classified as disadvantaged (86%), a finding consistent with the impact of South Africa's political history on this group (refer to Table 3).

Students in the advantaged group all spoke the language of the test in contrast to only 9% of the students from the disadvantaged group who spoke the language of the test (Table 4). More than 20% of students in the semi-advantaged and disadvantaged groups never spoke the language of the test.

Predicting performance of students in science within groups

Overall, as Table 5 shows, 40% of this sample achieved scores of 200 or below (out of 800 points), which is very low. Of the disadvantaged group, 44% attained this score in contrast to 8% of the advantaged group. Sixty percent of the advantaged group attained scores between 401 and 500 out of 800, while the other groups performed substantially below this level (only 20% of the semi-advantaged and 6% of the disadvantaged groups attained scores higher than 400 points).

Predicting the performance of advantaged students

The sub-sample of advantaged students was very small and comprised only 225 students (see Appendix B). The performance of the advantaged students was significantly above that of both the other groups and of the overall performance (422 points (*SD* 144) compared to 248 points). The performance, however, of the students in the advantaged group on the English proficiency test was below expectation (23 out of 40 marks). This group, at 14.9 years, was also younger than the average for the South Africa sample (15.5 years). Students in this group had many books at home and tended to speak English or Afrikaans at home, and they were also more likely to be Indian or white. By definition, they had most of the possessions listed for the index of SES.

The outer model results of the PLS model revealed that these results fell within the recommended parameters set for the loading, communality, and tolerance. A number of relationships for the loading and communality were significant, providing confirmation of the strength of the model. To evaluate the strength of the inner model (see Table 6), 0.8 was regarded as the minimum acceptable coefficient to reveal a relationship. Any coefficient found below this was not regarded as having a relationship (see Howie, 2002).

Overall, 66% of the variance of the science score could be explained by 14 factors. Of these factors, three had a direct effect on the science score. These factors were SES (-.35), self-concept in science (-.11), and (the most dominant factor) the English test score (.49). Several factors had an indirect effect, namely age (.10), language spoken at home (.14), and books in the home (-.09). SES, in addition to the direct effect, also had an indirect effect (-.17). Factors that had no

Table 3: Distribution of Students in Advantaged, Semi-advantaged, and Disadvantaged Groups by Racial Grouping

Categories	African	Asian	Colored	Indian	White
Advantaged	0%	0%	53%	4%	44%
Semi-advantaged	59%	1%	24%	4%	13%
Disadvantaged	86%	2%	5%	1%	6%
Overall percentage	71%	1%	15%	3%	10%

Note: The figures have been rounded off, and in some cases may exceed 100%.

Table 4: Percentage of Students and the Frequency of Students Speaking the Language of the Test at Home

	Language of the test spoken at home (%)		
	Never	Sometimes	Always or almost always
Advantaged	0	0	100
Semi-advantaged	23	38	39
Disadvantaged	22	68	9
Overall percentage	**22**	**52**	**26**

Table 5: Performance of Students by Category

Categories	Achievement in science within a group (%)					
	0–200	201–300	301–400	401–500	501–600	601–800
Advantaged	8	13	19	27	25	8
Semi-advantaged	38	29	14	11	7	2
Disadvantaged	44	35	16	5	1	0
Overall percentage	40	31	15	8	5	1

effect, although tested, were "if science is important," language of learning, and location of the school.

The strength of the English test score far surpassed that of any other predictor, and therefore was examined more closely than were the other predictors. Forty-eight percent of the variance of the English test score was explained by seven factors, of which five had direct effects. These were age, language spoken at home, SES, aspirations of students, and books in the home. The strongest predictor was SES. The only indirect effect observed was that of SES. The number of parents in the home and the location of the school had no effect on the English test score.

Predicting the performance of semi-advantaged students

The semi-advantaged students (*n* = 3,736) achieved a score on the science test (264 points) that was above the national average (248 points). The score achieved for the English test was 17 marks (out of 40). Many of these students spoke an indigenous African language at home (by definition of the classification). The semi-advantaged students on average had more listed possessions than those in the disadvantaged group. They were also more likely to live in towns and cities.

The results of the PLS outer model met the criteria for all the parameters. The results are presented in Table 7. The model explained 63% of the variance in the science score. Fourteen factors were included in the model, of which seven were tested for direct effects. Only four proved to have a direct effect (language

Table 6: Inner Model for Advantaged Group of Students

Variable	Direct	Total	Indirect	*R*-squared
LANLEARN				.023
LANG	.1527	.1527	–	
ASPIRE				.022
AGE	-.1392	-.1392	–	
LOCATION	.0401	.0401	–	
SUCATTRB				.101
LANG	.1069	.1069	–	
SES	-.2331	-.2331	–	
LOCATION	.0437	.0437	–	
ATTEND				.078
AGE	.2307	.2485	.0177	
LOCATION	–	-.0051	-.0051	
ASPIRE	-.1274	-.1274	–	
ATTITUDE				.010
AGE	-.1001	-.0995	.0005	
LOCATION	–	.0000	.0000	
ASPIRE	–	-.0003	-.0003	
ATTEND	.0021	.0021	–	

Table 6 (contd.): Inner Model for Advantaged Group of Students

Variable	Direct	Total	Indirect	*R*-squared
SELFCNPT				.208
AGE	–	.0246	.0246	
LANG	–	-.0373	-.0373	
SES	–	.0814	.0814	
LOCATION	–	-.0153	-.0153	
ASPIRE	–	.0001	.0001	
SUCATTRB	-.3492	-.3492	–	
ATTEND	–	-.0005	-.0005	
ATTITUDE	-.2474		-.2474	
BOOK				.842
SES	-.9192	-.9192	–	
LOCATION	-.0055	-.0055	–	
ENGTEST				.481
AGE	-.1703	-.1910	-.0207	
LANG	.2700	.2700	–	
SES	-.4915	-.3282	.1633	
HOME	.0515	.0515	–	
LOCATION	.0672	.0742	.0069	
ASPIRE	.1488	.1488	–	
BOOK	-.1777	-.1777	–	
SCISCR				.655
AGE		–		
LANG	-.0245	.1200	–	
SES		-.3465	–	
HOME	–	.0254	–	
SCIIMPT	-.0065	-.0065	–	
LANLEARN	.0489	.0489	–	
LOCATION	.0423	.0805	–	
ASPIRE	–	.0733	–	
SUCATTRB	–	.0372	–	
ATTEND	–	.0001	–	
ATTITUDE	–	.0263	–	
SELFCNPT	-.1065	-.1065	–	
BOOK	–	-.0876	–	
ENGTEST	.4929	.4929	–	

spoken at home, SES (greater than .20), language of learning (.08) and (again the strongest effect) the English test score (.37). Language spoken at home and SES also had an indirect effect on the science score. No other factors had an indirect effect.

A closer look at the English test score revealed that 48% of the variance in the students' English test score could be explained by seven factors in this model. Five factors had a direct effect (age, language spoken at home, SES, location of the school, and aspirations of students), with the strongest being SES (-.22). No indirect effects were found.

Predicting the performance of disadvantaged students

The disadvantaged students (*n* = 4,151) achieved the lowest scores for science (224.5 out of 800) (see Appendix B). This was also the case for the English test score (16 out of 40). These children tended to be older (15.6) than the mean age for the South Africa sample (15.5). By definition (of the classification), their home language differed from the medium of instruction and therefore was most likely to be one of the African languages. Children in this group were also more likely to come from rural areas than were the students in the other groups. They had fewer books in the home and had fewer listed possessions in the home than had the children in the other groups.

As with the advantaged group, the PLS outer model results for the disadvantaged students proved to be similar and conformed to the set parameters. Table 8 presents the results of the PLS inner model.

The model for the disadvantaged students explained only 35% of the variance in the science score—much less than the variance explained in the other two groups. Of the 14 factors included in the model, seven had a direct effect (language spoken at home, SES, science is important, language of learning in the classroom, location of the school, self-concept in science, and the

Table 7: PLS Inner Model for Semi-advantaged Students

Variable	Direct	Total	Indirect	Fit
LANLEARN				.371
LANG	.6090	.6090	–	
ASPIRE				.021
AGE	-.1342	-.1342	–	
LOCATION	.0383	.0383	–	
SUCATTRB				.081
LANG	.1299	.1299	–	
SES	-.1615	-.1615	–	
LOCATION	.0243	.0243	–	
ATTEND				.015
AGE	.0840	.0944	.0104	
LOCATION	–	-.0030	-.0030	
ASPIRE	-.0776	-.0776	–	
ATTITUDE				.020
AGE	-.1200	-.1262	-.0061	
LOCATION	–	.0002	.0002	
ASPIRE	–	.0051	.0051	
ATTEND	-.0651	-.0651	–	
SELFCNPT				.200
AGE	–	.0348	.0348	
LANG	–	-.0423	-.0423	
SES	–	.0525	.0525	
LOCATION	–	-.0080	-.0080	
ASPIRE	–	-.0014	-.0014	
SUCATTRB	-.3251	-.3251	–	
ATTEND	–	.0179	.0179	
ATTITUDE	-.2755	-.2755	–	
BOOK				.169
SES	-.4357	-.4357	–	
LOCATION	-.0483	-.0483	–	
ENGTEST				.483
AGE	-.1237	-.1364	-.0127	
LANG	.3871	.3871	–	
SES	-.2167	-.2210	-.0043	
HOME	.0076	.0076	–	
LOCATION	.0965	.0997	.0032	
ASPIRE	.0949	.0949	–	
BOOK	.0098	0098	–	

Table 7 (contd.): PLS Inner Model for Semi-advantaged Students

Variable	Direct	Total	Indirect	Fit
SCISCR				.626
AGE	–	-.0535	-.0535	
LANG	.2031	.4001	.1969	
SES	-.2179	-.3043	-.0864	
HOME	–	.0029	.0029	
SCIIMPT	.0002	.0002	–	
LANLEARN	.0804	.0804	–	
LOCATION	.0263	.0642	.0379	
ASPIRE	–	.0357	.0357	
SUCATTRB	–	.0222	.0222	
ATTEND	–	-.0012	-.0012	
ATTITUDE	–	.0188	.0188	
SELFCNPT	-.0683	-.0683	–	
BOOK	–	.0037	.0037	
ENGTEST	.3748	.3748	-	

English test score). Only one of these, the English test score, had a strong effect (.40). Age, language spoken at home, and location of the school had weak indirect effects on the science score.

It is worthwhile to examine the paths explaining the variance in the English test score, given the strength of its effect on the students' science achievement. Twenty-three percent of the variance was explained by the above seven factors, of which five had direct effects on the English score (age, language spoken at home, location of the school, aspirations of students, and books in the home). SES and number of parents in the home had no effect on the English score, and no indirect effects were found for these factors.

Conclusions

The results from the mathematics-related data from TIMSS-R (1999) analyzed previously for South Africa (Howie, 2002) showed a strong relationship with location, English language proficiency, language of the home, SES, self-concept of the student (in terms of having difficulty with mathematics), and the importance that mathematics held for the student (according to mother, friends, and the student). The analysis in this paper revealed a similar pattern for science.

The performance of South African students in science, as in mathematics, was very low. The pattern evident suggests a strong relationship between levels of advantage and performance in terms of the more advantaged the students, the better they tended to perform both in science and on the English language proficiency test (see Table 9).

In terms of the predictors of science, the models that were applied uniformly across the groups revealed interesting differences among the groups (see Table 10). The models were able to explain high levels of variance for each model (above 50%), except for the disadvantaged group. In the case of the latter, it is clear that the factors that were included are too limited for this group, although there could also be a bottom effect due to the extremely low scores attained by this group. Further exploration of this group is needed to fully understand which factors predict achievement in conditions of most disadvantage. However, one factor remains consistent across all groups and that is the English test score. It has been suggested previously that the strength of this factor could be a result of measuring intelligence or aptitude rather than language ability (Howie, 2002), although disentangling these factors could be rather difficult.

Given the language policy of and challenges in South Africa, the strength of the relationships found is not surprising. If children's access to science knowledge is denied through inadequate communication and comprehension skills, then poor conceptual understanding is inevitable and has disastrous consequences. However, this finding may in part explain the very low performance of South African students in science, particularly those students from disadvantaged backgrounds (as defined in this paper). The finding may also suggest that part of the

Table 8: PLS Inner Model for Disadvantaged Students

Variable	Direct	Total	Indirect	*R*-squared
LANLEARN				.055
LANG	.2341	.2341	–	
ASPIRE				.028
AGE	-.1542	-.1542	–	
LOCATION	.0480	0480	–	
SUCATTRB				0.12
LANG	.0548	.0548	–	
SES	-.0271	-.0271	–	
LOCATION	.0724	.0724	–	
ATTEND				.014
AGE	.0710	.0839	.0128	
LOCATION	–	-.0040	-.0040	
ASPIRE	-.0832	-.0832	–	
ATTITUDE				.029
AGE	-.1431	-.1501	-.0070	
LOCATION	–	.0003	.0003	
ASPIRE	–	.0069	.0069	
ATTEND	-.0832	-.0832	–	
SELFCNPT				.164
AGE	–	.0393	.0393	
LANG	–	-.0160	-.0160	
SES	–	.0079	.0079	
LOCATION	–	-.0212	-.0212	
ASPIRE	–	-.0018	-.0018	
SUCATTRB	-.2914	-.2914	–	
ATTEND	–	.0218	.0218	
ATTITUDE	-.2617	-.2617	–	
BOOK				.161
SES	-.3881	-.3881	–	
LOCATION	.0366	.0366	–	
ENGTEST				.228
AGE	-.1955	-.2113	-.0159	
LANG	.2061	.2061	–	
SES	-.0664	-.1060	-.0396	
HOME	.0158	.0158	–	
LOCATION	.1900	.1987	.0087	
ASPIRE	.1028	.1028	–	
BOOK	.1020	.1020	–	

Table 8 (contd.): PLS Inner Model for Disadvantaged Students

Variable	Direct	Total	Indirect	*R*-squared
SCISCR				.346
AGE	–	-.0889	-.0889	
LANG	.0942	.1996	.1055	
SES	-.0969	-.1402	-.0433	
HOME	–	.0063	.0063	
SCIIMPT	.0790	.0790	–	
LANLEARN	.0905	.0905	–	
LOCATION	.0898	.1717	.0819	
ASPIRE	–	.0414	.0414	
SUCATTRB	–	.0317	.0317	
ATTEND	–	-.0024	-.0024	
ATTITUDE	–	.0285	.0285	
SELFCNPT	-.1088	-.1088	–	
BOOK	–	.0408	.0408	
ENGTEST	.4005	.4005	–	

Table 9: Performance in Science and English Overall and Within Groups

	Advantaged		Semi-advantaged		Disadvantaged		Overall	
	Score	*SD*	Score	*SD*	Score	*SD*	Score	*SD*
Science score out of 800	422	144	264	145	224.5	107	249	132
English score out of 40	23.5	7.8	17	6.9	16	5.5	17	6.4

Table 10: Predictors of Science Score Overall and Within Groups

	Advantaged	Semi-advantaged	Disadvantaged	Overall
R-squared	.66	.63	.35	.54
Language spoken at home	X	.20	.09	.21
SES	-.35	-.22	-.10	-.14
Science is important	X	X	.09	X
Language of learning	X	.08	.09	.11
Location of school	X	X	.09	X
Self concept in science	X	X	-.11	.09
English test	.49	.37	.40	.42

Note: X indicates that no relationship was found.

solution in developing the sound knowledge and skills base of students in South Africa lies in interventions related to language at both the student level and the teacher level.

Final words

Few nationally representative studies (Tshabalala, 2006, is one of them) have been conducted in South Africa that have the aim of analyzing differences in rural–urban performance. The identification of predictors of science achievement is critical given the paucity of nationally generalizable data in South Africa. It is crucial to identify factors beyond the obvious inheritance of the apartheid era that could really contribute to the advancement of science teaching and learning, and that may also be invaluable to those involved in teacher education and development in the country.

Appendix A

Table A1: Latent and Manifest Variables Included in the Student-level PLS Analysis

Latent variables	Manifest variables	TIMSS-R variables	Description	Scoring
SCISCR	SCIEN	BSSCIT01	Student mean score on TIMSS-R science test	Score out of 800 points
ENGTEST	TOTSCORE	N/a	Student mean score on English language proficiency test	Score out of 40 points
RACE	RACE_1	POPULAT	Race of student: African, Colored, Indian, White,	1. African 2. Asian 3. Colored 4. Indian 5. White
AGE	AGE_1	BSGAGE	Age of student	Number of years
LANG	RALANG_1	RADIO	Language on favorite radio station	1. All other languages 2. Afrikaans 3. English
	HOMELANG	INGUA	Language on favorite radio station	0. Other languages 1. African languages 2. Afrikaans 3. English
HOME	PARENT (COMPOSITE)	GENADU 1,2, GENADU 5,6	Whether students have two parents	0 = no 1 = yes
BOOK	BOOKS	BSBGBOOK	Number of books	1. 0–10 2. 11–25 3. 26–100 4. 101-200 5. more than 200
SES	POSSES10 (COMPOSITE)	BSBGPS02,5-14 BSBGPS02,5-14	Computer, electricity, tap water, TV, CD player, radio, own bedroom, flush toilets, car (9 items)	0. no 1. yes
LANLEARN	LANLEARN (COMPOSITE)		Extent to which both student and teacher speak language of instruction in science class at home if not English/Afrikaans	1. Language spoken 2. Sometimes English/ Afrikaans 3. Most of the time English/Afrikaans 4. Always English/ Afrikaans
SCIIMPT	SCIIMPT (COMPOSITE)	BSBGMIP2 BSBGSIP3 BSBGFIP2	Extent to which students' respective mothers and friends think that science is important (3 items)	Scale of (+) 1_4 (-) strongly agree to strongly disagree
ASPIRE	SELFED_1	GENEDSE	Aspirations to education	1. Some secondary 2. Finished secondary 3. Finished technikon 4. Some university 5. Finished university

Table A1 (contd.): Latent and Manifest Variables Included in the Student-level PLS Analysis

Latent variables	Manifest variables	TIMSS-r variables	Description	Scoring
ATTEND	ATTEND (COMPOSITE)	BSBGSSKP BSBGFSKP	Extent to which student or students' friends bunk school (2 items)	Scale of (+) 0-4 (-) Never, once or twice, three or four times, five or more
SELFCNPT	DIFSCI	SCIMYT_1-5	The extent to which student reports having difficulty with science (5 items)	Scale of (-) 1–4 (+) Strongly agree to strongly disagree
SUCATTRB	LUCK	MATHDOW2	If the student attributes success to luck	Scale of (-)1–4 (+) Strongly agree to strongly disagree
ATTITUDE	BORES	SCIBORE	If student finds science boring	Scale of (-)1–4 (+) Strongly agree to strongly disagree

Appendix B Univariate Statistics of Factors from PLS Models per Groups

Table A2: Univariate Statistics for Overall Model (n = 8,142)

Variable	Mean	*SD*	Minimum	Maximum
bsssci01	248.816	132.237	5.000	775.400
totscore	17.007	6.396	.000	40.000
diction	.756	.430	.000	1.000
age_1	15.520	1.803	9.420	28.800
lesson_1	1.212	1.098	.000	4.000
selfed_1	3.899	1.420	1.000	5.000
books	2.019	1.158	1.000	5.000
luck	1.987	.990	1.000	4.000
homelang	1.314	.601	.000	3.000
race_1	1.802	1.359	1.000	5.000
ralang_1	1.583	.855	1.000	3.000
bores_1	2.860	1.040	1.000	4.000
posses10	14.439	2.479	10.000	20.000
parent	1.079	.997	.000	2.000
difsci	11.034	2.876	4.000	16.000
sciimp	10.217	1.977	3.000	12.000
lanlearn	5.104	1.447	2.000	7.000
attend	1.446	1.492	.000	6.000
studact	15.704	5.158	.000	24.000

Table A3: Descriptive Statistics for Advantaged Students (n = 225)

Variable	Mean	*SD*	Minimum	Maximum
bsssci01	421.809	144.181	5.000	775.400
totscore	23.545	7.859	4.000	40.000
diction	1.110	.313	1.000	2.000
age_1	14.919	1.057	13.400	20.800
lesson_1	.518	.867	.000	4.000
selfed_1	4.208	1.265	1.000	5.000
books	3.780	1.604	1.000	5.000
luck	2.612	1.000	4.000	1.007
homelang	2.169	.374	2.000	3.000
race_1	3.910	.976	3.000	5.000
ralang_1	2.573	.555	1.000	3.000
bores_1	3.102	.857	4.000	1.000
posses10	13.125	3.159	10.000	19.000
parent	1.522	.853	.000	2.000
difsci	9.427	2.977	4.000	16.000
sciimp	1.745	3.000	12.000	10.086
lanlearn	6.635	.906	4.000	7.000
sccomm_1	3.224	.845	2.000	4.000
attend	1.518	1.497	.000	6.000
studact	14.078	5.119	1.000	24.000

Table A4: Univariate Statistics for Semi-advantaged Students (n = 3,656)

Variable	Mean	*SD*	Minimum	Maximum
bsssci01	264.020	145.318	5.000	737.500
totscore	17.496	6.917	.000	40.000
diction	1.301	.459	1.000	2.000
age_1	15.442	1.754	9.500	26.600
lesson_1	1.105	1.098	.000	4.000
selfed_1	3.900	1.409	1.000	5.000
books	2.270	1.297	1.000	5.000
luck	2.034	1.008	1.000	4.000
homelang	1.451	.653	.000	3.000
race_1	2.105	1.456	1.000	5.000
ralang_1	1.684	.872	1.000	3.000
bores_1	2.830	1.017	1.000	4.000
posses10	15.309	2.885	10.000	20.000
parent	1.111	.994	.000	2.000
difsci	10.940	2.912	4.000	16.000
sciimp	10.211	1.902	3.000	12.000
lanlearn	5.284	1.531	2.000	7.000
sccomm_1	2.673	.904	1.000	4.000
attend	1.428	1.496	.000	6.000
studact	15.354	5.079	.000	24.000

Table A5: Univariate Statistics for Disadvantaged Students (n = 4,151)

Variable	Mean	*SD*	Minimum	Maximum
bsssci01	224.506	106.666	5.000	675.600
totscore	16.166	5.463	.000	38.000
diction	1.218	.413	1.000	2.000
age_1	15.628	1.867	9.420	28.800
lesson_1	1.322	1.075	.000	4.000
selfed_1	3.833	1.445	1.000	5.000
books	1.664	.758	1.000	4.000
luck	1.874	.947	1.000	4.000
homelang	1.138	.473	.000	3.000
race_1	1.400	1.067	1.000	5.000
ralang_1	1.432	.800	1.000	3.000
bores_1	2.850	1.065	1.000	4.000
posses10	13.736	1.635	10.000	16.000
parent	1.023	1.000	.000	2.000
difsci	11.219	2.802	4.000	16.000
sciimp	10.230	2.054	3.000	12.000
lanlearn	4.821	1.304	2.000	7.000
sccomm_1	2.892	.894	1.000	4.000
attend	1.426	1.482	.000	6.000
studact	16.120	5.189	.000	24.000

References

Alexander, R., Badenhorst, E., & Gibbs, T. (2005). Intervention programme: A supported learning programme of disadvantaged students. *Medical Teacher, 27*(1), 66–70.

Bacharack, V. R., Baumeister, A. A., & Furr, R. M. (2003). Racial and gender science achievement gaps in secondary education. *The Journal of Genetic Psychology, 164*(1), 115–126.

Bleeker, M. M., & Jacobs, J. E. (2004). Achievement in math and science: Do mother's beliefs matter 12 years later? *Journal of Educational Psychology, 96*(1), 97–109.

Dimitrov, D. M. (1999). Gender differences in science achievement: Differential effects of ability, response formats and strands of learning outcomes. *School Science and Mathematics, 99*(8), 445–450.

Howie, S. J. (2002). *English language proficiency and contextual factors influencing mathematics achievement of secondary school pupils in South Africa.* Unpublished doctoral thesis, University of Twente, The Netherlands.

Ikpa, V. W. (2003). The mathematics and the science gap between resegregated and desegregated schools. *Education, 124*(2), 223–229.

International Association for the Evaluation of Educational Achievement (1998). *The IEA guidebook: Activities, institutions and people.* Amsterdam: Author.

Kutnick, P., & Jules, V. (1988). Antecedents affecting science achievement scores in classrooms in Trinidad and Tobago. *International Journal of Educational Development, 8*(4), 305–314.

Lockheed, M. E., & Zhao, Q. (1993). *International Journal of Educational Development, 13*(1), 45–62.

Mabogoane, T. (2004). The challenge of distributing knowledge. *EduSource, 44*, 1–7.

McNeal, R. B. (2001). Differential effects of parental involvement on cognitive and behavioral outcomes by socioeconomic status. *Journal of Socio-Economics, 30*, 171–179.

National Department of Education. (2005). *Grade 6 intermediate phase systemic evaluation report.* Pretoria: Author.

Onocha, C., & Okpala, P. (1987). Family and school environmental correlates of integrated science achievement. *Journal of Psychology, 121*(3), 281–287.

Papanastasiou, E. C., & Zembylas, M. (2002). The effect of attitudes on science achievement: A study conducted among high school pupils in Cyprus. *International Review of Education, 48*(6), 469–484.

Shavelson, R. J., McDonnell, L. M., & Oakes, J. (1987). *Indicators for monitoring mathematics and science education: A sourcebook.* Santa Monica, CA: The RAND Corporation.

South Africa.info. (2006, December). *South Africa: Fast facts.* Retrieved 14 February, 2007, from http://www.southafrica.info/ess_info/sa_glance/facts.htm

Statistics South Africa. (2001). *Census data.* Pretoria: StatsSA.

Tamir, P. (1989). Home and school effects on science achievement of high school students in Israel. *Journal of Educational Research, 8*(1), 30–39.

Travers, K. J., & Westbury, I. (Eds.). (1989). *The IEA study of mathematics I: Analysis of mathematics curricula*. Oxford: Pergamon Press.

Tshabala, P. (2007). *Grade 3 numeracy achievement: A comparative analysis of rural and urban school learners in South Africa.* Unpublished Master's dissertation, University of Pretoria.

van der Berg, S., & Burger, R. (2003). Education and socio-economic differentials: A study of school performance in the Western Cape. *The South African Journal of Economics, 71*(3), 496–522.

Van Voorhis, F. L. (2003). Interactive homework in middle school: Effects on family involvement and science achievement. *Journal of Educational Research, 96*(6), 323–338.

Von Secker, C. (2004). Science achievement in social contexts: Analysis from national assessment of educational progress. *The Journal of Educational Research, 29*(2), 67–78.

Yang, Y. (2003). Dimensions of socio-economic status and their relationship to mathematics and science achievement at the individual and collective levels. *Scandinavian Journal of Educational Research, 47*(1), 21–42.

Young, D. J. (1998). Rural and urban differences in student achievement in science and mathematics: A multilevel analysis. *School Effectiveness and School Improvement, 9*(4), 386–418.

Young, D. J., & Fraser B. J. (1990). Science achievement of girls in single-sex and co-educational schools. *Research in Science and Technological Education, 8*(1), 5–21.

Young, D. J., Reynolds, A. J., & Walberg, H. J. (1996). Science achievement and educational productivity: A hierarchical linear model. *The Journal of Educational Research, 86*(5), 272–278.

Gender differences in mathematics learning in Malaysia

Halimah Awang and Noor Azina Ismail
University of Malaya, Malaysia

Abstract

This paper is a secondary analysis of TIMSS conducted in 2003 by the International Association for the Evaluation of Educational Achievement (IEA). This is the second participation from Malaysia since TIMSS 1999 involving students in the eighth grade. The paper examines gender differentials in terms of the overall mathematics average achievement as well as the main content areas of mathematics topics and skills, namely fraction and number sense; data representation, analysis, and probability; and geometry and algebra. It also examines several factors that could affect mathematics achievement, including students' self-confidence in learning mathematics, their attitudes toward mathematics, the value they place on mathematics, and the time they spend doing mathematics homework.

Introduction

The importance of having a strong foundation in mathematics as a prerequisite for admission into institutions of higher learning in most disciplines is well recognized. In Malaysia, the medium of teaching mathematics and science subjects in secondary schools was changed from the national language, Malay, to English, in 2002. The change was made because of the need for students to grasp scientific and mathematical understanding and learning in the universal language so that they can compete academically and vocationally in this borderless world on leaving school. These two subjects (mathematics in particular, as a subject taught in every tuition centre across all levels of schooling, and outside of the school system) receive considerable attention from teachers and parents.

Achievement in mathematics varies across nations, regions, and a variety of socioeconomic and demographic characteristics. One of the factors most discussed by educators and researchers is the role of gender in mathematics learning. For example, the literature shows that females generally score lower than males on standardized tests of mathematics achievement (Cleary, 1992; Gallagher & Kaufman, 2005), and that more males than females score at the two extreme ends of the range of scores (Wang & Maxey, 1995; Willingham & Cole, 1997). In their analysis of data from the second International Assessment of Educational Progress in Mathematics and Science, Beller and Gafni (1996) found that, across countries, boys outperformed girls in mathematics performance. (The sample of students included 9-year-olds and 13-year-olds.) Engelhard (1990) similarly found that among 13-year-old students, boys tended to perform better than girls, and increasingly so as the level of complexity in mathematics content increased. However, a study by Alkhateeb (2001) found that among high school students in the United Arab Emirates, girls scored higher than boys on tests of mathematics achievement. The purpose of this present paper is to examine the differences in mathematics achievement of Grade 8 male and Grade 8 female students in Malaysian schools who participated in IEA's Third International Mathematics and Science Study (TIMSS) 2003.

Method

TIMSS 2003 was designed to provide data on trends in Grade 8 mathematics and science achievement in an international context involving 46 countries, including Malaysia. This present study uses mathematics achievement data from Malaysia, where all testing for the study was carried out at the end of the 2002 school year. In particular, this paper examines gender differences in terms of the average overall mathematics scores as well as average achievement in each of the five content areas of mathematics. These are fractions and number sense; measurement; data representation, analysis, and probability; geometry; and algebra. The paper also examines several factors that could affect differences in mathematics achievement between girls and boys. The ones examined here are students' confidence in their mathematics ability, the value students place on mathematics, and the amount of time they spend doing mathematics homework

Chi-square tests and *t*-tests were used to examine the associations between variables and to assess the differences in the average achievement between the two genders, respectively. The analyses are based on data collected from 5,314 students, of whom 3,071 were girls and 2,243 were boys.

Results

The overall average mathematics score of 508 for the Malaysian students on TIMSS 2003 placed Malaysia at 10th place on the international ranking and set it as one of the 26 countries with an average achievement significantly higher than the international average of 467 from the 46 participating countries. This average was slightly lower than the international average score of 519 obtained in TIMSS 1999 (Mullis, Martin, Gonzalez, & Chrostowski, 2004). The 2003 international average achievement scores of 512 for girls and 504 for boys were lower than the averages for 1999 (521 and 517, respectively), although both differences were not statistically significant.

As Table 1 shows, the Malaysian students' average achievement score in mathematics was significantly higher than the international average in all five content areas. When we consider performance by gender, however, we can see that the average achievement of girls was significantly higher than the international average in only two mathematics content areas—number and algebra. Table 1 also indicates that, in relation to the five mathematics content areas, Malaysian students' highest achievement was in the number and content area, and that this pattern was the same for the girls and the boys.

Table 2 shows that the overall average achievement for Malaysian girls was significantly higher than the average achievement for Malaysian boys. Also, although the average achievement for girls was greater than that for the boys in all five content areas, the difference was significant in three of them, namely, algebra, data, and number.

We also looked at gender differentials in terms of students' self-confidence in learning mathematics. The results of the chi-square test, based on four statements the students made about their mathematics ability, is shown in Table 3, which shows significant differences between boys and girls in terms of their level of agreement and disagreement with the belief statements. More boys than girls "agreed a lot" that they usually did well in mathematics and that they learned things quickly in mathematics. However, boys also tended to agree that mathematics was not one of their strengths and that they found the subject more difficult than did many of their classmates.

To obtain the overall confidence measure of mathematics ability among students, TIMSS used three levels of index—high, medium, and low. The high index was assigned to students who, on average, "agreed a little" or "agreed a lot" with the four statements; the low index was assigned to students who "disagreed a little" or "disagreed a lot" with all four, on average. All other combinations of responses were assigned the medium index (Mullis et al., 2004). Note that the computation of index is reversed for the two negative statements in Table 3.

For Malaysia, the students' reports of their mathematics ability placed 39% of the students within the high index, 45% within the medium, and 16% within the low index. Table 4 shows that the average achievement in mathematics decreased substantially with decreasing level of index. Table 4 also shows that

Table 1: Average Achievement in Mathematics Content Areas by Gender

Mathematics content areas	Malaysia average achievement			International average scale scores		
	Total	Girls	Boys	Total	Girls	Boys
Overall	508*	512	504	467	486	485
Number	524*	529*	518	467	467	467
Algebra	495*	501*	488	467	471	462
Measurement	504*	505	502	467	464	470
Geometry	495*	495	494	467	466	467
Data	505*	507	503	467	467	467

Note: * Denotes a score significantly higher than the international average.

Table 2: Comparison of Average Achievement in Mathematics Content Areas by Malaysian Girls and Boys

	Sex	*N*	Mean	*SD*	*p*-value
Overall Score	Girl	3,071	512	70.73	0.000
	Boy	2,243	504	74.20	
Number	Girl	3,071	529	67.60	0.000
	Boy	2,243	518	72.56	
Algebra	Girl	3,071	501	68.54	0.000
	Boy	2,243	488	72.77	
Measurement	Girl	3,071	505	76.99	0.157
	Boy	2,243	502	81.50	
Geometry	Girl	3,071	495	75.06	0.867
	Boy	2,243	494	80.38	
Data	Girl	3,071	507	56.61	0.008
	Boy	2,243	503	60.15	

Table 3: Malaysian Students' Reports of Self-Confidence in Mathematics Ability by Gender

Statements	Gender	Agree a lot	Agree a little	Disagree a little	Disagree a lot	*p*-value
I usually do well in mathematics	Girl	11.5	46.2	40.9	1.4	0.001
	Boy	15.1	45.1	38.0	1.8	
Mathematics is more difficult for me than for many of my classmates	Girl	7.8	34.6	45.3	12.2	0.003
	Boy	9.4	35.0	41.1	14.5	
Mathematics is not one of my strengths	Girl	5.7	29.4	44.4	20.5	0.032
	Boy	7.1	30.5	40.9	21.4	
I learn things quickly in mathematics	Girl	9.3	49.1	39.0	2.6	0.000
	Boy	13.1	46.3	37.2	3.4	

the average achievement of the Malaysian students was much higher than the international average for all three levels of index of self-confidence. A slightly higher proportion of the boys than girls came within the high and the low indices. On comparing the average achievement within the same level of index of self-confidence, we can see that girls outperformed boys at every level.

The seven statements presented in Table 5 provide some measure of how the Malaysian students valued the importance of mathematics and the need to learn the subject. The chi-square statistics showed significant differences between responses and gender for all statements except for the statement, "Learning mathematics will help me in my daily life." Just over 95% of both girls and boys agreed that learning mathematics would help them in their daily lives.

Girls were more interested in taking more mathematics in school than were the boys (94% and 90%, respectively). Similarly, the proportion of students who agreed that they enjoyed learning mathematics was higher for girls (90%) than for boys (85%). More girls than boys agreed they needed mathematics to learn other school subjects (85% and 81%, respectively). The need for students to do well in mathematics—to get them into the university of their choice and to get the job they wanted—found agreement among 93% and 88% of the girls, respectively, and among 90% and 86% of the boys, respectively. One reason for these findings may relate to

Table 4: Average Achievement in Mathematics by Index of Students' Self-Confidence in Learning Mathematics

	High index	Medium index	Low index
Malaysia			
Percentage of students'	39	45	16
average achievement	546	490	471
Girls			
Percentage of students'	38	46	16
average achievement	549	493	479
Boys			
Percentage of students'	39	44	17
average achievement	542	486	462
International			
Percentage of students'	40	38	22
average achievement	504	453	433

Table 5: Attitude Toward and Value Malaysian Students Place on Mathematics, by Gender

Statements	Gender	Agree a lot	Agree a little	Disagree a little	Disagree a lot	*p*-value
I would like to take more mathematics in school	Girl Boy	54.3 46.8	39.3 43.3	5.9 9.4	0.5 0.6	0.000
I enjoy learning mathematics	Girl Boy	43.9 37.7	43.1 47.1	11.8 13.9	1.2 1.3	0.000
I think learning mathematics will help me in my daily life	Girl Boy	58.0 54.9	38.3 40.5	3.5 4.5	0.1 0.1	0.071
I need mathematics to learn other school subjects	Girl Boy	28.8 24.9	55.9 56.4	14.6 17.7	0.8 1.0	0.001
I need to do well in mathematics to get into the university of my choice	Girl Boy	58.8 54.3	34.2 36.0	6.6 8.6	0.5 1.1	0.000
I would like a job that involved using mathematics	Girl Boy	17.6 17.3	46.5 47.3	33.3 32.5	2.5 2.9	0.738
I need to do well in mathematics to get the job I want	Girl Boy	48.1 45.2	40.0 41.0	11.4 12.6	0.5 1.2	0.009

the fact that girls generally tend to mature earlier than boys. Girls may therefore be more likely than boys to appreciate learning mathematics not only as a subject but also for its relevance within the wider context, notably its relationship to other subjects in school and to further study and to vocational opportunities after high school.

The seven statements relating to the value students placed on mathematics were converted into an index of the value students placed on mathematics. This index used the same categories as for the index of self-confidence in learning mathematics (Mullis et al., 2004). The results of the analysis, summarized in Table 6, showed that more than three-quarters of the Malaysian students fell within the high index, followed by just over 20% within the medium index. The average achievement decreased substantially with each drop in level of index. Also, for these two levels, the Malaysian students' average achievement was much higher than the international averages. Girls also tended to have a more positive attitude than boys toward learning mathematics; 21% of the girls, compared with 28% of the boys, agreed they would never really understand a new topic if they did not initially understand it. This finding may suggest a difference in the level to which boys and girls are willing to learn something new, and prepared to persevere in that learning. Table 6 also shows a higher proportion of girls than boys in the high index category, although no gender difference is discernible for those students in the low index. Finally, we can see that, in each index, the achievement of girls was higher than that of the boys.

We also examined gender differences in average achievement in terms of the amount of time students spent doing mathematics homework during a normal school week. This factor was included in TIMSS on the premise that improvement in mathematics learning requires practice involving numerous exercises. The index of time spent on mathematics as computed by TIMSS is shown in Table 7. Thirty-three percent of the total sample fell within the high index, 56% the medium index, and the remaining 11% within the

Table 6: Average Achievement by Index of Students' Valuing Mathematics

	High index	Medium index	Low index
Malaysia			
Percentage of students	78	21	1
average achievement	515	486	–
Girls			
Percentage of students	80	20	1
average achievement	517	494	456
Boys			
Percentage of students	75	24	1
average achievement	512	478	455
International			
Percentage of students	55	35	10
average achievement	479	458	458

Table 7: Average Achievement by Index of Time Students Spent Doing Mathematics Homework in a Normal School Week

	High index	Medium index	Low index
Malaysia			
Percentage of students	33	56	11
average achievement	515	510	485
Girls			
Percentage of students	3	57	10
average achievement	519	512	493
Boys			
Percentage of students	34	54	12
average achievement	512	506	474
International			
Percentage of students	26	54	19
average achievement	468	471	456

low index. The average level of achievement of the Malaysian students was substantially better than the international average level of achievement for every index level. It was somewhat surprising, though, to find more boys than girls in the high index category. However, the proportion of boys in the low index category was also much larger than the proportion of girls. At each level of index of time students spent doing mathematics homework, girls achieved higher average scores than boys on the TIMSS mathematics test.

Discussion and conclusion

The results from this secondary analysis of the TIMSS 2003 data for Malaysia show that girls gained significantly higher scores than boys in overall average mathematics achievement as well as in three areas of mathematics content—number, algebra, and data. These findings are similar to those found by Alkhateeb (2001), but contrast with those from several other studies, in which boys generally outperformed girls (Beller & Gafni, 1996; Cleary, 1992; Engelhard, 1990; Gallagher & Kaufman, 2005).

Several dimensions pertaining to students' level of self-confidence in mathematics ability, the value that students placed on mathematics, and the amount of time they spent doing mathematics homework were explored as possible factors affecting these gender differences for the Malaysian students. As expected, average achievement increased with increasing level of index of self-confidence, index of valuing of mathematics, and index of time spent doing mathematics homework. This result held for the total sample as well as for girls and for boys. Within each index level of each of the three dimensions mentioned, girls outperformed boys, although the differences were quite small for a couple of categories.

Among the five mathematics content areas, both sexes scored highest in number, a result that is unsurprising given that students are introduced to number sense much earlier than they are introduced to other content areas. And it is also not surprising that, despite having lower average achievement than girls in almost all the mathematics content areas, the boys had a higher self-confidence level, as perceived by them, than did the girls. The boys seemed to think that they tended to do well in mathematics and that they quickly understood new learning in the subject. The study also revealed that boys spent slightly more time doing mathematics homework. What, then, are the reasons for the girls' greater success on the mathematics test? The reasons could relate to the girls scoring much higher than the boys in terms of enjoyment in mathematics learning, interest in taking up more mathematics subjects, and appreciating the importance of mathematics in relation to other school subjects, university education, and employment. Girls seemed to have a better and wider understanding than the boys of the need to take mathematics learning seriously.

References

Alkhateeb, H. M. (2001). Gender differences in mathematics achievement among high school students in the United Arab Emirates, 1991–2000. *School Science and Mathematics, 101*(1), 1–5.

Beller, M., & Gafni, N. (1996). The 1991 International Assessment of Educational Progress in Mathematics and Sciences: The gender differences perspective. *Journal of Educational Psychology, 88*(2), 365–377.

Cleary, T. A. (1992). *Gender differences in aptitude and achievement test scores.* Paper presented at the 1991 ETS Invitational Conference on Sex Equity in Educational Opportunity, Achievement and Testing, Princeton, New Jersey.

Engelhard, G. (1990). Gender differences in performance on mathematics items: Evidence from the United States and Thailand. *Contemporary Educational Psychology, 15*, 13–26.

Gallagher A. M., & Kaufman, J. C. (Eds.) (2005). *Gender differences in mathematics: An integrative psychological approach.* Cambridge: Cambridge University Press.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth Grades.* Chestnut Hill, MA: Boston College.

Wang, X., & Maxey J. (1995). *Gender differences in the ACT mathematics test: A cross cultural comparison.* Paper presented at the 1995 Annual NCME Meeting, San Francisco, California.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, NJ: Lawrence Erlbaum.

Profiles of scientific competence in TIMSS 2003: Similarities and differences between countries[1]

Marit Kjærnsli and Svein Lie
Department of Teacher Education and School Development
University of Oslo, Norway

Abstract

The aim of the present contribution is to investigate similarities and differences of strengths in science competences between countries, based on TIMSS 2003 data. Analyses are based on systematic investigation of patterns of *p*-values (percentage correct) for individual science items. Hierarchical cluster analysis was applied to establish meaningful groups of countries. The resulting pattern of how countries cluster together into groups of increasing size, based on similarities of strengths and weaknesses, is presented and discussed. As a measure of similarity between countries, we applied the Pearson correlation coefficient to the *p*-value residuals (i.e., each country's set of *p*-values, corrected for the country's average of all items and the international item difficulty). For each group of countries, average *p*-value residuals were calculated to investigate characteristic features. These features are described in terms of separate measures of relative strengths according to item format, subject domain, and cognitive domain. Finally, data on relative emphases in the intended curriculum (curriculum documents) and in the implemented curriculum (percentage of topics taught) explained to a considerable degree the patterns of achievements within the different content domains.

Introduction

Several studies have used achievement data from the TIMSS (Third International Mathematics and Science Study) 1995 assessment to undertake cross-national analyses of students' cognitive strengths and weaknesses (Angell, Kjærnsli, & Lie, 2006; Grønmo, Kjærnsli, & Lie, 2004). Similar analyses have been carried out using data from the OECD Programme in International Students Achievement (PISA) study (Kjærnsli & Lie, 2004; Lie & Roe, 2003; Olsen, 2005). These analyses have been based on systematic investigations of patterns of *p*-values (percentage correct) for individual achievement items. Following a method proposed by Zabulionis (2001), the researchers in these studies applied hierarchical cluster analysis as a tool to establish meaningful groups of countries based on similar areas of relative strengths and weaknesses.

The aim of the present contribution is to further investigate these patterns of cognitive strengths in science, using TIMSS 2003 data (Martin, Mullis, Gonzalez, & Chrostowski, 2004). In 2003, more countries participated in this study than in the earlier studies of mathematics and science achievement, which has allowed us to say more about how countries seem to group together. We also investigate the characteristic features for each of these country groups, and discuss the findings in light of cultural traits and traditions in science education. Finally, we present more evidence to help us determine which factors underlie the mechanism of this clustering of countries. In particular, we try to understand how curricular factors influence this pattern.

In large-scale international studies like TIMSS, uni-dimensional models are applied for scaling student competencies. Consequently, pronounced different item functioning (DIF) across countries is often regarded as a source of measurement error, and is thus used as a criterion for item exclusions. However, test-curriculum analyses (for TIMSS 2003, see Martin et al., 2004, Appendix C) have frequently shown that the exact selection of items for the test has only a minor influence on the countries' relative standing. The position in the present study is that the differential item functioning brings some very interesting information on the strengths and weaknesses of

[1] An update of this paper will be published in a special issue of the journal *Educational Research and Evaluation*.

individual countries. We decided to investigate this effect in a systematic but simple way based on the *p*-value residuals mentioned above. To prevent too many cases (countries), we used country groups as our unit of analysis. In an earlier analysis based on data from TIMSS 1995 (Angell et al., 2006), we applied a similar strategy to construct country groups, but in further analyses, we used one country from each group as the unit of analysis. Because the 1995 assessment offered the most detailed data available on curricular factors (Schmidt, Raizen, Britton, Bianchi, & Wolfe, 1997), we were able to go into more detail for these countries. In this present investigation, we use data from questionnaires given to the science teachers and the national research coordinators (NRCs)who participated in TIMSS 1995.

Clusters of countries

The basis for our analysis was a complete matrix of *p*-values by country, covering 190 items (or rather, score points) and 50 countries (including a few regions within a country). For each cell in this matrix, we calculated the *p*-value residual, that is, how much better or worse (in percentage correct) the students in the particular country performed on the particular item compared to what was expected from the overall achievement of the country (for all items) and the overall difficulty of the item (for all countries). By applying hierarchical cluster analysis to the *p*-value residual matrix, we obtained a pattern of relationships between countries. This pattern can be displayed in a so-called "dendrogram" that, from left to right, illustrates how countries cluster together into groups of increasing size, based on similarities of strengths and weaknesses. As we move from left to right in the dendrogram—that is, from high positive to negative correlations—we see that countries cluster into even larger groups until they all unite. To measure the similarity between two countries, or among already established groups of countries, we applied the (Pearson) correlation coefficient to the *p*-value residuals. Alternative criteria are possible (Olsen, 2005), but the results are similar, even if the details depend on the exact method being applied. Thus, the picture presented in the following represents a reasonably stable solution.

The dendrogram in Figure 1 immediately draws our attention to the remarkable pattern of meaningful groups, thus allowing us to define certain groups of countries that can be identified and labeled according to either location (political or regional unit) or cultural trait (e.g., language). For our further analysis, we concentrated on the following rather distinct country groups, with each group containing at least three countries:

- *English-speaking:* Australia, Canada (Ontario and Quebec), England, New Zealand, Scotland, the USA
- *East-Central Europe:* Estonia, Hungary, Latvia, Lithuania, Russia, Slovakia, and Slovenia
- *East Asia:* Chinese Taipei, Hong Kong SAR, Japan, Korea, Malaysia and Singapore
- *South-East Europe:* Bulgaria, Macedonia, Moldova, Romania, and Serbia
- *Arabic:* Bahrain, Egypt, Jordan, Palestine, and Saudi-Arabia
- *Southern Africa:* Botswana, Ghana, and South Africa
- *Latin:* Italy, Spain (Basque province), and Chile.

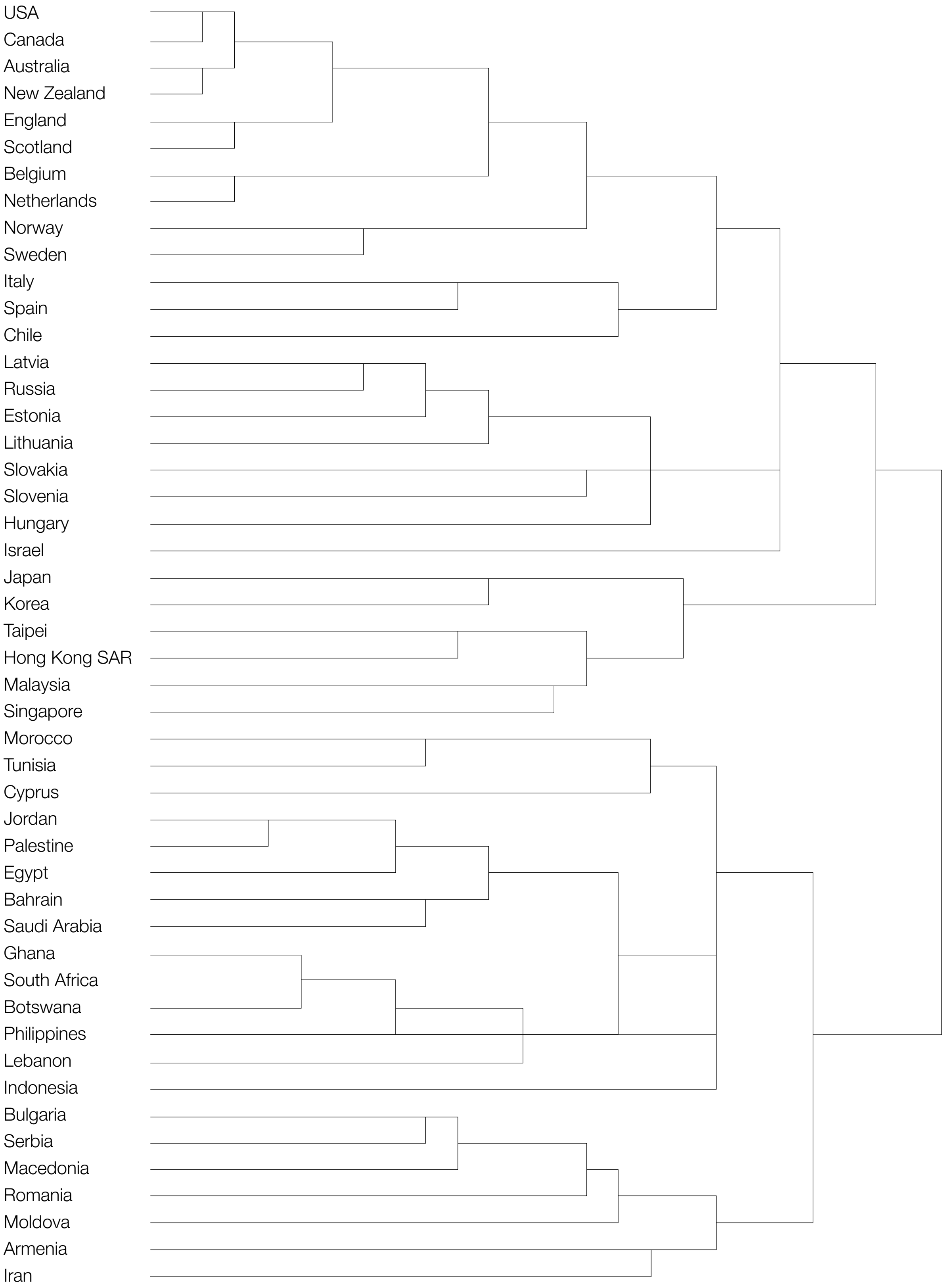
In addition, we included two pairs of countries of particular interest to us:

- *Nordic:* Norway, Sweden
- *Dutch:* The Netherlands and Flemish Belgium.

The labels used above should not be taken too literally, but rather as labels of reference. These nine groups of countries are our focus for the rest of this paper. By calculating average *p*-value residuals for each group, we could evaluate how these values related to characteristic features for items. This, in turn, allowed us to investigate some main characteristics of cognitive strengths and weaknesses for each country group.

We first compared the pattern described above with findings obtained by the same method from other data sets, and from this found the same general patterns as those established from earlier analyses of science achievement data. These earlier analyses related to TIMSS 1995 (Angell et al., 2006), PISA 2000 (Grønmo et al., 2004; Kjærnsli & Lie, 2004), and PISA 2003. Even though the details could differ, mainly due to the fact that countries' participation in the various studies varied, the pronounced linkages within each of the English-speaking countries, the East-(Central) European countries, the East-Asian countries, the Nordic countries, and the Dutch "countries" were confirmed. In addition, three other distinct groups appeared, the Arabic countries, the Southern African countries, and the South-East European (or Balkan)

Figure 1: Dendrogram for Clustering of Countries According to Similarities between Countries in Patterns across Science Items



countries. It was not so easy, however to label the group of "Latin" countries linked together by our data, and we stress that the label applied should not be taken literally.

Characteristic features of country groups

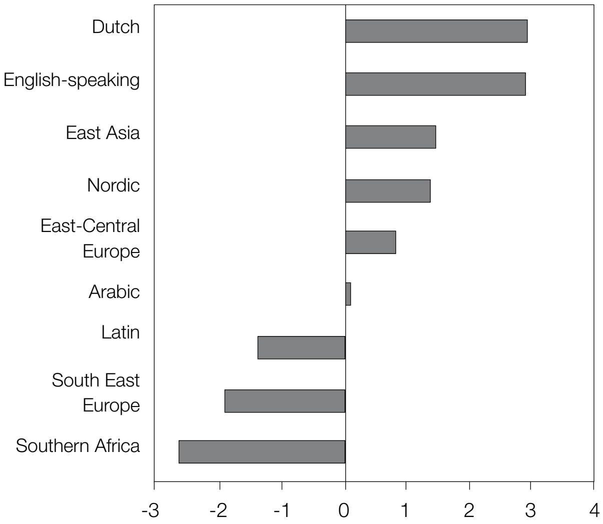
We now turn our attention to the essential features characteristic of each of the above groups. What did the countries in each group have in common? Our approach consisted of classifying all science items according to selected criteria, and investigating how these classifications related to the patterns of *p*-value residuals for each of the country groups (Olsen, 2005). The following item criteria applied:

- *Item format:* constructed response versus multiple choice
- *Science content:* life science, chemistry, physics, earth science, or environmental science
- *Cognitive domain:* factual knowledge, conceptual understanding, or reasoning and analysis.

Item format

The TIMSS 2003 achievement test consisted of both multiple-choice items and constructed-response items. The distribution between these two formats was about 60% multiple choice and 40% constructed response. In the following, we look more closely at how the country groups performed in relation to these two item formats. Figure 2 compares the relative strengths within the two formats. From this figure, we can see in particular that the Dutch and the English-speaking groups of countries performed relatively better on the constructed-response items than did the other groups of countries. On the other hand, the groups from Southern Africa and South-East Europe performed substantially better on multiple-choice items.

Figure 2: Constructed Response Items versus Multiple Choice Items for Each Country Group



Note: Positive values in favour of constructed response items.

Science content domains

The TIMSS 2003 Assessment Framework defined five content domains in Population 2 (i.e., 13- or 14-year-olds): *life science, chemistry, physics, earth science,* and *environmental science*. Each content domain had several main topic areas, presented as a list of specific assessment objectives. For a more detailed description, see *TIMSS Assessment Framework* (Mullis et al., 2001).

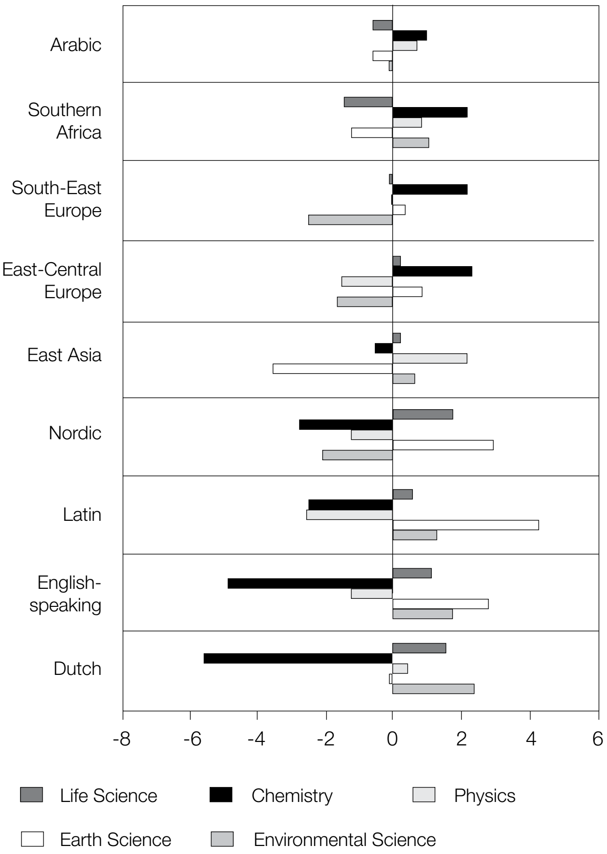
Figure 3 displays the relative achievement strengths in each content domain. The figure shows the country groups sorted by increasing spread among the domains, so that those groups with the most distinguished profiles appear toward the bottom. Some remarkable characteristics stand out in this figure. The most pronounced is the fact that the variation among the groups is much larger in chemistry and somewhat larger in earth science than it is in life science and physics. We can also see that the Dutch and English-speaking countries performed relatively much worse in chemistry than in the other content domains. The Nordic, Latin, and English-speaking groups performed particularly well in earth science; the performance of the East-Asian group was weak in this content domain.

One extreme case may illustrate the situation for chemistry in the Dutch and English-speaking countries. The following item (S022202) represents the main topic "Particular structure of matter" within the cognitive domain of factual knowledge.

What is formed when a neutron atom gains an electron?
A, A mixture
B. An ion
C. A molecule
D. A metal

The *p*-value residuals for this item for the Dutch and the English-speaking groups were as low as -30 and -25, respectively, which means as many as 30 and 25 percentage points respectively lower than what was expected based on the overall abilities for these country

Figure 3: Achievement in Science Content Domains for Each Country Group, Sorted by Increasing Spread among the Domains



groups and the overall difficulty of this item. The item simply required students to recognize the correct term; that many apparently did not signals that the countries in question do not regard such information as an important part of the chemistry curriculum.

Cognitive domains

All science items in TIMSS 2003 were classified into one of three cognitive domains according to how the students were expected to act or the type of cognitive activity they needed to engage in to reach a correct response. These three domains were assessed across the science content domains: *factual knowledge, conceptual understanding*, and *reasoning and understanding* (Mullis et al., 2001).

Within the category *factual knowledge*, the students needed to demonstrate knowledge of relevant science facts, information, tools, and procedures. Thus, this category involves more than just memorization and recall of isolated bits of information. *Conceptual understanding* required students to extract and use scientific concepts and principles to find solutions and develop explanations, to support statements of facts or concepts, and to demonstrate relationships, equations, and formulas in context. The problems in this cognitive domain involve more straightforward applications of concepts and require less analysis than is the case with items in the reasoning and analysis domain. *Reasoning and analysis* covers challenges like solving problems, developing explanations, drawing conclusions, making decisions, and extending knowledge to new situations. The students were, for example, expected to evaluate and make decisions based on their conceptual understanding. Some items required students to bring knowledge and understanding from different areas and apply it to new situations.

Figure 4 displays the relative strengths and weaknesses concerning cognitive domains among the country groups. As in Figure 3, the most distinguished profiles appear near the bottom of the figure. And again, the Dutch and English-speaking groups stand out, with similar profiles, including a particular relative weakness in *factual knowledge*. This domain appears to have the most variation among the groups; the variation with the *conceptual understanding* domain is much less.

The role of the intended curriculum and the implemented curriculum

Having discussed some characteristics of each country group, we now focus, for the remaining part of this paper, on the extent to which other data from TIMSS explained these characteristics. In particular, we investigate the role of what, in TIMSS, are called the *intended* and the *implemented* curriculum, respectively. By intended curriculum, we mean the curricular documents and how these give prescriptions for distribution of emphasis across subject and cognitive domains. The implemented curriculum refers to what is actually taught and the emphasis given to the different aspects.

Item format

The science teacher questionnaire included a question on the relative frequency of different item formats in the assessment of students in science. These data (given in Exhibit 7.13 in Martin et al., 2004) are graphically displayed for our country groups in Figure 5. This

Figure 4: Achievement in the Cognitive Domains for Each Country Group, Sorted by Increasing Spread among the Domains

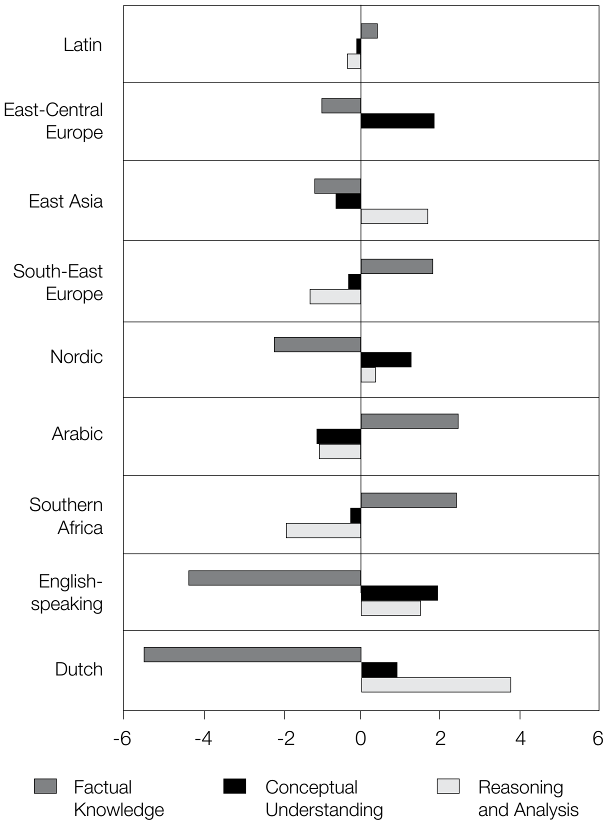


figure shows that the dominant response is an even distribution of the two item formats. However, the Nordic group (i.e., Norway and Sweden) stands out with a very different profile, with multiple-choice items playing essentially no role in assessment practice. This is interesting information in itself, but it does not provide explanation for the pattern shown in Figure 3. The other features in Figure 5 also offer little explanation for the pattern in Figure 3. Our conclusion, therefore, is that the item formats applied did not play a strong role in shaping the results in the TIMSS science test.

Science content domains

The intended curriculum

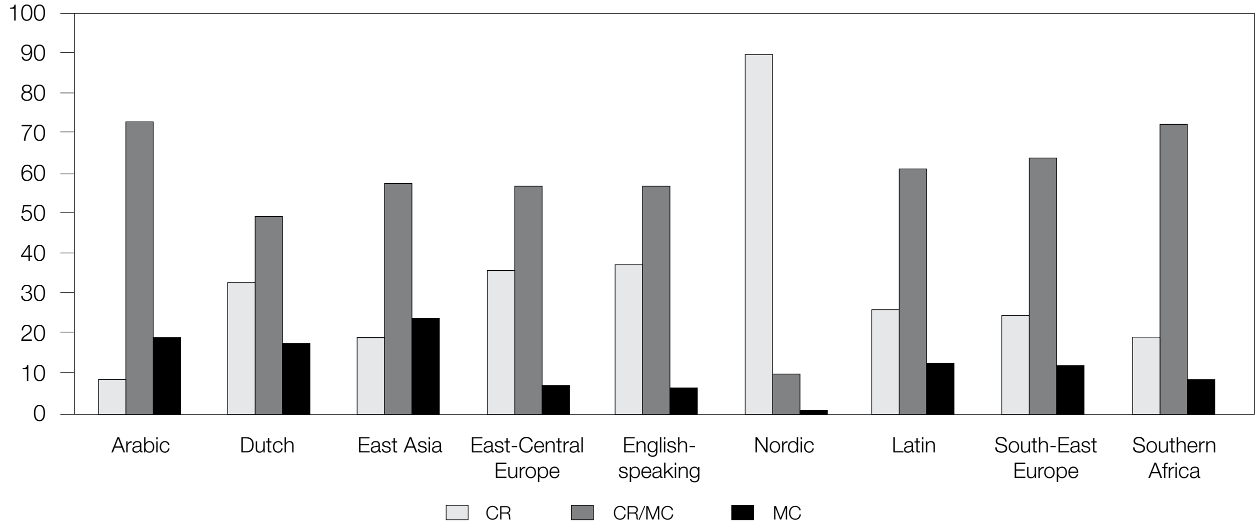
The NRCs for TIMSS 2003 responded to a questionnaire on the national context for mathematics and science in their respective countries. They responded to specific questions on curricular coverage for a series of science (and mathematics) topics as they were described in the assessment framework (Mullis et al., 2001). For each of these 44 science topics, there are data on whether NRCs expected the topic to be taught up to and including the actual grade (Grade 8 for most countries). We have taken these data as our measure of the intended curriculum, and applied them in the form of the *percentage of the given topics* covered, as given for each subject domain in Martin et al. (2004, Exhibit 5.7).

Table 1 shows the intended curriculum by the above method averaged within each country group. Note that the environment domain contains only three topics, which means the data in Table 1 are less reliable for this domain than they are for the others. The far-right column of the table gives the (Pearson) correlation of these five numbers, with the corresponding *p*-value residuals displayed in Figure 2. These correlations are all positive and of medium size for most groups.

Table 1: Percentage of Science Topics in the TIMSS Framework Covered by the National Intended Curriculum (To Be Taught Up to and Including Grade 8), and Correlation with Achievement

Groups of countries	Life Science	Chemistry	Physics	Earth Science	Environmental Science	Correlation with *p*-value residuals
Arabic	93	83	96	75	100	0.13
Dutch	81	41	36	25	83	0.54
East Asia	69	68	79	55	56	0.71
East-Central Europe	79	92	91	92	90	0.03
English-speaking	76	67	75	78	61	0.27
Nordic	92	67	71	78	100	0.16
Latin	82	73	63	67	78	0.10
South-East Europe	82	93	96	87	80	0.65
Southern Africa	56	66	73	64	72	0.64

Figure 5: The Relative Distribution of Item Formats Used in Science Assessments



Note: * CR = mostly Constructed Response; MC = mostly Multiple Response; CR/MC = about half of each.

The average across all groups is 0.36. Thus we notice a clear and positive relationship between (relative) achievement and curricular emphasis in the data.

The implemented curriculum

Next, we investigated the parallel relationship between achievement and the *implemented* curriculum. In the science teacher questionnaire, the teachers were asked to state which out of the list of 44 framework topics would actually be taught up to and including the present school year. Exhibit 5.8 in Martin et al. (2004) gives this information in the form of the percentage of students taught the topics within each of the content areas. Table 2 gives this information for each of the country groups, in addition to the Pearson correlation, with *p*-value residuals. A few countries did not provide comparable data for environmental science, so there are three empty cells in the table. The correlations in these cases are calculated for the four other domains only.

Not unexpectedly, we find that the correlations are generally somewhat higher in Table 2 than in Table 1. The average is 0.48. Our data on emphasis in the classrooms explain more of the relative strengths and weaknesses than do the intended curriculum.

Also, note in particular, the negative correlation for the English-speaking group in Table 2. Whereas the low coverage for chemistry in Table 1 is reflected in Figure 3, this is not paralleled for Table 2. Here, the English-speaking "profile" shows no particular

Table 2: Average Percentages of Students Taught the TIMSS Science Topics, and Correlation with Achievement

Groups of countries	Life science	Chemistry	Physics	Earth Science	Environmental Science	Correlation with *p*-value residuals
Arabic	74	76	83	61	54	0.55
Dutch	72	33	39	42		0.68
East Asia	59	71	71	34	38	0.59
East-Central Europe	73	80	61	88		0.80
English-speaking	65	65	64	66	57	-0.28
Nordic	54	54	48	68	33	0.70
Latin	83	72	67	76	69	0.46
South-East Europe	87	93	92	91		0.55
Southern Africa	51	44	42	28	45	0.23

drop for chemistry. It appears that even when the teaching covers chemistry topics reasonably well, the characteristics of how those topics are addressed may be, to some degree, at odds with what is required by the TIMSS chemistry items. The item discussed above (S022202) appears to be an extreme example.

Intended, implemented, and achieved curricula

In Figure 6, we have tried to illustrate all three curriculum levels—*intended, implemented*, and *achieved*—for some of the country groups. Here we have applied the TIMSS notation of *achieved curriculum* for the assessment results, in the form of *p*-value residuals. To simplify comparison between the shapes of the three curves, the achieved curriculum data have somewhat arbitrarily been multiplied by five. Although the numbers are not directly comparable, it is nevertheless interesting to compare the shape of the three curves for selected groups of countries. We have here selected the three clusters with most countries in addition to the Nordic group, which is of special interest to us.

Figure 6 clearly shows the similarities between the three curves for each of the displayed country groups. The situation for the other groups of countries is similar. However, there are some interesting differences, especially in relation to the results for the domain environmental science for the Nordic group. Here, we can see a clear difference between the intended curriculum on the one hand and the implemented curriculum and the achieved curriculum on the other hand. The gap between intended and implemented cannot be explained by our data, but the question it poses will be an interesting one to consider further. It may be easier to give this area emphasis in the intended curriculum than to follow it up in the classroom.

Science cognitive domains

This final section of our paper considers the emphasis each group of countries put on the three cognitive domains in their intended science curricula. The data for the cognitive domains are taken from Exhibit 5.6 in Martin et al. (2004). Columns 2 and 3 of Table 3 show the data relating to the first two categories. The findings presented here are identical with what Martin and his colleagues reported. For the third category (Column 4), we collapsed, with some doubt, three categories—"writing explanations about what was observed and why it happened," "formulating

Figure 6: Comparison among Intended, Implemented, and Achieved Curricula for Four Groups of Countries

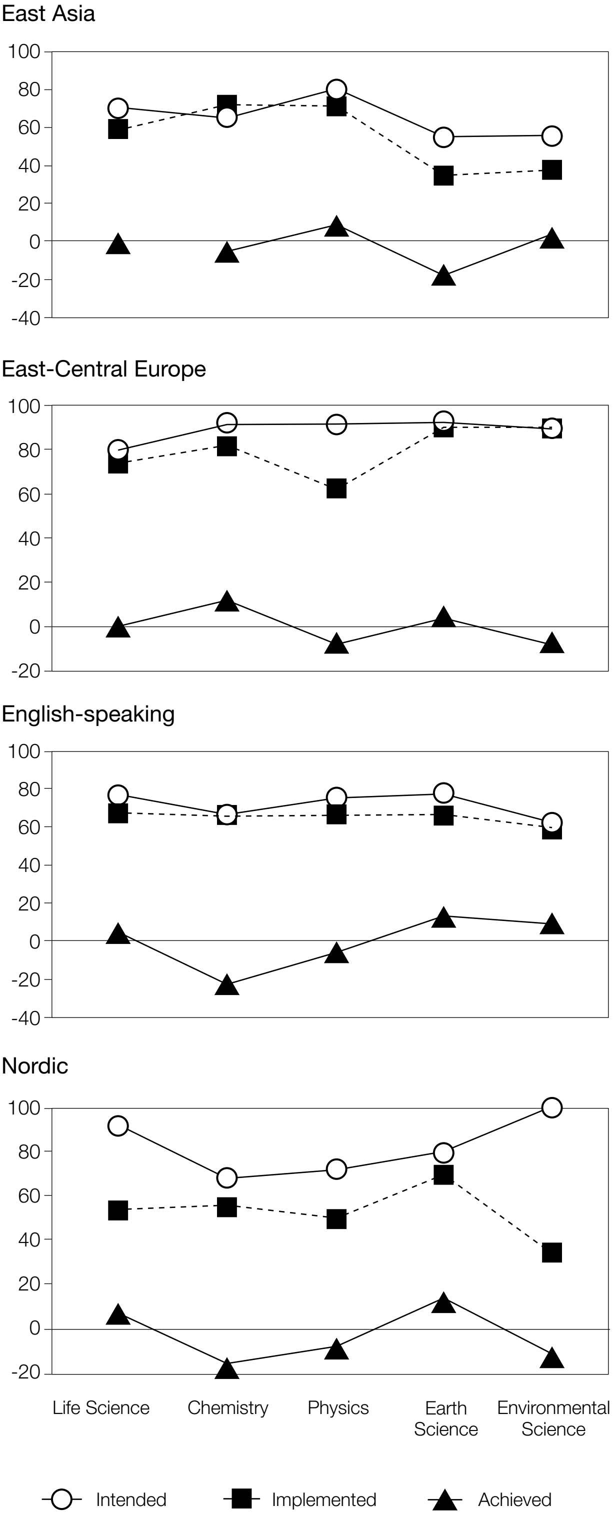


Table 3: Emphasis Countries Placed on Cognitive Domains in the Intended Curriculum

Groups of countries	Factual Knowledge	Conceptual Understandings	Reasoning and Analysis
Arabic	3.80	4.00	3.27
Dutch	3.50	3.00	2.67
East Asia	4.00	4.00	3.28
East-Central Europe	3.57	3.57	2.67
English-speaking	3.71	4.00	3.67
Nordic	3.50	3.50	3.33
Latin	4.00	4.00	3.00
South-East Europe	3.60	3.40	2.47
Southern Africa	3.67	3.67	2.67

hypotheses or predictions to be tested," and "designing and planning experiments or investigations"—into one category, which we called "reasoning and analysis." This merging of categories is not obvious and contributes to a degree of credibility of the data in Table 3 that is somewhat lower than that of the data in the other tables. The simple scale that we applied to measure "emphasis" (a lot of, some, very little, and no emphasis) also contributes to the lower quality of the data in Table 3. However, we nevertheless calculated the correlations with the achievement data for the pattern of the three different cognitive domains.

As we might have suspected, the data in Table 3 did not explain, to any extent, the profiles in Figure 4. The average correlation coefficient with *p*-value residuals was as low as 0.07. In an analysis of country differences regarding cognitive profiles in mathematics in TIMSS, Klieme and Baumert (2001) classified each item according to a set of cognitive demands. For each item, the dependence on each of these categories was coded by a group of coders. This multidimensional approach allowed the researchers to obtain data on differential item functioning and then to compare this information with what was expected from national analyses of various sources. Klieme and Baumert analyzed several countries in this way, and obtained meaningful results. However, in our analysis of items classified according to three mutually exclusive categories, we were unable to obtain a meaningful relationship between curriculum and achievement information.

Summary and conclusion

The analysis presented in this paper involved three steps. First, we identified some country groups with similar profiles of relative strengths from item to item. Second, we described some characteristic features for each of these country groups. And third, we looked into some other TIMSS data, that is, emphases in the intended curriculum as well as topics implemented by teachers in the classrooms. We found that these factors provided, to some extent, explanations for the patterns of features for country groups featured in the figures and tables in this paper. However, our search for relating the cognitive profiles to curriculum factors did not lead to further understanding.

To go deeper into the country (or group) profiles, one would need stronger tools to handle the differential functioning aspect. Using measures of item difficulties would be more appropriate than using (residuals of) *p*-values. Also, it may be possible, in the future, to scale the items by applying within-item multidimensionality to model the inner complexity of individual items. Such an approach might then allow us to build the cognitive profiles directly into the scaling procedure.

References

Angell, C., Kjærnsli, M., & Lie, S. (2006). Curricular effects in patterns of student responses to TIMSS science items. In S. J. Howie & T. Plomp (Eds.), *Contexts of learning mathematics and science* (pp. 277–290). Leiden: Routledge.

Grønmo, L. S., Kjærnsli, M., & Lie, S. (2004). Looking for cultural and geographical factors in patterns of responses to TIMSS items. In C. Papanastasiou (Ed.), *Proceedings of the IRC-2004 TIMSS Conference* (pp. 99–112). Lefkosia: Cyprus University Press.

Kjærnsli, M., & Lie, S. (2004). PISA and scientific literacy: Similarities and differences between the Nordic countries. *Scandinavian Journal of Educational Research, 48*(3), 271–286.

Klieme, E., & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education, 16*(3), 385–402.

Lie, S., & Roe, A. (2003). Exploring unity and diversity of Nordic reading literacy profiles. In S. Lie, P. Linnakylä, & A. Roe (Eds.), *Northern lights on PISA: Unity and diversity in the Nordic countries in PISA 2000* (pp. 147–157). Oslo: Department of Teacher Education and School Development, University of Oslo.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international science report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzales, E. J., Chrostowski, S. J., & O'Connor, K. M. (2001). *TIMSS assessment frameworks and specifications 2003.* Chestnut Hill, MA: Boston College.

Olsen, R. V. (2005). An exploration of cluster structure in scientific literacy in PISA: Evidence for a Nordic dimension? *NorDiNa, 1*(1), 81–94.

Schmidt, W. H., Raizen, S. A., Britton, E., Bianchi, L. J., & Wolfe, R. G. (1997): *Many visions, many aims: A cross-national investigation of curricular intentions in school science.* Dordrecht: Kluwer Academic Publishers.

Zabulionis, A. (2001). Similarity of mathematics and science achievement of various nations. *Educational Policy Analysis Archives, 9*(33). Retrieved from http://epaa.asu.edu/epaa/v933/

Test-taking motivation on low-stakes tests: A Swedish TIMSS 2003 example

Hanna Eklöf
Department of Educational Measurement
Umeå University, Sweden

Abstract

The objective of this study was to investigate the test-taking motivation of students in the Swedish TIMSS 2003 context. Swedish Grade 8 students formed the study sample, and the focus was on mathematics. Test-taking motivation was measured using questionnaire items and interviews, and reported level of test-taking motivation was regressed on test score. The questionnaire study showed that the Swedish students in general reported that they were well motivated to do their best in TIMSS. According to regression analysis, test-taking motivation was positively and significantly, though rather weakly, related to mathematics achievement. The interview study mainly corroborated the major results from the questionnaire study but added some complexity to the results. Although most interviewed students reported they were well motivated to do their best in TIMSS and that they valued a good performance, they nevertheless seemed aware of the fact that the test was low-stakes for them personally. Many students also reported competitive, comparative, or social responsibility reasons as motivating, while other students seemed more intrinsically motivated to do their best. Findings from quantitative as well as qualitative analyses suggest that the Swedish TIMSS result is unlikely to be negatively affected by a lack of student motivation. However, nothing is known about student test-taking motivation in other countries participating in TIMSS, and further research exploring this issue in an international context is warranted.

Introduction

The present paper presents parts of a research project conducted in the Swedish Trends in Mathematics and Science Study (TIMSS) 2003 context. The project is exploring different aspects of student achievement motivation from a measurement perspective and a validity perspective (see Eklöf, in press; Eklöf, 2006a, 2006b).

Student motivation is a core issue in educational settings, as achievement motivation is assumed to interact with achievement behavior in important ways (Pintrich & Schunk, 2002; Wigfield & Eccles, 2002). Achievement motivation can be conceptualized and measured on different levels of generality. The most common type of motivational measure is domain-specific and measures achievement motivation for a particular domain (e.g., mathematics, science). However, achievement motivation can also be conceptualized and measured on a situation-specific level, that is, motivation to perform well in a given situation, or on a given test. Situation-specific motivation or test-taking motivation is the focus of the present paper.

A positive motivational disposition toward the test is often assumed to be a necessary though not sufficient condition for a good test performance (Cronbach, 1988; Robitaille & Garden, 1996; Wainer, 1993; Zeidner, 1993). Messick (1988) noted that a poor test performance could be interpreted not only in terms of test content and student ability, but also in terms of lack of motivation. If different groups of students differ systematically in level of motivation, and if less motivated students are disadvantaged in that they score below their actual proficiency level, then test-taking motivation is a possible source of bias (Baumert & Demmrich, 2001; Mislevy, 1995; O'Leary, 2002; O'Neil, Sugrue, Abedi, Baker, & Golan, 1997; Robitaille & Garden, 1996; Wainer, 1993; Zeidner, 1993) and is hence a threat to the validity of score interpretation and use (Messick, 1995).

The issue of student test-taking motivation thus is an issue of validity, and an issue of the trustworthiness of test results. However, knowledge of how individuals perceive the tests they are designated to complete, and their motivation to do their best on these tests, is scarce

(Baumert & Demmrich, 2001; Nevo & Jäger, 1993), not least in the context of large-scale, comparative studies.

Test-taking motivation and low-stakes tests

Tests that have no personal consequences, that is, low-stakes tests, are often assumed to cause a decrease in motivation and performance (Wise & DeMars, 2003; Wolf & Smith, 1995; Wolf, Smith, & Birnbaum, 1995). TIMSS is, in several aspects, a low-stakes test, and the issue of test-taking motivation is therefore highly relevant in the TIMSS context. Indeed, a rather common concern in the TIMSS context is that not all students are motivated to do their best on the test and that the results therefore can be an underestimation of student knowledge (Baumert & Demmrich, 2001).

First, the result on the TIMSS test has no impact on student grades in mathematics or science. Second, the results in TIMSS are mainly summarized at a national level and no individual results are given to the students or the schools. Thus, the students and their teachers, parents, and peers never know the result of an individual student. However, one may argue that the fact that the students represent their country in a world-wide comparative study is motivating for the students. One may also argue that the low stakes of the test make the students less anxious, and that they therefore achieve as well as they would on an ordinary test, although they are not maximally motivated.

Previous research on test-taking motivation

A vast amount of research has investigated various aspects of general and domain-specific achievement motivation. The research on situation-specific motivation or test-taking motivation is anything but vast. Studies are scattered in time and place, theoretically and methodologically. However, the expectancy-value theory of achievement motivation (Eccles & Wigfield, 2002; Pintrich & De Groot, 1990; Wigfield & Eccles, 2002) has been applied to a number of studies investigating test-taking motivation (Baumert & Demmrich, 2001; Sundre & Kitsantas, 2004; Wolf & Smith, 1995; Wolf et al., 1995), and was the theoretical framework used in the investigation of test-taking motivation in the Swedish TIMSS 2003 context as well (see Eklöf, 2006b). The expectancy-value theory is comprehensive in order to mirror as many as possible of the processes underlying motivated behavior and includes many contextual and psychological aspects that have been shown to interact with and influence achievement choices and achievement behavior.

Although comprehensive, the model has two core components: one expectancy component that corresponds to the question "Can I do this task?" and one value component that corresponds to the question "Do I want to do this task and why?" The expectancy component in the model thus refers to the individual's beliefs and judgments about his or her capabilities to do a task and succeed at it. The value component in the model refers to the various reasons individuals have for engaging in a task or not (see Eccles & Wigfield, 2002; Eklöf, 2006b). In the present paper, the focus is mainly on the value component in the model (see Eklöf, 2006b, for an elaborated presentation of the expectancy-value theory).

The results from earlier studies that focus on test-taking motivation have been somewhat inconclusive and, in many cases, the link between reported level of motivation and actual achievement has been weak. Studies have found that students are quite motivated even when the test is low-stakes for them (Center for Educational Testing and Evaluation, 2001), that raising the stakes does not always contribute to a corresponding rise in motivation and achievement (Baumert & Demmrich, 2001; O'Neil, Abedi, Miyoshi, & Mastergeorge, 2005), and that reported level of test-taking motivation is weakly associated with subsequent performance (O'Neil et. al., 2005; Zeidner, 1993). Other studies, however, have found that the stakes of the test do have an impact on motivation and performance (Chan, Schmitt, DeShon, Clause, & Delbridge, 1997; Wolf & Smith, 1995; Wolf et al., 1995).

In summary, it is not clear from previous empirical studies whether the validity of low-stakes tests like TIMSS is threatened by a lack of motivation among the participants because it is not clear if (a) the participating students are lacking motivation, and/or (b) rated level of test-taking motivation interacts with test performance. The present study explores these issues in a Swedish TIMSS 2003 context.

The study presented in this paper is also concerned with issues of measurement validity. According to validity theorist Samuel Messick, an important aspect to consider from a validity viewpoint is the "social psychology of the assessment setting ... [which]

requires careful attention" (Messick, 1989, p. 14). How do respondents react to the tasks that are presented to them in TIMSS 2003? Do they perceive the tests they are about to complete as valid? As important? These are issues vital for the validity of interpretation and use of test scores. This consideration is particularly true for studies like TIMSS, which involve so many students from so many countries and cultures. Unfortunately, these are also issues that have been all but forgotten in the research on the validity of large-scale, comparative studies.

Study objective

The main objectives of the present study were to investigate the reported level of test-taking motivation and the relationship between test-taking motivation and mathematics test performance, as well as to explore student perceptions of test stakes and task value in a sample of the Swedish Grade 8 students who participated in TIMSS 2003.

Method

Participants

A sample (*n* = 350) of the Swedish eighth-graders who participated in TIMSS 2003 took part in the study. They completed a questionnaire on test-taking motivation before they took the TIMSS test. Of these students, 343 were valid cases in the TIMSS database, and this sample was the sample used in the present study. The sample consisted of 174 boys (50.7%) and 169 girls (49.3%). Students came from 17 randomly sampled classes that participated in TIMSS. Approximately half the sample was 14 years old at the time of testing; the other half was 15 years old. A previous study based on the same sample of students showed that the present sample was representative of the Swedish TIMSS 2003 participants (Eklöf, 2006a). Of these 343 students, 329 completed the open-ended questionnaire item also analyzed in the present study. Further, 30 students (15 boys and 15 girls) from this sample agreed to be interviewed about their experience of participating in TIMSS 2003.

Measure, procedure, and data analysis

No established measures of test-taking motivation were available, and the TIMSS student background questionnaire had no items that asked about test-taking motivation. Therefore, a test-taking motivation questionnaire was developed and applied in the Swedish TIMSS context (see Eklöf, 2006a, for a description of the development and validation of this questionnaire). The students completed the test-taking motivation questionnaire before they took the TIMSS test. Also, two items asking about test-taking motivation were added to the TIMSS student background questionnaire as national options. The students completed the student background questionnaire after the TIMSS mathematics and science test. For validation purposes, post-test interviews were performed with a smaller sample of the students (*n* = 30) (see Eklöf, 2006b, for a detailed study design).

The test-taking motivation questionnaire

This questionnaire is a self-report instrument developed to measure aspects related to student test-taking motivation (Eklöf, 2006a, see also Eklöf, in press). The *Expectancy-Value Model of Achievement Motivation* (Eccles & Wigfield, 2002) was used as the general theoretical basis for the development and interpretation of this questionnaire.

According to exploratory factor analysis on data from the present sample (see Eklöf, 2006a), four items (translated from Swedish) formed a scale that was used as a measure of mathematics test-taking motivation in the present study. These items were:

Item 1: How motivated are you to do your best on TIMSS' mathematics items? (pre-test measure)

Item 2: How important is it for you to do your best in TIMSS?

Item 3: How much effort will you spend on answering the mathematics items in TIMSS?

Item 4: How motivated were you to do your best on TIMSS' mathematics items? (post-test measure)

This four-item scale was named Test-Taking Motivation (TTM). All items in the scale were measured on a four-point scale, with ratings ranging from a highly unfavorable attitude to a highly favorable attitude (e.g., 1 = *not at all motivated*, 4 = *very motivated*) (see Eklöf, in press, for a more detailed presentation of the items in the scale).

The test-taking motivation questionnaire (TTMQ) also contained an open-ended item that read "*Describe in your own words how motivated you feel to do your best in TIMSS and why*" (translated from Swedish).

The item was assumed to generate answers revealing something about the students' perceptions of task value and of the stakes of the TIMSS-test. All students responded to this item before they completed the TIMSS mathematics and science test.

The interviews

Short semi-structured interviews were conducted with 30 students. The interview guide contained a list of topics that were to be explored in each interview. These focused on the students' perceptions of TIMSS 2003 in terms of test stakes and task value, their reported level of test-taking motivation before and during the test, their perceived importance of a good performance, and the effort they reported investing when completing the TIMSS mathematics and science items. The students were also asked to compare the TIMSS test to regular achievement tests in school.

TIMSS 2003 mathematics test

In TIMSS, each student completes a booklet containing only a sample of the total number of mathematics/science items that are used in the study. It is therefore impossible, when using raw data, to calculate a total score that can be compared over populations and sub-populations. To obtain comparable achievement scores, each student obtains a scaled score, which represents an estimation of his or her score if the student had answered all items (see Martin, Mullis, & Chrostowski, 2004). Two different kinds of scores are estimated for each student. One is the national Rasch score (see Eklöf, in press, for analyses of the relationship between test-taking motivation and test performance using the Rasch score as a dependent variable). The other score consists of five "plausible values" for each student. These values are imputed values obtained through complex item response modeling. The five values an individual obtains are random excerpts from the distribution of possible values for that individual. The mean plausible value is set to 500, with a standard deviation of 100. All achievement results reported in TIMSS internationally are based on these plausible values; in the present study, the first plausible value was used as the dependent variable.

Quantitative data analysis

First, descriptive statistics and correlations between variables were computed. Then, student ratings of test-taking motivation were regressed on mathematics score. In this regression, two motivational scales used in TIMSS internationally—mathematics self-concept (MSC) and student valuing of mathematics (VoM)—were included in the analysis. They were held constant in order to investigate whether the test-taking motivation scale explained any variance in the mathematics score not explained by these two motivational variables. All analyses were performed in SPSS. All tests of significance were two-tailed and the alpha level was set to .05.

Qualitative data analysis

The wording of the open-ended item in the TTMQ (*Describe in your own words how motivated you feel to do your best in TIMSS and why*) was not intended to lead the students into specified response categories. Accordingly, the analysis of this item initially followed a rather unbiased bottom-up procedure. Here, response categories were not specified in advance. Instead, the students' answers were content-analyzed and interpreted mainly in terms of the students' perceptions of the stakes of the TIMSS test and their perceptions of task value. Common themes in the students' responses were then identified, and students giving similar responses were joined together in one category. Only the main common themes are reported below.

The analysis of the interviews was mainly descriptive in nature. The students' responses to, and elaborations on, the topics included in the interview guide were generally taken at face value and interpreted in terms of task-value perceptions and perceptions of test stakes. Students' responses to the open-ended item and the interview topics were assumed to reveal more about the students' perceptions of the TIMSS test than were questionnaire items with a closed item format. The open-ended item and the interviews were also used as tools for validating the questionnaire results (for more details, see Eklöf, 2006b).

Findings

The findings are presented as follows:

1. A description of the students' ratings on the TTM scale as well as their ratings of the individual items.
2. An account of the multiple linear regressions exploring the relationship between the TTM scale and mathematics score, with students' ratings of mathematics self-concept and of their valuing of the mathematics subject held constant.
3. A brief summary of the results from the open-ended item in the TTMQ.
4. A summary of the main findings from the interviews.

Reported level of test-taking motivation

The TTM scale, which included the four TTMQ items listed above, had a score reliability coefficient of α = .79, which is acceptable given that the scale consisted of only four items. The maximum value of the TTM scale was 4.0, and the mean value for the present sample was 3.09 (*SD* = .55), which indicates that the students in the sample on average reported a fairly high level of test-taking motivation.

In regard to the individual items in the TTM scale, a majority of the students in the sample (*n* = 343) reported that they were either very motivated or somewhat motivated to do their best on TIMSS mathematics items before (89%) as well as after (76%) taking the test. A majority of the students said that it was either very important or rather important for them to do their best in TIMSS (74%), and that they would spend a lot of effort or a fair amount of effort (90%) when answering the TIMSS mathematics tasks (see Eklöf, in press, for more detailed results).

Relationships between ratings of test-taking motivation and mathematics score

For the total sample, the TTM scale was significantly but rather weakly correlated with the mathematics score (the first plausible value) (r = .25, p < .01).

As noted above, to investigate whether the TTM scale accounted for any variation in the TIMSS mathematics score when other relevant variables were held constant, a regression model was built with the TTM scale and two motivational scales used in TIMSS internationally—mathematics self-concept (MSC) and student valuing of mathematics (VoM)—as independent variables. The first plausible value in mathematics was used as the dependent variable

According to this model, the three independent variables together explained about 39% (R^2) of the variation in the mathematics score for the present sample. Most of this variation was explained by the MSC variable (β = .60, t = 12.87, p < .01). The TTM variable had a positive and statistically significant, though rather weak, relationship to the mathematics score when the other independent variables were partialed out (β = .11, t = 2.36, p < .05). The VoM variable was weakly negatively related to the mathematics score when the effect of the other independent variables was partialed out (β = .-.02, t = -.44, p = n.s.).

The open-ended item

There were 329 valid responses to the open-ended item in the TTMQ that was administered to the students (*n* = 343) before they completed the TIMSS mathematics and science test. In their answers, 238 of these 329 students (72%) expressed themselves in positive terms as regards their participation in TIMSS and their motivation to do their best. Forty-eight students (15%) gave rather indifferent answers to this open-ended item, and 43 students (13%) reported a negative motivational disposition toward the TIMSS test.

In relation to the question of what motivated the students to do their best on a test like TIMSS, those students reporting a positive motivational disposition toward the test in their answers to the open-ended item were grouped into three major categories:

1. This category contained students who mainly gave comparative/competitive reasons (CR) for their motivation to do well: they wanted to do their best because they were to be compared to other countries; they wanted to show that Sweden is a prominent country.
2. The students in this category expressed a social responsibility (SR) as the main reason for why they wanted to do their best: they wanted to do their best because they had been chosen for this study; they wanted to do their best to help with the research.
3. The students in this category gave more personal reasons (PR) for their motivation to do well: they always did their best; they wanted to do their best to test themselves; they wanted to see how much they knew.

Examples of what the students in each of these categories said follow:

- *I am fairly motivated to do my best, as it is a competition. And you'd rather win.*
- *I am motivated to do my best. I think it is an important test to see how children in different parts of the world work and how they solve problems.*
- *I want to do my best to see how much I have learned over the years.*

About half of the students reporting a positive motivational disposition toward TIMSS were categorized either in the CR category (67 students) or in the SR category (50 students). Nineteen percent (44 students) of the students reporting a positive motivational disposition toward TIMSS were categorized in the PR category.

Among the 43 students who claimed that they were *not* well motivated to do their best, two main categories were identified. The first category included students who reported the low stakes of the test as the reason for why they were not maximally motivated (the result did not count for their grades; they would never know the results). Sixteen students were coded as belonging to this category. The students in the second group were those who reported they were not motivated because they did not like school, the school subjects tested, or tests in general. Twelve students were coded as belonging to this category. The remaining students gave various reasons as to why they felt motivated/not motivated to do their best in TIMSS.

The interviews

Students' motivation to do their best on the test

During the interviews, all students were asked how motivated they were to do their best on the TIMSS test, before and during it. Most interviewed students said they had been fairly motivated to do their best on the TIMSS test. Most students said their level of motivation was approximately the same during the time they took the test. A couple of students maintained they got even more motivated once they started to work with the mathematics and science tasks. However, a number of students reported that their motivation to do their best decreased during the test.

Value of a good performance and amount of effort invested

When the students were asked if a good performance in TIMSS was important to them and why, a majority of them reported that they thought TIMSS was a fairly important study and that it was quite important for them to do their best, perhaps not always for their own sake, but more for reasons external to themselves. In accordance with the answers to the open-ended item, several of the respondents gave comparative/competitive reasons for why they felt that a good performance in TIMSS was important:

- *It felt rather important to kind of show what we can.*
- *You don't want Sweden to look stupid.*

However, a few students did not think that TIMSS or a good performance in TIMSS was that important because of the low stakes of the test:

- *It didn't feel as important, because it wasn't—it didn't count for—the grades.*

The students were also asked how much effort they invested when answering the mathematics and science items in TIMSS. The typical answer was that they tried the best they could, because there was little point in not trying. Many students also reported that the amount of effort they invested shifted from item to item. Some students exerted more effort on items they found interesting and thought they had a chance on, while a few students exerted more effort on items they found difficult. Again, a number of students gave comparative/competitive reasons for why they tried hard to do their best, while others put forward social responsibility reasons:

- *I wanted to do the best I could because the investigation gets wrong if not everyone tries their best.*

TIMSS test compared to other tests in school

In connection with the discussion about the value of a good performance, the students were also asked to compare the TIMSS test to their regular school tests. Here, a number of students said that, although they tried to do their best on TIMSS, they were more motivated to perform well on their regular tests and to exert more effort on these tests:

- *I think that you try harder on a regular test because then you are to be graded and then you want to do a better result.*

The students seemed to be well aware of the fact that the TIMSS test did not have any consequences for them personally. Most of them still claimed that they tried their best, however. One student even said he was more motivated to do well and exerted more effort on the TIMSS test than on the regular school tests because "It's so big, like, Sweden in the world." Other students said that they worked as hard on the TIMSS test as they would have on any other test and that it did not matter that the test did not count for their grades. Rather, they felt no pressure and could work in a more relaxed manner:

- *It wasn't like a test; it was like an investigation, and therefore you weren't under stress and so on.*

Summary and concluding remarks

The main purpose of the work presented in this paper was to study aspects related to student test-taking motivation in the TIMSS 2003 context. The issue of test-taking motivation is an issue about validity and about the trustworthiness of test results. As such, it is an issue worthy of attention in the context of large-scale, comparative studies. Students' perceptions of TIMSS 2003 and their motivation to perform well in the study were investigated through questionnaire items and interviews, and the association between test-taking motivation and performance was explored through correlation and regression. Swedish Grade 8 students participating in TIMSS 2003 formed the study sample. From the obtained results, some conclusions can be drawn.

First, the questionnaire study as well as the interview study indicated that the Swedish students participating in TIMSS 2003 in general were well motivated to do their best on the TIMSS test and that they valued a good performance on the test.

Second, the Swedish mathematics result in TIMSS 2003 does not seem to have been affected by a lack of motivation among the participating students, as the students reported that they were well motivated to do their best. Also, ratings of test-taking motivation were positively but rather weakly related to performance.

Third, although the relationship between test-taking motivation and test performance was rather weak, the test-taking motivation scale still explained some of the variance in the result that could not be explained by the domain-specific motivational variables actually measured in TIMSS.

Findings from the present study thus indicate that a majority of the students valued a good performance and did not perceive TIMSS as a low-stakes test in the sense that it was unimportant and not worth spending effort on. It seems that the fact that the test does not have any personal consequences for the individual does not preclude the individual from attaching value to a good test performance. Students might be intrinsically motivated to do their best, even when there is no feedback on their performance. Students might also be extrinsically motivated by the fact that they are participating in a large international study where student populations are being compared. In their answers to the open-ended questionnaire item and the interviews, some students gave mainly intrinsic reasons for their motivation or lack of motivation, while others gave mainly extrinsic utility reasons. Many students put forward comparative/competitive reasons or social responsibility reasons as the reason for their motivation. Some students reported the low stakes of the test as detrimental to their motivation, but these students constituted a minority of the total sample.

It also seems that most students do seem to care about how they perform in studies like TIMSS, but that they nevertheless are aware of the fact that the test result does not count for them personally. Even though most of the students claimed they had been well motivated to do their best on the TIMSS test and that they tried to do their best, some of these students still claimed they would have tried harder had the test been a regular one.

This present paper explores an issue that has largely been ignored in the literature: how motivated students are to do their best in low-stakes contexts. Test-taking motivation can be relevant to the validity of interpretation and use of test results, and ignoring the test-takers and their views is incompatible with modern conceptions of validity (Messick, 1989). If test-taking motivation does contaminate responses to tests, this is an example of construct-irrelevant variance that affects the validity of test score interpretation (Benson, 1998; Messick, 1989). It follows that students' reactions to tests and their task-specific motivation to do well should be acknowledged in the interpretation and use of test scores, including those for TIMSS.

An obvious limitation of the present paper is that the study includes only Swedish TIMSS participants and that is not possible to study potential bias in the international comparisons due to varying levels of test-taking motivation. Nothing is known about student test-taking motivation in other countries participating in TIMSS and in other large-scale international studies. It is possible that level of test-taking motivation differs between countries and cultures. If so, cross-country comparisons of test-taking motivation and of the effect of test-taking motivation on test achievement constitute an urgent area of future research, especially in terms of strengthening the validity of interpretation of results from studies like TIMSS. Other systematic group differences in test-taking motivation, such as gender differences and differences between ethnic and social groups in national and international contexts, are also worthy of systematic investigation.

It should be noted that the instrument used in the present study is somewhat tentative and needs continued development and continued validation. Nevertheless, the obtained results imply that even a short measure of test-taking motivation can provide important information, and it would be possible to include a measure of student effort, performance expectancies, perceived importance of a good performance, or level of motivation in the TIMSS test battery. Including such a measure could contribute to the understanding of score meaning, and to the validity of score-based inferences.

The present paper also illustrates the possibility of adapting or adding national options to the TIMSS test battery. Each participating country has the possibility of adding national options to the questionnaires, and more nations, or possibly collaborating clusters of nations, should take advantage of this possibility. All nations participating in TIMSS have unique characteristics; adding national options to the anchor instruments administered by TIMSS could mirror these unique characteristics and enable large-scale investigation of questions of particular interest.

References

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*, 441–462.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice, 17*, 10–17.

Center for Educational Testing and Evaluation. (2001). *Student test taking motivation and performance: Grade 10 mathematics and science and grade 11 social studies* (research report). Lawrence, KS: School of Education, University of Kansas.

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82*, 300–310.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109–132.

Eklöf, H. (in press). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing.*

Eklöf, H. (2006a). Development and validation of scores from an instrument measuring student test-taking motivation. *Educational and Psychological Measurement, 66*, 643–656.

Eklöf, H. (2006b). *Motivational beliefs in the TIMSS 2003 context: Theory, measurement and relation to test performance.* Doctoral dissertation, Department of Educational Measurement, Umeå University, Umeå.

Martin, M. O., Mullis, I. V. S., & Chrostowski, S. J. (2004). *TIMSS 2003 technical report.* Chestnut Hill, MA: Boston College.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33–46). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (Vol. 3, pp. 13–103). New York: Macmillan/American Educational Research Association.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Mislevy, R. J. (1995). What can we learn from international assessments? *Educational Evaluation and Policy Analysis, 4*, 419–437.

Nevo, B., & Jäger, R. S. (1993). *Educational and psychological testing: The test taker's outlook.* Stuttgart: Hogrefe & Huber Publishers.

O'Leary, M. (2002). Stability of country rankings across item formats in the Third International Mathematics and Science Study. *Educational Measurement: Issues and Practice, 21*, 27–38.

O'Neil, H. F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment, 10*, 185–208.

O'Neil, H. F., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1997). *Final report of experimental studies on motivation and NAEP test performance* (CSE Technical Report 427). Los Angeles, CA: University of California, CRESST.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*, 33–40.

Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications* (2nd ed.). Englewood Cliffs, NJ: Merrill Prentice Hall.

Robitaille, D. F., & Garden, R. A. (1996). *Research questions and study design.* TIMSS Monograph No. 2, Vancouver: Pacific Educational Press.

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29*, 6–26.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement, 30*, 1–21.

Wigfield, A., & Eccles, J. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield & J. Eccles (Eds.), *Development of achievement motivation* (pp. 92–120). New York, NY: Academic Press.

Wise, S. L., & DeMars, C. E. (2003). *Examinee motivation in low-stakes assessment: Problems and potential solutions.* Paper presented at the annual meeting of the American Association of Higher Education Assessment Conference, Seattle, WA.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*, 227–242.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education, 8*, 341–351.

Zeidner, M. (1993). Essay versus multiple-choice type classroom exams: The student's perspective. In B. Nevo & R. S. Jäger (Eds.), *Educational and psychological testing: The test taker's outlook* (pp. 85–111). Stuttgart: Hogrefe & Huber Publishers.

Grade 4 Norwegian students' understanding of reproduction and inheritance[1]

Jorun Nyléhn
Department of Teacher Education and School Development
University of Oslo, Norway

Abstract

This study presents an analysis of two open-ended and two multiple-choice items from the TIMSS 2006 field test and the TIMSS 2003 main survey. The items relate to biology and concern aspects of reproduction and inheritance. The open-ended questions are classified in detail in order to investigate students' understanding and to reveal their preconceptions of these topics. The students' answers varied considerably, from those that accorded with school science to those seemingly without any scientific reasoning. Many of the "wrong" answers revealed interesting understandings of the topic in question, although with plain and inaccurate attempts at explanation. Such answers might be useful in classroom situations if the teacher is able to sort out the relationships in the answers to biological theory.

Introduction

All content in biology could be simplified to aspects of survival and reproduction. Although biology is fundamental to life, this consideration does not imply the subject has equal importance to other subject matters in school science.

Reproduction is part of the Norwegian curriculum from Grade 3, when a simplified version of the human life cycle is introduced along with a few other lifecycles (KUF, 1996). More details concerning reproduction are not given until Grades 5 and 6. Thus, students' answers to questions in assessments on reproduction and inheritance in Grade 4 might reveal their understandings (i.e., their existing, everyday knowledge) of these areas before they receive further formal instruction in them.

Many researchers point to the robust body of literature showing that students develop ideas about "nature" before receiving school instruction in topics related to it (see, for example, Morrison & Lederman, 2003; Palmer, 1999). The underlying concern in this kind of research is to gain a better understanding of how students acquire new knowledge, and thus how to teach them in the best way (see, for example, Warren, Ballenger, Ogonowski, Rosebery, & Hudicourt-Barnes, 2001). Students interpret what they learn in school in terms of their existing ideas, which they may in turn modify in light of the newly acquired knowledge. This "constructivist view" is the dominant paradigm of learning in science (Driver, Asoko, Leach, Mortimer, & Scott, 1994).

How students interpret the unknown in terms of the known can be exemplified through the notion of giving human characteristics to other organisms, or offering cultural practices as scientific explanations. Whether students' everyday understanding agrees with the scientific worldview is highly variable (Helm & Novak, 1983; Novak, 1987). Surveys of students' conceptions and misconceptions relating to science topics are therefore important. The research on students' preconceptions often shows that these differ considerably from the accepted scientific viewpoints (Palmer, 1999). Many surveys of students' conceptions have been conducted in relation to topics in physics (see, for example, Angell, 2004; Liu & McKeough, 2005). Studies on topics in biology have also been conducted, although to a much lesser extent than in physics (Bishop & Anderson, 1990; Dagher & Boujaoude, 2005; Hagman, Olander, & Wallin, 2001; Sinatra, Southerland, McConaughy, & Demastes, 2003; see also Tanner & Allen, 2005). Duit (2006) provides a comprehensive bibliography of studies on students' and teachers' conceptions of science topics.

[1] The author sincerely thanks Svein Lie, Rolf Vegar Olsen, and Inger S. Throndsen for comments on drafts of this paper.

Identification of students' "misconceptions" or "alternative conceptions" is especially important whenever students learn by linking new information to their existing worldview. Thus, science teachers should be able to identify students' scientifically acceptable conceptions as well as their alternative or scientifically wrong conceptions (Morrison & Lederman, 2003; Palmer, 1999; Tanner & Allen, 2005). But to what degree do students' preconceptions form a base for learning science or present obstacles? Some authors point to a large gap between everyday thinking and scientific viewpoints (Aikenhead, 1996; Cobern, 1996). Others regard the everyday sense-making of students as a useful basis for learning (Clement, 1989). In this tradition, the relationship between students' and scientists' worldviews is assumed to be complex, ranging from similarity to difference, from complementary or just generalizations (Warren et al., 2001). Authors in this tradition regard some of students' preconceptions from daily life as useful "starting points" for understanding scientific topics or theories (Brown, 1997).

These considerations informed the present study, which investigated students' answers to questions concerning reproduction and inheritance in Grade 4, and classified the answers according to correspondence with the scientific worldview. The study also endeavored to determine if the students' ideas could serve as a basis for further learning, and to provide some reasons for any non-scientific beliefs evident in the students' answers.

Method

Materials

The Trends in Mathematics and Science Study (TIMSS) contains a considerable number of open-ended and multiple-choice questions. Several thousand students answered this test in Norway, and their answers provide a unique data set relating to Norwegian students' achievements in and understanding of science and mathematics topics. Two open-ended biology items given in the field test in TIMSS 2006 are analyzed in detail in the present study. These items concern aspects of reproductive biology in Grade 4 and are not included in the TIMSS 2007 main study. Two multiple-choice items from TIMSS 2003 concerning (genetic) inheritance are also analyzed.

The students' answers on these items were categorized in terms of being "right" or "wrong," and the "wrong" answers were further separated, based on the presence of relationships with biological scientific theories. Common answers that seemed to have no relation to science were also given categories. Note that in presenting the students' answers, I have tried to transfer the misspellings in them from Norwegian to English.

Findings

Items on reproductive biology

Rats and mice
Item no. S041170, 2006 field test (deleted from the main test in 2007):
Can a rat and a mouse mate and produce offspring?
(Check one box.)
□ *Yes*
□ *No*
Explain your answer.

As stated in the field test scoring guide, a correct answer required students to check "No" and to explain that the rat and the mouse are different species, different types of rodents, and/or different kinds of animals. Answers like "Yes, they are in the same family" were classified incorrect. In the present study, a far more detailed categorization was accomplished in order to investigate the students' knowledge and understanding of hybridization and mating between closely related taxa.

All together, 225 students answered the booklet containing this question. Two students did not answer the question (they left the answer space blank) and 63 made only what can be best described as "stray" marks in the answer space. In addition, 26 answers were not classified (these included misunderstandings of the question and answers like "Because a plant can't move" or "No, they are not alive"). The remaining 134 answers were classified according to the criteria given in Table 1. The percentages of answers within each category are given in Table 2, calculated from the total number of participating students, including the two who did not answer the question. Figure 1 gives the proportion of answers within each category. Note that although the answers in Category F were not necessarily the result of thinking from a human viewpoint, and the

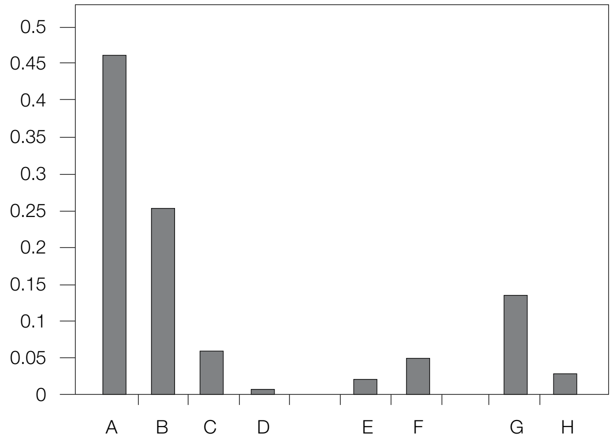
Table 1: Classification of Students' Answers to Item No. S041170: Can a rat and a mouse mate and produce offspring?

	Kind of explanation
	Biological explanations
A	No. Refers to different species, types, and/or kinds of animals. ("No, because it is not the same species.")
B	Yes. Refers to close relationship between the animals. ("Yes, because they are in the same animal family" or "Yes, because they are almost totally the same animal.")
C	No. Refers to unsuccessful hybrid offspring. ("No, then it could be too strange offspring.")
D	No. Refers to habitat differences. ("No, because the rat lives in the sewer.")
	Human viewpoint
E	Yes. Refers to marriage or dating before conceiving offspring. ("Yes, because they can marry.")
F	Yes. Refers to rats being males and mice being females. ("Yes, because mouse is a woman and the rat is man.")
	Other common explanations
G	No. Refers to rats being larger than mice. ("No, because the rat is larger.")
H	No. Refers to rats eating mice. ("No, I think the rat can eatt the mouse.")

answers in Category G might be seen as the result of some biological reasoning (such as student thinking relating to morphology), I have treated the categories separately in the discussion below.

Twenty-eight percent of the students gave answers that fell into Category A, that is, they chose "No" and referred to different species or kinds of animals. Another common answer was "Yes" in terms of rats and mice being closely related, chosen by 15% of the students. Some students (4%, Category C) chose "No" and explained that the hybrid offspring would be too strange or did not exist in nature (Category C). Only one student gave an answer related to habitat differences (Category D). Some students chose "Yes," explaining that dating or marriage was the necessary action before producing offspring (1%, Category E). Three percent of the students believed that rats are males and mice are females (Category F). A fairly common kind of answer was "No," with reference to rats being larger than mice (8%, Category G); 2% of the students believed that rats eat mice and answered "No" (Category H).

Figure 1: Proportions of Answers in Each Category for Item No. S041170, Calculated from the Number of Classified Answers



Note: Explanations of the categories are given in Table 1.

Table 2: Numbers and Percentages of Answers in Each of the Categories for Item No. S041170

	Number of answers	Percentage
A	62	28
B	34	15
C	8	4
D	1	0.4
E	3	1
F	7	3
G	18	8
H	4	2

Note: The percentages are calculated from the total number of participating students, and they include students who did not answer the question. Explanations of the categories are given in Table 1.

Animals in plant reproduction
Item no. S041021, 2006 field test (deleted from the main test in 2007):
Name an animal that plays a part in plant reproduction.
Describe what the animal does.

The field test scoring guide stipulated that a fully correct answer was one in which students named an animal and described what the animal did to aid pollination or dispersal. The answers, including the wrong ones, were again categorized in this present study in order to investigate the students' knowledge and understanding of animals that contribute to plant reproduction.

One hundred and eighty-eight students answered the booklet containing this question. Of these, 69 answers were blank and two had only stray marks. Thus, only 117 pupils answered the question. In addition, 36 answers were not classified (these included misunderstandings of the question and answers like "A large animal brown with horns" or "Ant. Have no idea"). The remaining 81 answers were classified according to the criteria given in Table 3. The percentages of answers (including the blank answers) within each category are given in Table 4. The percentages are calculated from the total number of participating students.

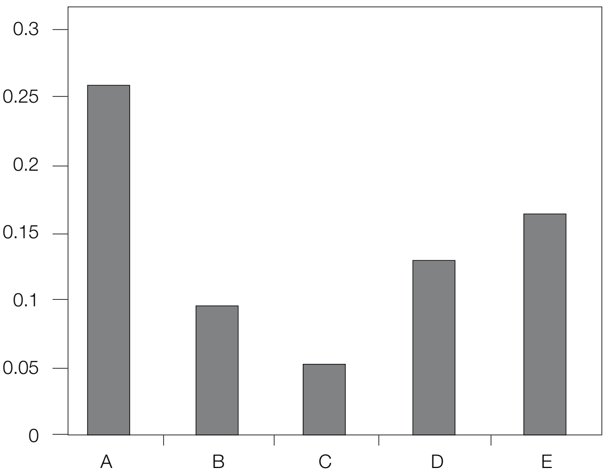
Sixteen percent of the students gave answers related to pollination (Category A in Tables 3 and 4 and in Figure 2). Answers relating to soil or manure were relatively common (6%, Category B). Three percent of the students answered "earthworm" and explained that worms dig tunnels in the ground for plant roots

Table 4: Numbers and Percentages of Answers in Each of the Categories for Item No. S041021

Number of answers		Percentage
A	62	28
A	30	16
B	11	6
C	6	3
D	15	8
E	19	10

Note: The percentages are calculated from the total number of participating students, and they include students who did not answer the question. Explanations of the categories are given in Table 3.

Figure 2: Proportions of Answers in Each Category for Item No. S041021, Calculated from the Number of Classified Answers



Note: Explanations of the categories are given in Table 3.

Table 3: Classification of Students' Answers to Item No. S041021: Name an animal that plays a part in plant reproduction. Describe what the animal does.

Kind of explanation	
A	Refers to pollination or taking nectar. ("The bees. When they shall suck nectar from a flower they get pollen on themselves which fall off a little by little.")
B	Refers to animals making the soil better for growth or making manure. ("Horse. It eats a lot of strange things like dry plants etc. Then it shits. It can be used as good manure.")
C	Refers to animals (especially earthworms) digging tunnels for the plant roots or to aid water trickling through the ground. ("Earthworm. It makes holes where wate can sep through" or "Digs tunnels.")
D	Refers to animals eating plants. ("The snail can eat the flower.")
E	Refers to animals doing something else with flowers, like camouflaging themselves or dancing around the plants. ("Tiger. It plays with the flower. It runs around the flower.")

or water (Category C). Eight percent mentioned animals eating plants (Category D; here, the students did not mention seeds and were not referring to dispersal). Some students, 10%, mentioned animals doing something else with plants, like camouflaging themselves (Category E).

Items on inheritance

Yellow flowers
Item no. S031269, 2003 main survey (released):
A plant has yellow flowers. What best explains why the flowers are yellow?
A The sunshine colored the flowers yellow.
B The flowers of the parent plants were yellow.
C It was very warm when they flowered.
D It rained every day.

The correct answer is B—the flowers of the parent plants were yellow. All together, 677 Norwegian students answered this question. Of these, 23.1% chose A, 55.9% chose B, 14.0% C, and 4.0% D. The international average frequencies were 18.5% for A, 53.5% for B, 18.4% for C, and 6.2% for D.

Adult height
Item no. S031269, 2003 main survey (released):
What will most likely affect your adult height?
A The height of your parents
B The height of your brothers and sisters
C Your hair color
D Your weight

The correct answer is A—the height of your parents. Seven hundred and eleven Norwegian students answered this item. Of these, 62.0% chose A, 5.2% chose B, 1.2% C, and 29.3% D. The international average frequencies were 42.1% for A, 8.0% for B, 4.9 % for C, and 42.4% for D.

Discussion

Rats and mice item

Students were asked whether a rat and a mouse can mate and produce offspring. In most cases, different species cannot breed and produce successful and fertile offspring unless the result is a new species (hybridization). This is actually the definition of a species, according to the biological species concept. Twenty-eight percent of the Norwegian students' answers corresponded with this concept (Category A in Tables 1 and 2 and in Figure 1).

Interestingly, 15% of the students answered that rats and mice can produce offspring, with their explanations referring to the close relationship between these animals (Category B in Tables 1 and 2). These students certainly had knowledge about the species in question, and referred to them as mammals or as being part of the same family. Although the theory (the biological species concept) sets the limit for the ability to produce fertile offspring at the species level, there are not so definite limits in nature. In general, the chance of hybridization increases because the parent species are more closely related.

Furthermore, the "successful and fertile offspring" in the biological species concept is not the same as "offspring" solely. There are well-known examples of crosses between different species resulting in offspring, like the sterile mule (parents are horse and donkey). Some students (4%, Category C in Tables 1 and 2) referred to the possible offspring being "too strange" or that such intermediates between rats and mice do not exist in nature. One student described a possible offspring as "being as little as a mouse with a tail like a rat."

A problem with the biological species concept is that the limits between species have to be tested experimentally. We do not know from the present distribution of species whether they are able to reproduce with one another or not. For example, geographically isolated groups might belong to the same species even though they look somewhat different. There is a theoretical possibility that some species of rats and mice might produce fertile offspring, if brought together. Habitat differences also contribute to species isolation. One student referred to this: "No, because the rat lives in the sewer" (Category D in Tables 1 and 2).

Also, there are many organisms that definitely do not fit into the biological species concept. A striking example is asexual organisms. Biologists have different species concepts, which are more or less suitable for the various species groups. Sometimes the limit between two species is clear-cut in one part of the world and blurry elsewhere. There might be gradual differences within a species along a geographical gradient, whereas the individuals in the endpoints belong to different species according to the biological species concept. A well-known example is that of two species of gulls in Northern Europe that are related through a gradient of intermediates crossing Siberia, Alaska, and northern Canada.

In scientific language, it is more specific to state that an organism belongs to the same genus than to the same family. In Norwegian, the word for family (*familie*) means a closer relationship than does the word for genus (*slekt*) in non-scientific connections. These shades of meaning may have confused the Norwegian students.

Some students' interpretations of the animals rested on human behavior. Approximately 1% answered that the rat and the mouse can have children if they first marry (Category E in Tables 1 and 2). In Norwegian, the verb "can" also has the connotation of "being allowed to." Some of these students may have been invoking moral or religious justifications in mentioning the need for marriage before having children. Another example of interpretation from a human viewpoint is "Yes. They can met each otther They can eat together and become sweethearts and mate."

Another common type of answer is exemplified by "Yes. A rat is a man and a mouse is a lady and hav they children together." Three percent of the students gave answers that were variants of this notion (Category F in Tables 1 and 2). The underlying reason seems unclear. Did the students associate rats with men and mice with women based on what they look like, that is, masculine-looking rats and feminine-looking mice? Mice are usually perceived as cute, while sewer-living rats do not have similar positive associations. Is the idea caused by the size differences of the animals? Answers relating to size differences were frequent (8% of the students, category G in Tables 1 and 2).

The last category of answers suggested that rats eat mice (2%, Category H in Tables 1 and 2). People stay away from wild rats living in the sewer, and are afraid rats might bite or spread diseases. Mice can also spread diseases, but this fact is not so well known. Rats are also known as being able to eat "everything." It is therefore not surprising that students think rats are a danger to mice.

Animals in plant reproduction item

This item asked students to identify an animal that takes part in plant reproduction, and to describe what the animal does. Animals contribute to two processes in plant reproduction—pollination and dispersal. Few Norwegian students gave answers that could be construed as relating to dispersal, and these were not classified in the present study. The few answers with some relevance were "Horse. It eats flowers and defecate it out another place" and "The shiraf [giraffe]. The animals shake it then it will reproduce," but it is doubtful that these students were actually thinking of dispersal.

The students seemed to have a better understanding of pollination. Pollination might also be the first consideration students associate with the notion of animals contributing to plant reproduction. Thus, they might have answered dispersal as well if asked for two kinds of interactions. Overall, 16% of the answers related to pollen, nectar, or honey and naming bees, wasps, or bumblebees as the transfer agents (Category A in Tables 3 and 4 and in Figure 2). These answers were sometimes simple ("Bumblebee. Carries honey") or somewhat more elaborate ("The bumblebee takes pollon from flower to flower" or "The bees. When they shall suck nectar from a flower they get pollen on themselves which fall off a little by little"). Other pollination agents, for example, flies, were totally absent in the material. A few students mentioned butterflies, but as they did not give any explanation, their answers were not included in this category.

Some students seemed to confuse reproduction with growth. They referred to animals making the soil better for growth or making manure (6%, Category B in Tables 3 and 4). Examples were "Horse. Theirs dung is used as manur" and "The hen. It makes manure." This category of answers relates to the next, in which 3% of the students said that earthworms contribute to improved reproduction by digging tunnels, so helping the plant to grow (Category C in Tables 3 and 4). Other examples included "Animales makes breathing holes in the earth (earthworms)" and "Earthworm. The earthworm digs down in the earth and cleans it."

Plant reproduction is not always mediated through seeds. Vegetative propagation through stolons or rhizomes is an important mode of reproduction in a great many plant species. Thus, plant growth might contribute directly to reproduction: when a plant splits in two, the result may be two functional individuals. This is how many plants in the garden (e.g., strawberries and potted plants) reproduce, and students might have seen this occurring at home. Because the borderline between plant growth and reproduction is unclear in "real life," we should not be surprised that students also find this notion unclear.

Also, "growth and reproduction" is a common, set phrase, as is "contribute to growth and reproduction." In the Norwegian translation, the word used for reproduction (*formering*) is one not commonly used by nine-year-olds, and lack of knowledge of this word clearly would have made it more difficult for the students to figure out what they should answer. However, the other word that might have been used—*reproduksjon*—is even more unfamiliar. *Reproduksjon* is a direct translation of "reproduction."

Another common kind of answer rested on variants of the notion that animals eat plants: "Frog. It takes the tongue out and eats it" and "Roe deer. It eats bark on the treens. And eats tulips" (8%, Category D in Tables 3 and 4). These students at least referred to a plant–animal interaction, even though this did not contribute to reproduction (i.e., the answers did equate eating seeds with dispersal).

The last category contains rather different kinds of animal actions, which have nothing to do with reproduction, growth, or eating (10%, Category E in Tables 3 and 4). Examples are "Bumblebee. They smear something on the berries so they will get large," "Earthworm. It keeps insect away from the flower," and "Cat. The cat likes to go out in nice weather and pass along flowers and likes to play." Some students mentioned hedgehogs, tigers, and elephants—answers that seem to signify guessing rather than reasoning. Also, the students might simply have selected animals they found fascinating.

Inheritance items

The first of the two multiple-choice items relating to heritage asked students for the best explanation of why flowers are yellow. The correct answer was "The flowers of the parent plants were yellow." All together, 55.9% of the Norwegian students chose this answer, close to the international average of 53.5%. The alternative—"The sunshine colored the flowers yellow"—was also popular, being chosen by 23.1% of the Norwegian students (international average, 18.5%). The yellow color is indeed partly caused by the sunshine (light), with the yellow-colored wavelengths being reflected. The other two alternative answers attracted relatively few students. Fourteen percent chose "It was very warm when they flowered" and 4% chose "It rained every day" (international averages, 18.4% and 6.2%, respectively).

The second multiple-choice item asked the students to choose among variables that affect adult height. The most frequently chosen alternative was the correct one, "The height of your parents." Sixty-two percent of the Norwegian students selected this answer. The international average was 42.1%. Two alternatives were seldom checked, either by the Norwegian students or by the international cohort. These items were "The height of your brothers and sisters" and "Your hair color." However, many students selected the alternative "Your weight" (29.3% in Norway and 42.4 % internationally). While weight is, on average, correlated with height, it seems that a misunderstanding was behind the students' decision to choose the answer that weight affects height.

Concluding remarks

The open-ended questions considered in this present analysis contained scientific ambiguities, and so were excluded from the main TIMSS survey. But ambiguities such as these can contribute positively by increasing our understanding of the reasons behind students' often richly varied answers. This was the premise behind the study reported here.

What can we learn from the collection of answers documented in this paper? As mentioned in the introduction, two traditions are evident in the research on students' preconceptions of topics they study at school (e.g., Warren et al., 2001). Should students' misunderstandings be erased and replaced by correct theory, or rather used as a resource in the learning of science? The detailed analyses of selected items in the present study reveal a range in the "wrong" answers, from scientifically meaningful reasoning to misunderstandings. Clearly, a teacher could use many of these answers in classroom situations. Tanner and

Allen (2005), for example, regard an understanding of students' wrong answers as a useful resource in science instruction.

In the "rat and mouse" question, the answer Categories A to D could be a stepping stone to far more interesting and fascinating considerations of biology than that for Category A, which is strictly linked to the definition of species in the biological species concept. Categories B and C could be similarly used in the "animals in plant reproduction" question. Of course, many of the students' answers seemingly had no connection to science and scientific thinking. In general, the extent to which students' answers are fruitful for classroom discussions will differ, but answers that seem to be wrong compared to the key might nonetheless contain scientifically correct ideas.

As also mentioned in the introduction, this consideration relies on science teachers' ability to recognize the meaningful parts of the students' often plain and inaccurate attempts at explanations (Morrison & Lederman, 2003; Palmer, 1999). Teachers need to have a deeper understanding than that denoted merely by the contents of the curriculum at the given level. In Norwegian schools, however, teachers at Grade 4 level usually lack specific education in science after obligatory schooling (Grønmo, Bergem, Kjærnsli, Lie, & Turmo, 2004). Teachers with poor knowledge of science may also rely heavily on the textbook, curriculum, and key, or they might confuse scientific theories themselves.

Despite these concerns, telling students that their ideas are useful is probably more motivating for them and their learning than telling them that they are wrong. Such an approach, however, should not be construed as telling them that all their ideas are scientifically meaningful. Although some "wrong answers" might be useful in helping students refine their scientific understanding, other preconceptions that impede their comprehension of science and should be replaced (Brown, 1997).

References

Aikenhead, G. (1996). Border crossings into the subculture of science. *Studies in Science Education, 27*, 1–52.

Angell, C. (2004). Exploring students' intuitive ideas based on physics items in TIMSS 1995. In C. Papanastasiou (Ed.), *Proceedings of the IEA International Research Conference 2004*, Cyprus. Nicosia: Cyprus University Press.

Bishop, B., & Anderson, C. (1990). Student conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching, 27*, 415–427.

Brown, A. (1997). Transforming schools into communities of thinking and learning about serious matters. *American Psychologist, 52*, 399–413.

Clement, J. (1989). Not all preconceptions are misconceptions: Finding "anchoring conceptions" for grounding instruction on students' intuitions. *International Journal of Science Education, 11*, 554–565.

Cobern, B. (1996). Worldview theory and conceptual change in science education. *Science Education, 80*, 579–610.

Dagher, Z. R., & Boujaoude, S. (2005). Students' perceptions of the nature of evolutionary theory. *Science Education, 89*, 378–391.

Driver, R., Asoko, H., Leach, J., Mortimer, E., & Scott, P. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher, 23*, 5–12.

Duit, R. (2006). *Bibliography: Students' and teachers' conceptions and science education database.* Kiel: University of Kiel. Available on http://www.ipn.uni-kiel.de/aktuellstcse/stcse.html (most recent version, February 2006).

Grønmo, L. S., Bergem, O. K., Kjærnsli, M., Lie, S., & Turmo, A. (2004). Hva i all verden har skjedd i realfagene? Norske elevers prestasjoner i matematikk og naturfag i TIMSS 2003 [What on earth has happened in the natural sciences? Norwegian students' achievements in mathematics and science in TIMSS 2003]. *Acta Didactica* (Issue 5). Oslo: Department of Teacher Education and School Development, University of Oslo.

Hagman, M., Olander, C., & Wallin, A. (2001). *Teaching and learning about biological evolution: A preliminary teaching-learning sequence.* Paper presented at the Third International Conference on Science Education Research in the Knowledge Based Society (ESERA), Thessaloniki, Greece.

Helm, H., & Novak, J. D. (Eds.). (1983). *Proceedings of the international seminar "Misconceptions in Science and Mathematics," June 20–22, 1983.* Ithaca, NY: Cornell University.

KUF (Kirke-, utdannings- og forskningsdepartementet) [Ministry of Church Affairs]. (1996). *Læreplanverket for den 10-årige grunnskolen [The curriculum for the 10-year compulsory school].* Oslo: Author.

Liu, X., & McKeough, A. (2005). Developmental growth in students' concept of energy: Analysis of selected items from the TIMSS database. *Journal of Research in Science Teaching, 42*, 493–517.

Morrison, J. A., & Lederman, N. G. (2003). Science teachers' diagnosis and understanding of students' preconceptions. *Science Education, 87*, 849–867.

Novak, J. D. (1987). *Proceedings of the second international seminar "Misconceptions and Educational Strategies in Science and Mathematics," July 26–29, 1987.* Ithaca, NY: Cornell University.

Palmer, D. H. (1999). Exploring the link between students' scientific and nonscientific conceptions. *Science Education, 83*, 639–653.

Sinatra, G. M., Southerland, S. A., McConaughy, F., & Demastes, J. W. (2003). Intentions and beliefs in students' understanding and acceptance of biological evolution. *Journal of Research in Science Teaching, 40*, 510–528.

Tanner, K., & Allen, D. (2005). Approaches to biology teaching and learning: Understanding the wrong answers—teaching toward conceptual change. *Cell Biology Education, 4*, 112–117.

Warren, B., Ballenger, C., Ogonowski, M., Rosebery, A. S., & Hudicourt-Barnes, J. (2001). Rethinking diversity in learning science: The logic of everyday sense-making. *Journal of Research in Science Teaching, 38*, 529–552.

Examining the problem-solving achievement of Grade 8 students in TIMSS 2003

Alka Arora
TIMSS and PIRLS International Study Center, Boston College
Chestnut Hill, Massachusetts, USA

Abstract

Using data from the Trends in Mathematics and Science Study (TIMSS) 2003, this paper examines the problem-solving achievement of Grade 8 students in the countries that participated in the study. Data from TIMSS 2003 were used to create a new problem-solving scale, primarily based on the special, multi-part problem-solving tasks in mathematics and inquiry tasks in science. The work builds on a development project recently completed by the TIMSS and PIRLS International Study Center to create scales for the TIMSS 2003 mathematics cognitive domains. This study capitalizes on the TIMSS design, whereby students respond to both mathematics and science items in the same test booklet. The problem-solving achievement results for the Grade 8 students overall and by gender are compared to their mathematics and science achievement generally.

Introduction

The TIMSS 2003 data provide an as yet untapped and unique opportunity to examine the problem-solving achievement of students in the countries that participated in the study. As an initiative supported by the United States National Science Foundation (NSF), TIMSS 2003 included a special development effort to assess problem solving and inquiry. Specifically, at the eighth grade, TIMSS 2003 included four extended multi-part tasks assessing problem solving in mathematics and three such tasks assessing inquiry in science. Also, with support from several participating countries, TIMSS 2003 recently completed a development project designed to create valid and reliable scales for the mathematics cognitive domains.

This study combines information from the TIMSS 2003 mathematics and science assessments to develop a new cross-cutting achievement scale that measures problem solving at the eighth grade. The overall methodological approach was to use what was learned from the development project to scale the TIMSS 2003 mathematics cognitive domains (Mullis, Martin, & Foy, 2005). All the TIMSS 2003 items for the eighth grade, including the special NSF-funded tasks that assess reasoning, problem solving, and inquiry in both mathematics and science, were examined. Based on the literature review and experts' judgment, a pool of items was selected that were considered to assess problem solving. The next step was to combine the response data from the selected items and tasks into a single scale to measure Grade 8 students' achievement in problem solving.

Once the scale was developed, the problem-solving performance of Grade 8 students from across the participating countries was described overall and by gender. It was also compared with overall achievement in mathematics and science.

Review of the literature

Problem solving is important in instruction because it provides valuable contexts for learning. More importantly, however, it enables students to be independent thinkers and to find solutions in all areas of life. Much research in the area of problem solving concerns the cognitive processes involved (Pellegrino, Chudowsky, & Glaser, 2001). For example, Wicklegren (1974) described the phases of the problem-solving process as clarifying the givens, identifying the needed operations, drawing inferences, and recognizing goals. According to current research and literature, students exhibit some key attributes when performing at a high level of problem solving. These include conceptual understanding, reasoning, communication, computation and execution, and insights (Case, 1985; Mayer, 1985; Resnick & Ford, 1981).

Because problem solving is such an important outcome of mathematics and science instruction, it has been measured in a number of assessments. However, there are many challenges in developing valid and reliable problem-solving assessments, including appropriate complexity of tasks, number of steps, interesting non-routine situations, and administrative feasibility (Speedie, Treffinger, & Feldhusen, 1971).

Importance of the study

TIMSS 2003 included a special set of tasks in the mathematics and science assessment devoted to having students do the best job possible in assessing problem solving and inquiry in an international setting. Mayer (1992), for example, noted that assessments should require students to find solutions to realistic, non-routine problems that require the invention of novel solution strategies. The tasks should extend into situations, drawing on prior knowledge and requiring the integration of concepts, representations, and processes on the part of the test-takers. With NSF support, TIMSS 2003 employed panels of international experts, extensive field testing, and cognitive laboratories to develop such tasks. However, the achievement results of these tasks have yet to be fully mined because TIMSS 2003 reported the problem-solving tasks as integrated into the overall mathematics results and because the inquiry tasks were reported as integrated into the overall science results.

Given the amount of effort and energy expended in TIMSS 2003 on collecting valuable information about problem-solving achievement in an international context, it is important to analyze and report the findings. This study uses TIMSS 2003 data to develop a robust problem-solving scale. It capitalizes on the measures in both the mathematics and science assessments. Even though students participating in TIMSS responded to both mathematics and science items, it is rare for research to take advantage of this situation. Typically, mathematics and science results are reported separately.

Methodology

This study used the TIMSS 2003 database for eighth-grade mathematics and science (Martin, 2005). The database included responses from nationally representative samples of students (approximately 360,000 in total) from 48 countries. In particular, the database used the student responses to the TIMSS 2003 mathematics and science items that assessed problem solving as it pertains to the higher-level cognitive processing strategies described in the literature.

The first step was to identify the set of items and tasks in both mathematics and science to form the basis of the scaling. Thirty-four science items and 42 mathematics items were included in the scaling. The next step was to construct the problem-solving scale, using the scaling methods routinely applied in TIMSS. TIMSS 2003 relies on item-response theory (IRT) scaling to describe student achievement. The TIMSS IRT scaling approach uses multiple imputation or plausible values methodology to obtain proficiency scores in mathematics and science for all students, even though, because of the rotated block/booklet design, each student responded to only part of the assessment item pool.

To enhance the reliability of the student assessment scores, the TIMSS scaling combined student responses to the administered items with information about students' backgrounds, a process known as "conditioning." Also, since TIMSS is based on representative samples of students from the participating countries, the use of sampling weights is an integral part of the analysis. TIMSS scaling methodology is described in detail in Gonzalez, Galia, and Li (2004).

Once the scale was developed, the problem-solving achievement results for the eighth-grade students, overall and by gender, were compared to their performance in TIMSS 2003 to determine if their relative strengths and weaknesses in problem solving compared to their mathematics and science achievement generally.

Results

Reliability of the problem-solving scale

Reliability was measured as the ratio of sampling variance to sampling variance plus imputation variance. This method was first used in TIMSS for the TIMSS 2003 cognitive domain report (Mullis, Martin, Gonzalez, & Chrostowski, 2004). This approach is considered better for multiple-matrix-sampling designs, where students respond to relatively few items, than classical reliability methods (such as the Kuder-Richardson formulas), which are affected by the number of items taken by the student (Johnson, Mislevy, & Thomas, 1994). A value of 0.80 for the reliability coefficient is generally considered acceptable for such designs.

Figure 1 presents the reliability coefficients for the TIMSS 2003 countries for the problem-solving scale. Despite some variation, the reliability was generally high for most of the countries. Of the 46 countries, 40 had reliabilities that were above 0.80. The international median (the median of the reliability coefficients for all countries) was 0.94.

Construct validity of the scale

The items comprising the problem-solving scale are a subset of the entire pool of TIMSS 2003 eighth-grade mathematics and science items. Items were selected for the problem-solving scale if they were judged to assess the kinds of skills described in the literature on problem solving. The TIMSS 2003 Science and Mathematics Item Review Committee (SMIRC) classified many of these items as belonging to the reasoning cognitive domain, which called for higher-level problem-solving skills.

In choosing items for the problem-solving scale, a preliminary selection was made by staff of the TIMSS and PIRLS international study center, assisted by the TIMSS mathematics and science coordinators. As a validity check, the members of the TIMSS 2003 SMIRC reviewed the selected items and suggested a number of adaptations. The members of SMIRC agreed that the resulting items were included in the problem-solving domain.

Problem solving is a crucial aspect of achievement in both mathematics and science. Accordingly, achievement in problem solving should correlate positively with achievement in these subjects. Figures 2 and 3 show the relationship between problem solving and mathematics and science, respectively. As would be expected for scales sharing some of the same items, the Pearson correlation is substantial in each case: 0.89 between problem solving and mathematics achievement and 0.86 between problem solving and science achievement. However, as can be seen in Figure 4, which shows the relationship between mathematics and science achievement, these correlations are about the same as the correlation between mathematics and science achievement, which is 0.86. Thus, because we see variations in performance between mathematics and science, we can expect to see differences in relative performance in problem solving for the TIMSS 2003 countries.

Figure 1: Reliabilities of the Problem-solving Scale

Countries	Reliabilities
Armenia	0.88
Australia	0.96
Bahrain	0.75
Belgium (Flemish)	0.99
Botswana	0.62
Bulgaria	0.94
Chile	0.91
Chinese Taipei	0.97
Cyprus	0.84
Egypt	0.93
England	0.96
Estonia	0.97
Ghana	0.94
Hong Kong SAR	0.96
Hungary	0.94
Indonesia	0.97
Iran, Islamic Rep. of	0.84
Israel	0.88
Italy	0.94
Japan	0.78
Jordan	0.96
Korea, Rep. of	0.94
Latvia	0.96
Lebanon	0.88
Lithuania	0.96
Macedonia, Rep. of	0.88
Malaysia	0.99
Moldova, Rep. of	0.96
Morocco	0.79
Netherlands	0.95
New Zealand	0.95
Norway	0.88
Palestinian Nat'l Auth.	0.86
Philippines	0.93
Romania	0.96
Russian Federation	0.98
Saudi Arabia	0.88
Scotland	0.95
Serbia	0.86
Singapore	0.98
Slovak Republic	0.91
Slovenia	0.92
South Africa	0.98
Sweden	0.98
Tunisia	0.56
United States	0.95
International Median	0.94

Figure 2: Relationship between Problem-solving and Mathematics Scores

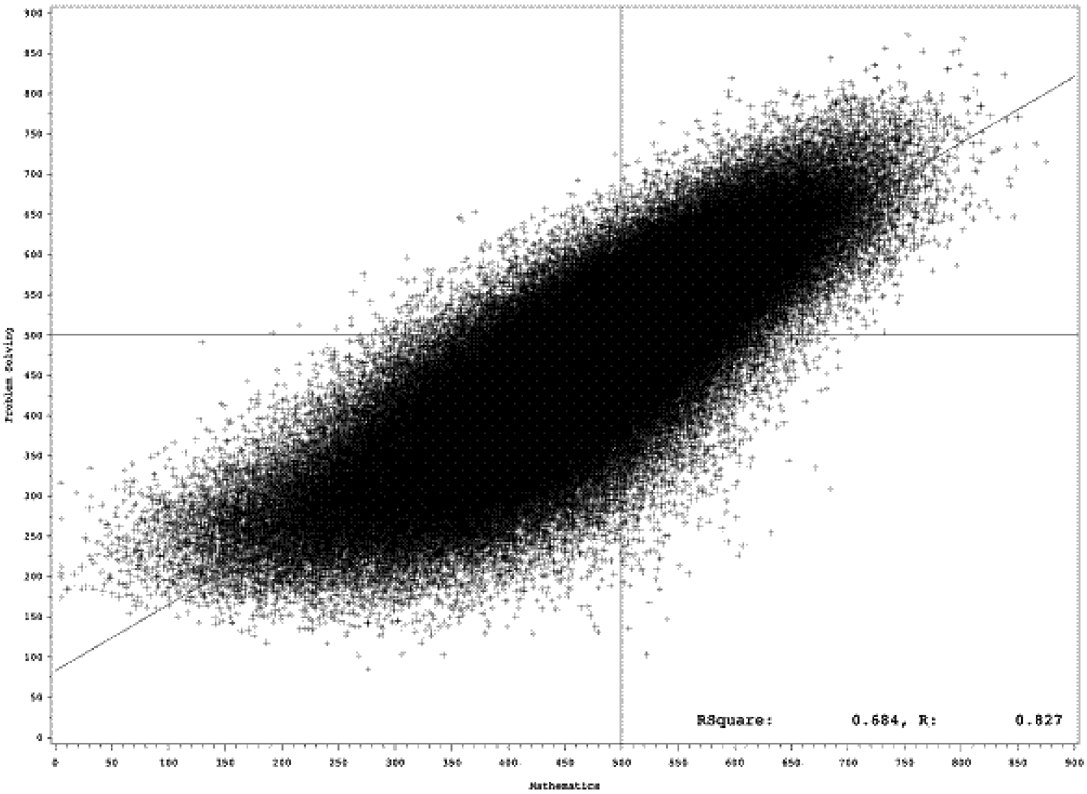


Figure 3: Relationship between Problem-solving and Science Scores

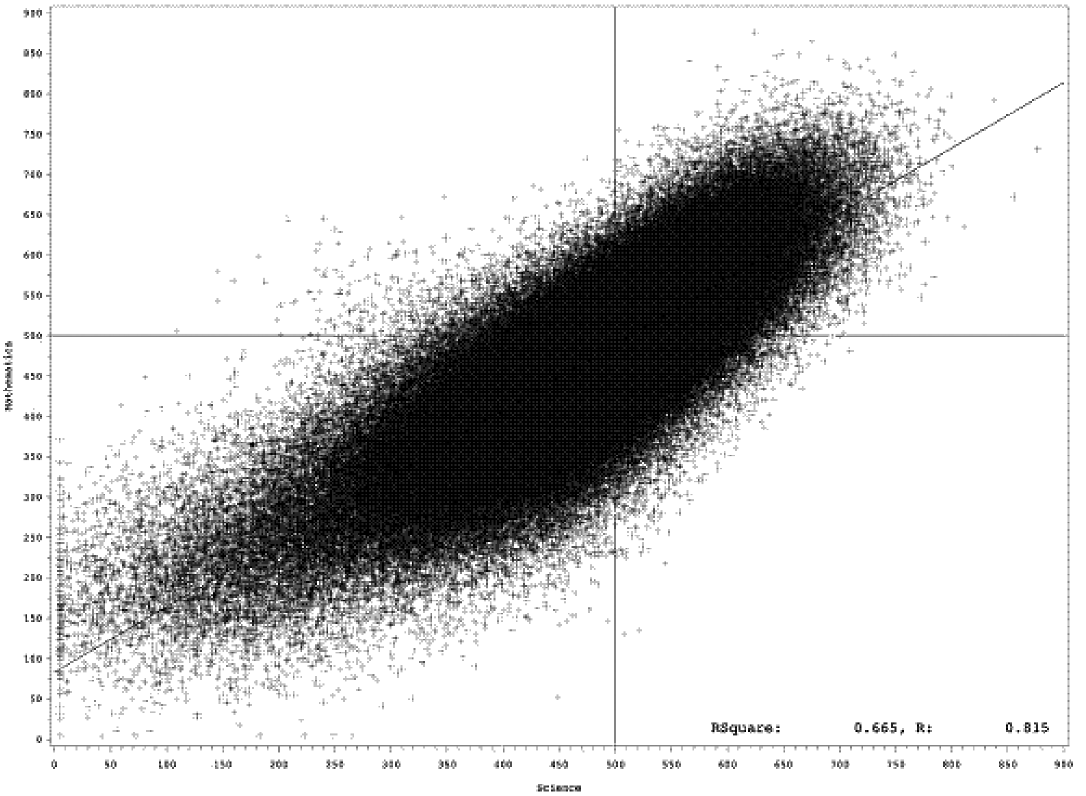
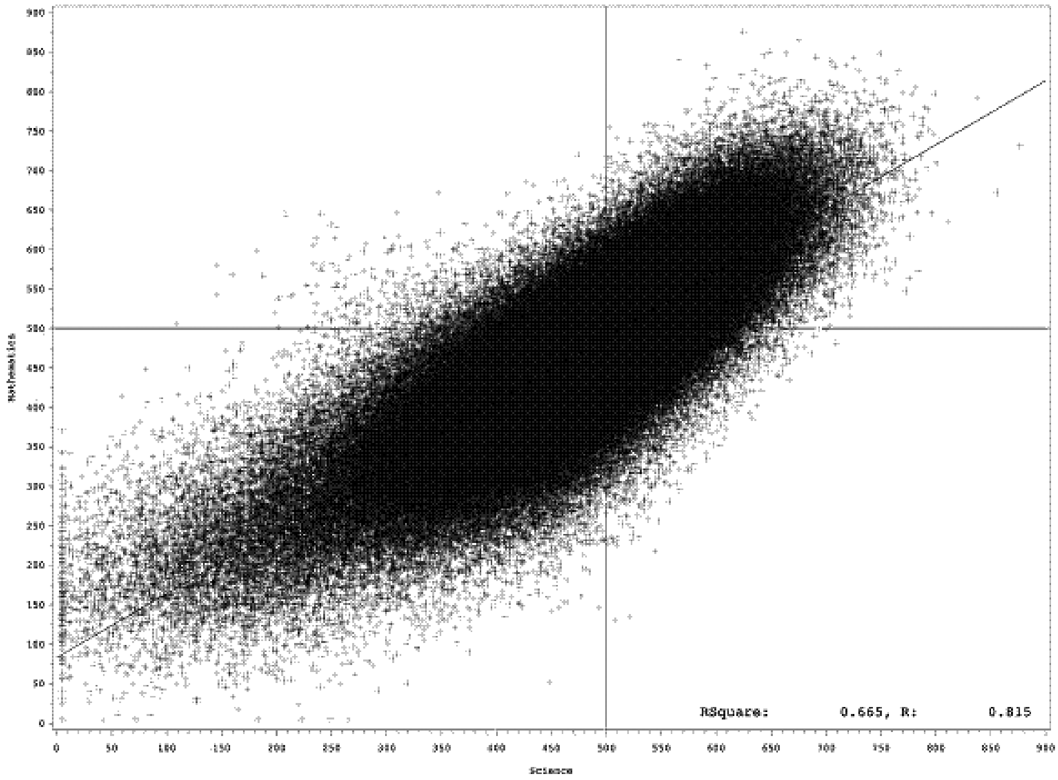


Figure 4: Relationship between Mathematics and Science Scores



Distribution of problem-solving achievement across countries

Figure 5 illustrates the broad range of problem-solving achievement both within and across the countries assessed. As shown in this figure, there was a wide range of performance across countries, with national average scale scores ranging from 588 in Singapore to 307 in Ghana, and an international average of 467. Twenty-four countries performed above the international average, and 22 countries scored below the international average.

In Figure 5, performance within each country is represented by a bar graph, which shows the 5th, 25th, 75th, and 95th percentiles, as well as the 95% confidence level for the mean. Each percentile point indicates the percentage of students below that point on the scale. For most participating countries, there was an enormous range between the highest and the lowest scores. This range was as large as 300 score points for some countries, which is approximately the same as the difference in mean achievement between the highest performing country and the lowest performing country.

Figure 6 shows how a country's average problem-solving achievement compared to achievement in the other participating countries. The figure shows whether or not the differences in average achievement between the pairs of countries are statistically significant. To read the table, select a country of interest from the first column and read across the row corresponding to the country. A circle with a triangle pointing up indicates a significantly higher performance than the performance for the comparison country listed across the top. Absence of the symbol indicates no significant difference in performance, and a circle with a triangle pointing down indicates a significantly lower performance.

As shown in Figure 6, Singapore outperformed all the other countries except the Republic of Korea. The Republic of Korea performed similarly to Singapore as well as to Japan and Chinese Taipei. The latter two countries were outperformed only by Singapore. On the other side of the achievement continuum, Ghana and South Africa were outperformed by all the participating countries.

Figure 5: Distribution of Problem-solving Achievement

	Countries	Years of schooling	Average age	Problem-solving achievement distribution	Average scale score index		Human development
	Singapore	8.0	14.3		588 (4.0)	▲	0.884
^	Korea, Rep. of	8.0	14.6		580 (2.0)	▲	0.879
	Japan	8.0	14.4		576 (2.3)	▲	0.932
	Chinese Taipei	8.0	14.2		572 (4.0)	▲	-
†	Hong Kong SAR	8.0	14.4		566 (3.7)	▲	0.889
	Estonia	8.0	15.2		541 (2.8)	▲	0.833
†	Netherlands	8.0	14.3		540 (4.3)	▲	0.938
	Hungary	8.0	14.5		536 (3.1)	▲	0.837
	Belgium (Flemish)	8.0	14.1		534 (3.0)	▲	0.937
	Australia	8 or 9	13.9		527 (4.5)	▲	0.939
†	Scotland	9.0	13.7		523 (4.1)	▲	0.930
	Sweden	8.0	14.9		518 (3.4)	▲	0.941
	New Zealand	8.5 - 9.5	14.1		513 (5.4)	▲	0.917
‡	United States	8.0	14.2		509 (3.1)	▲	0.937
	Latvia	8.0	15.0		507 (3.5)	▲	0.811
	Slovak Republic	8.0	14.3		505 (3.0)	▲	0.836
	Malaysia	8.0	14.3		505 (3.5)	▲	0.790
	Slovenia	7 or 8	13.8		502 (2.6)	▲	0.881
1	Lithuania	8.0	14.9		501 (2.6)	▲	0.824
	Russian Federation	7 or 8	14.2		500 (3.7)	▲	0.779
	Norway	7.0	13.8		488 (3.1)	▲	0.944
	Italy	8.0	13.9		486 (3.1)	▲	0.916
2	Israel	8.0	14.0		478 (3.9)	▲	0.905
	International Avg.	**8**	**14.5**		**467 (0.5)**		**-**
	Bulgaria	8.0	14.9		454 (4.6)	●	0.795
	Romania	8.0	15.0		453 (4.8)	●	0.773
	Moldova, Rep. of	8.0	14.9		450 (3.8)	●	0.700
1	Serbia	8.0	14.9		450 (2.8)	●	-
	Jordan	8.0	13.9		449 (3.2)	●	0.743
	Cyprus	8.0	13.8		441 (2.2)	●	0.891
	Armenia	8.0	14.9		433 (3.3)	●	0.729
2	Macedonia, Rep. of	8.0	14.6		428 (4.3)	●	0.784
	Bahrain	8.0	14.1		418 (2.2)	●	0.839
	Iran, Islamic Rep. of	8.0	14.4		416 (3.0)	●	0.719
	Chile	8.0	14.2		412 (3.4)	●	0.831
	Palestinian Nat'l Auth.	8.0	14.1		405 (3.2)	●	0.731

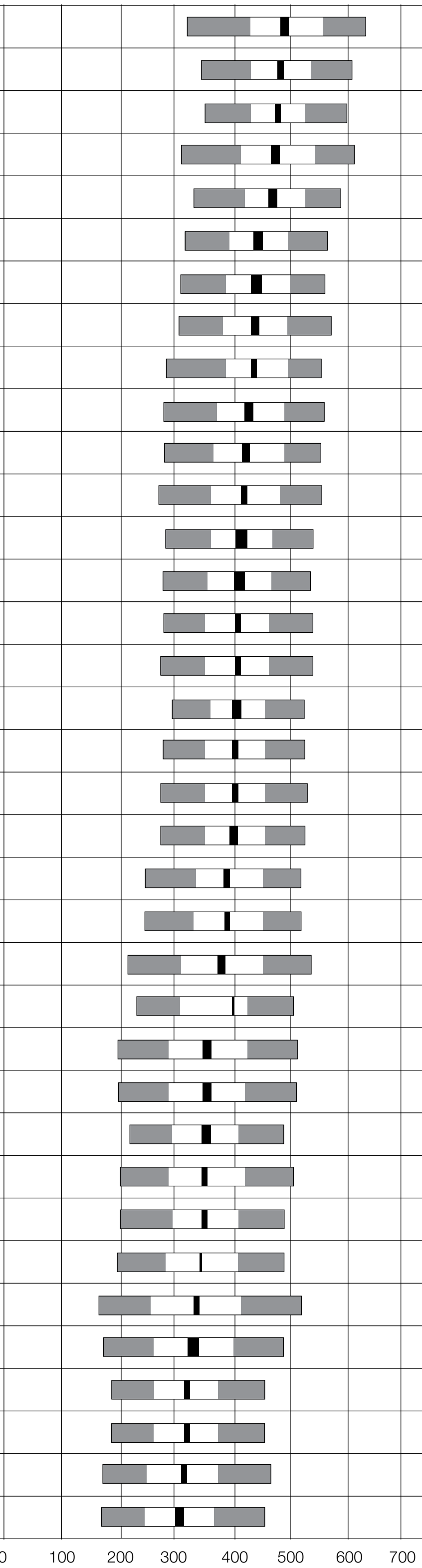
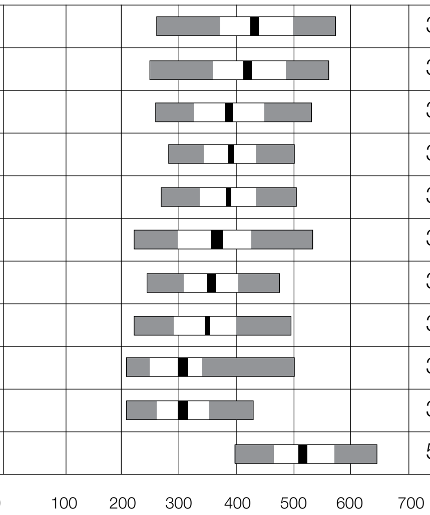


Figure 5 (contd.): Distribution of the Problem-solving Achievement

	Countries	Years of schooling	Average age	Problem-solving achievement distribution	Average scale score		Human development
					Index		
1	Indonesia	8.0	14.5		399 (4.1)	●	0.682
	Egypt	8.0	14.4		399 (3.4)	●	0.648
	Lebanon	8.0	14.6		390 (3.9)	●	0.752
	Tunisia	8.0	14.8		390 (2.9)	●	0.740
1 ‡	Morocco	8.0	15.2		385 (3.4)	●	0.606
	Philippines	8.0	14.8		364 (4.9)	●	0.751
	Saudi Arabia	8.0	14.1		358 (3.8)	●	0.769
	Botswana	8.0	15.1		349 (2.9)	●	0.614
	South Africa	8.0	15.1		308 (4.9)	●	0.684
	Ghana	8.0	15.5		307 (3.9)	●	0.567
	England	9.0	14.3		518 (4.6)	▲	0.930

5th
25th
75th
95th

95th% Confidence Interval for Average (±2SE)

▲ Country average significantly higher than international average

● Country average significantly lower than international average

Notes:

* Represents years of schooling counting from the first year of ISCED Level 1.
** Taken from United Nations Development Programme's *Human Development Report 2003*, pp. 237–240.
† Met guidelines for sample participation rates only after replacement schools were included.
‡ Nearly satisfied guidelines for sample participation rates only after replacement schools were included.
‡ Did not satisfy guidelines for sample participation rates.
1 National Desired Population does not cover all of International Desired Population.
2 National Desired Population covers less than 90% of International Desired Population.
^ Korea tested the same cohort of students as other countries, but later, in 2003, at the beginning of the next school year.
( ) Standard errors appear in parenthesis. Because results are rounded to the nearest whole number, some totals may appear inconsistent.
A dash (–) indicates comparable data are not available.

Difference between problem-solving achievement and achievement in mathematics and science

Figure 7 shows the difference between average achievement in problem solving and average achievement in mathematics for each country. Interestingly, even though some of the items in the two scales are the same, the results reveal that many countries performed relatively better or worse in one area compared to the other (a darkened bar indicates the difference is statistically significant). The range extends from South Africa, where students performed an average 44 score points higher on the problem-solving scale than on the mathematics scale, to Armenia, with 45 score points higher on the mathematics scale than on the problem-solving scale. There are 20 countries with significantly higher relative performance in problem solving and 20 countries with significantly higher relative performance in mathematics.

Figure 8 shows corresponding information for average achievement in problem solving and average achievement in science achievement. Students in 13 countries had a significantly higher relative performance in problem solving than in science, while students in 23 countries had a higher relative performance in science than in problem solving. The range moves from South Africa (71 score points higher in problem solving) to Saudi Arabia (33 score points higher in science).

Problem-solving achievement by gender

Figure 9 shows gender differences in problem-solving achievement. For each country, the figure presents average achievement separately for girls and for boys, as well as the difference between the means. Countries are shown in increasing order of their gender difference. The gender difference for each country is

Figure 6: Multiple Comparisons of Average Problem-solving Achievement

Instructions: Read across the row for a country to compare performance with countries listed along the top of the chart. The symbols indicate whether the average achievement of the country in the row is significantly lower than that of the comparison country, significantly higher than that of the comparison country, or if there is no statistically significant difference between the average achievement of the two countries.

Country	Singapore	Korea, Rep. of	Japan	Chinese Taipei	Hong Kong SAR	Estonia	Netherlands	Hungary	Belgium (Flemish)	Australia	Scotland	Sweden	England	New Zealand	United States	Latvia	Slovak Republic	Malaysia	Slovenia	Lithuania	Russian Federation	Norway	Italy	Israel	Bulgaria
Singapore			▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Korea, Rep. of					▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Japan	●				▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Chinese Taipei	●					▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Hong Kong SAR	●	●	●			▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Estonia	●	●	●	●	●					▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Netherlands	●	●	●	●	●					▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Hungary	●	●	●	●	●						▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Belgium (Flemish)	●	●	●	●	●						▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Australia	●	●	●	●	●	●	●							▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Scotland	●	●	●	●	●	●	●	●	●						▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Sweden	●	●	●	●	●	●	●	●	●							▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
England	●	●	●	●	●	●	●	●	●								▲	▲	▲	▲	▲	▲	▲	▲	▲
New Zealand	●	●	●	●	●	●	●	●	●	●											▲	▲	▲	▲	▲
United States	●	●	●	●	●	●	●	●	●	●											▲	▲	▲	▲	▲
Latvia	●	●	●	●	●	●	●	●	●	●	●	●										▲	▲	▲	▲
Slovak Republic	●	●	●	●	●	●	●	●	●	●	●	●	●									▲	▲	▲	▲
Malaysia	●	●	●	●	●	●	●	●	●	●	●	●	●									▲	▲	▲	▲
Slovenia	●	●	●	●	●	●	●	●	●	●	●	●	●									▲	▲	▲	▲
Lithuania	●	●	●	●	●	●	●	●	●	●	●	●	●									▲	▲	▲	▲
Russian Federation	●	●	●	●	●	●	●	●	●	●	●	●	●	●								▲	▲	▲	▲
Norway	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●				▲	▲
Italy	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●					▲
Israel	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●				▲
Bulgaria	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	
Romania	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	
Moldova, Rep. of	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	
Serbia	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	
Jordan	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	
Cyprus	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Armenia	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Macedonia, Rep. of	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Bahrain	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Iran, Islamic Rep. of	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Chile	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Palestinian Nat'l Auth.	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Indonesia	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Egypt	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Lebanon	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Tunisia	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Morocco	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Philippines	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Saudi Arabia	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Botswana	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
South Africa	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Ghana	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●

Note: 5% of these comparisons would be statistically significant by chance alone.

Figure 6 (contd.): Multiple Comparisons of Average Problem-solving Achievement

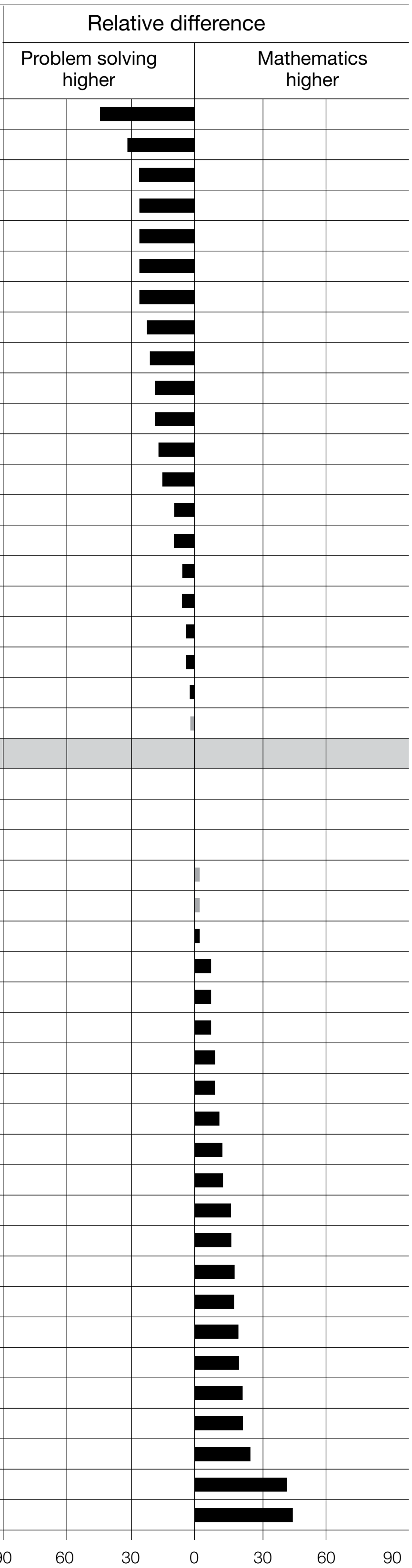
Instructions: Read across the row for a country to compare performance with countries listed along the top of the chart. The symbols indicate whether the average achievement of the country in the row is significantly lower than that of the comparison country, significantly higher than that of the comparison country, or if there is no statistically significant difference between the average achievement of the two countries.

Country	Romania	Moldova, Rep. of	Serbia	Jordan	Cyprus	Armenia	Macedonia, Rep. of	Bahrain	Iran, Islamic Rep. of	Chile	Palestinian Nat'l Auth.	Indonesia	Egypt	Lebanon	Tunisia	Morocco	Philippines	Saudi Arabia	Botswana	South Africa	Ghana
Singapore	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Korea, Rep. of	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Japan	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Chinese Taipei	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Hong Kong SAR	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Estonia	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Netherlands	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Hungary	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Belgium (Flemish)	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Australia	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Scotland	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Sweden	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
England	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
New Zealand	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
United States	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Latvia	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Slovak Republic	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Malaysia	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Slovenia	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Lithuania	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Russian Federation	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Norway	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Italy	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Israel	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Bulgaria					▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Romania					▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Moldova, Rep. of					▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Serbia					▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Jordan					▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Cyprus	●	●	●	●		▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Armenia	●	●	●	●	●			▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Macedonia, Rep. of	●	●	●	●	●			▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Bahrain	●	●	●	●	●	●	●				▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Iran, Islamic Rep. of	●	●	●	●	●	●	●				▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Chile	●	●	●	●	●	●	●					▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Palestinian Nat'l Auth.	●	●	●	●	●	●	●	●	●					▲	▲	▲	▲	▲	▲	▲	▲
Indonesia	●	●	●	●	●	●	●	●	●							▲	▲	▲	▲	▲	▲
Egypt	●	●	●	●	●	●	●	●	●	●	●					▲	▲	▲	▲	▲	▲
Lebanon	●	●	●	●	●	●	●	●	●	●	●						▲	▲	▲	▲	▲
Tunisia	●	●	●	●	●	●	●	●	●	●	●						▲	▲	▲	▲	▲
Morocco	●	●	●	●	●	●	●	●	●	●	●	●	●				▲	▲	▲	▲	▲
Philippines	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●			▲	▲	▲
Saudi Arabia	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●				▲	▲
Botswana	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●			▲	▲
South Africa	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		
Ghana	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		

▲ Average achievement significantly higher than comparison country ● Average achievement significantly lower than comparison country

Figure 7: Relative Difference in Performance between Problem Solving and Mathematics Achievement

Countries	Problem-solving average scale score	Mathematics average scale score	Relative difference
South Africa	308 (4.9)	264 (5.5)	44 (2.2)
Ghana	307 (3.9)	276 (4.7)	31 (2.3)
Norway	488 (3.1)	461 (2.5)	26 (1.3)
Saudi Arabia	358 (3.8)	332 (4.6)	26 (3.2)
Jordan	449 (3.2)	424 (4.1)	25 (1.9)
Chile	412 (3.4)	387 (3.3)	25 (2.2)
Scotland	523 (4.1)	498 (3.7)	25 (2.0)
Australia	527 (4.5)	505 (4.6)	22 (1.9)
England	518 (4.6)	498 (4.7)	19 (1.8)
New Zealand	513 (5.4)	494 (5.3)	19 (2.0)
Sweden	518 (3.4)	499 (2.6)	19 (1.5)
Bahrain	418 (2.2)	401 (1.7)	17 (2.3)
Palestinian Nat'l Auth.	405 (3.2)	390 (3.1)	15 (1.9)
Estonia	541 (2.8)	531 (3.0)	10 (1.4)
Slovenia	502 (2.6)	493 (2.2)	9 (2.1)
Hungary	536 (3.1)	529 (3.2)	6 (1.6)
Japan	576 (2.3)	570 (2.1)	6 (1.9)
United States	509 (3.1)	504 (3.3)	5 (1.0)
Iran, Islamic Rep. of	416 (3.0)	411 (2.4)	4 (2.2)
Netherlands	540 (4.3)	536 (3.8)	4 (1.8)
Italy	486 (3.1)	484 (3.2)	3 (1.6)
International Avg.	**467 (0.5)**	**467 (0.5)**	**0 (0.3)**
Lithuania	501 (2.6)	502 (2.5)	0 (1.4)
Morocco	385 (3.4)	387 (2.5)	1 (2.6)
Latvia	507 (3.5)	508 (3.2)	2 (1.6)
Belgium (Flemish)	534 (3.0)	537 (2.8)	2 (1.5)
Slovak Republic	505 (3.0)	508 (3.3)	2 (1.7)
Malaysia	505 (3.5)	508 (4.1)	3 (1.5)
Macedonia, Rep. of	428 (4.3)	435 (3.5)	7 (2.3)
Egypt	399 (3.4)	406 (3.5)	7 (1.7)
Russian Federation	500 (3.7)	508 (3.7)	8 (2.1)
Korea, Rep. of	580 (2.0)	589 (2.2)	9 (1.3)
Moldova, Rep. of	450 (3.8)	460 (4.0)	10 (2.7)
Indonesia	399 (4.1)	411 (4.8)	12 (2.0)
Chinese Taipei	572 (4.0)	585 (4.6)	13 (1.7)
Philippines	364 (4.9)	378 (5.2)	14 (2.5)
Singapore	588 (4.0)	605 (3.6)	17 (1.2)
Botswana	349 (2.9)	366 (2.6)	17 (3.0
Israel	478 (3.9)	496 (3.4)	18 (2.1)
Cyprus	441 (2.2)	459 (1.7)	18 (1.9)
Hong Kong SAR	566 (3.7)	586 (3.30	20 (1.7)
Tunisia	390 (2.9)	410 (2.2)	20 (2.4)
Bulgaria	454 (4.6)	476 (4.3)	22 (2.5)
Romania	453 (4.8)	475 (4.8)	23 (2.1)
Serbia	450 (2.8)	477 (2.6)	26 (2.3)
Lebanon	390 (3.9)	433 (3.1)	43 (2.4)
Armenia	433 (3.3)	478 (3.0)	45 (2.3)

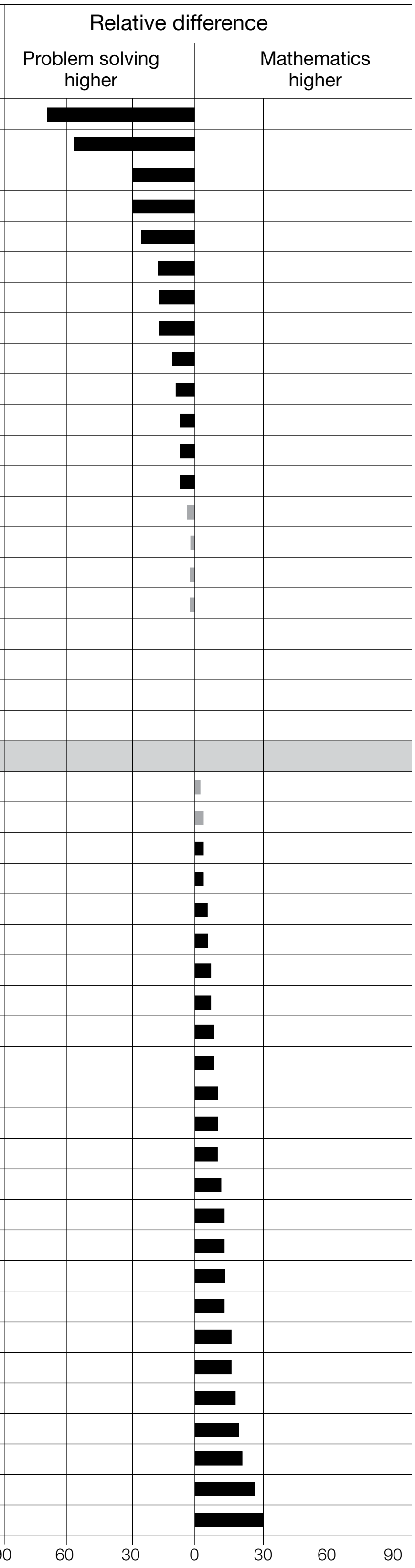


■ Gender difference statistically significant ■ Gender difference not statistically significant

( ) Standard errors appear in parentheses. Because results are rounded to the nearest whole number, some totals may appear inconsistent.

*Figure 8: Relative Difference in Performance between Problem Solving and Science Achievement **

Countries	Problem-solving average scale score	Mathematics average scale score	Relative difference
South Africa	308 (4.9)	237 (6.7)	71 (3.2)
Ghana	307 (3.9)	248 (5.9)	59 (3.1)
Japan	576 (2.3)	545 (1.7)	31 (1.9)
Korea, Rep. of	580 (2.0)	551 (1.6)	29 (1.1)
Belgium	534 (3.0)	509 (2.5)	26 (1.5)
Scotland	523 (4.1)	505 (3.4)	18 (1.8)
Singapore	588 (4.0)	571 (4.3)	18 (1.0)
Hong Kong, SAR	566 (3.7)	549 (3.0)	17 (1.6)
Netherlands	540 (4.3)	529 (3.1)	11 (1.9)
Chinese Taipei	572 (4.0)	564 (3.5)	8 (1.7)
Australia	527 (4.5)	520 (3.8)	7 (1.9)
Cyprus	441 (2.2)	434 (2.0)	7 (2.3)
Chile	412 (3.4)	406 (2.9)	6 (2.0)
Lebanon	390 (3.9)	386 (4.3)	4 (2.3)
Italy	486 (3.1)	484 (3.1)	2 (1.5)
Malaysia	505 (3.5)	503 (3.7)	2 (1.2)
Latvia	507 (3.5)	505 (2.6)	1 (2.0)
Norway	488 (3.1)	487 (2.2)	1 (2.1)
Sweden	518 (3.4)	517 (2.7)	1 (1.6)
New Zealand	513 (5.4)	513 (5.0)	0 (2.1)
Hungary	536 (3.1)	536 (2.8)	0 (1.6)
International Avg.	**467 (0.5)**	**467 (0.6)**	**0 (0.4)**
Israel	478 (3.9)	481 (3.1)	4 (2.6)
Morocco	385 (3.4)	389 (2.5)	4 (2.9)
Estonia	541 (2.8)	545 (2.5)	4 (1.6)
Slovak Republic	505 (3.0)	510 (3.2)	4 (2.1)
Philippines	364 (4.9)	370 (5.8)	6 (2.3)
Tunisia	390 (2.9)	397 (2.1)	6 (2.7)
Russian Federation	500 (3.7)	507 (3.7)	7 (2.0)
Botswana	349 (2.9)	358 (2.8)	8 (2.7)
Romania	453 (4.8)	463 (4.9)	10 (2.6)
Serbia	450 (2.8)	461 (2.5)	10 (2.2)
Lithuania	501 (2.6)	512 (2.1)	11 (1.7)
United States	519 (3.1)	520 (3.1)	11 (1.1)
Slovenia	502 (2.6)	513 (1.8)	12 (1.9)
Bahrain	418 (2.2)	431 (1.8)	13 (1.6)
Macedonia, Rep. of	428 (4.3)	442 (3.6)	14 (2.5)
Indonesia	399 (4.1)	413 (4.1)	14 (1.9)
Moldova, Rep. of	450 (3.8)	465 (3.4)	15 (2.6)
Egypt	399 (3.4)	414 (3.9)	15 (2.0)
Bulgaria	454 (4.6)	472 (5.2)	18 (3.3)
Jordan	449 (3.2)	468 (3.8)	18 1.9)
England	518 (4.6)	537 (4.1)	19 (1.8)
Armenia	433 (3.3)	454 (3.5)	21 (2.4)
Palestinian Nat'l Auth.	405 (3.2)	428 (3.2)	23 (1.9)
Iran, Islamic Rep. of	416 (3.0)	446 (2.3)	31 (2.0)
Saudi Arabia	358 (3.8)	391 (4.0)	33 (2.6)



■ Gender difference statistically significant ▨ Gender difference not statistically significant

Notes: ( ) Standard errors appear in parentheses. Because results are rounded to the nearest whole number, some totals may appear inconsistent.
* To facilitate comparisons among the mathematics, science, and problem-solving achievement scales, they all have been caliberated to have the same mean as the mathematics scale. For science, this meant adjusting the scale by subtracting seven points from each student's science score. Thus, each country's science achievement mean in Figure 8 is seven points lower than in the TIMSS 2003 international science report.

Figure 9: Average Problem-solving Achievement, by Gender

	Countries	Girls	Boys			Difference
		Percent of students	Average scale score	Percent of students	Average scale score	(Absolute value)
	Malaysia	50 (1.8)	505 (4.3)	50 (1.8)	505 (3.8)	0 (4.1)
‡	United States	52 (1.7)	509 (3.2)	48 (0.7)	509 (3.2)	0 (2.0)
	Bulgaria	48 (1.3)	454 (5.2)	52 (1.3)	454 (5.2)	0 (5.1)
	Russian Federation	49 (1.2)	500 (4.0)	51 (1.2)	499 (4.2)	0 (3.3)
1	Indonesia	50 (1.7)	399 (3.9)	50 (0.7)	399 (4.9)	0 (3.5)
	Italy	50 (0.9)	487 (3.0)	50 (0.9)	486 (3.7)	1 (2.7)
	Japan	49 (1.2)	577 (3.6)	21 (1.2)	576 (3.5)	1 (5.5)
	Egypt	46 (2.7)	400 (4.0)	54 (2.7)	398 (5.2)	1 (6.5)
	Slovak Republic	48 (1.3)	504 (3.1)	52 (1.3)	3.6 (3.7)	2 (3.5)
‡	England	50 (2.4)	519 (5.3)	50 (2.4)	516 (5.7)	3 (6.1)
	South Africa	51 (0.9)	307 (5.7)	49 (0.9)	309 (5.4)	3 (5.0)
†	Hong Kong SAR	50 (2.4)	568 (3.9)	50 (2.4)	565 (4.8)	3 (4.8)
†	Netherlands	49 (1.2)	539 (4.4)	51 (1.2)	452 (5.3)	3 (4.5)
^	Korea, Rep. of	48 (2.8)	578 (2.7)	52 (2.8)	582 (2.4)	4 (3.2)
	Estonia	50 (1.0)	543 (3.3)	50 (1.0)	539 (3.2)	4 (3.3)
	International Avg.	**50 (0.2)**	**469 (0.6)**	**50 (0.2)**	**464 (0.6)**	**5 (0.7)**
	Australia	51 (2.2)	525 (5.7)	49 (2.2)	529 (5.2)	5 (6.2)
	Botswana	51 (0.7)	352 (3.0)	49 (0.7)	347 (3.8)	5 (3.6)
	Lebanon	57 (1.8)	392 (4.4)	43 (1.8)	387 (5.1)	5 (5.5)
1	Serbia	49 (0.8)	453 (3.2)	51 (0.8)	448 (3.5)	5 (3.6)
	Romania	52 (0.9)	455 (5.6)	48 (0.9)	450 (5.3)	5 (5.0)
1	Lithuania	50 (0.9)	505 (2.9)	50 (0.9)	498 (3.1)	6 (2.5)
	Norway	50 (0.8)	491 (3.2)	50 (0.8)	484 (3.7)	6 (2.9)
	Scotland	50 (1.3)	526 (5.0)	50 (1.3)	519 (4.2)	7 (4.3)
	Hungary	50 (1.0)	532 (3.4)	50 (1.0)	539 (3.8)	7 (3.5)
1 ‡	Morocco	50 (1.8)	382 (3.8)	50 (1.8)	389 (4.5)	7 (4.5)
2	Israel	52 (1.6)	474 (3.8)	48 (1.6)	481 (5.6)	7 (5.4)
	Armenia	53 (0.7)	436 (4.0)	47 (0.7)	429 (3.9)	8 (4.2)
	Sweden	51 (0.9)	522 (3.8)	49 (0.9)	513 (3.7)	9 (3.2)
	Latvia	49 (0.8)	511 (3.4)	51 (0.8)	502 (4.2)	9 (3.2)
	Philippines	58 (0.9)	368 (5.1)	42 (0.9)	358 (5.7)	10 (4.3)
	Chile	48 (1.6)	407 (4.0)	52 (1.6)	417 (4.2)	10 (4.5
	Belgium (Flemish)	54 (1.2)	530 (4.0)	46 (2.1)	540 (3.9)	10 (5.3)

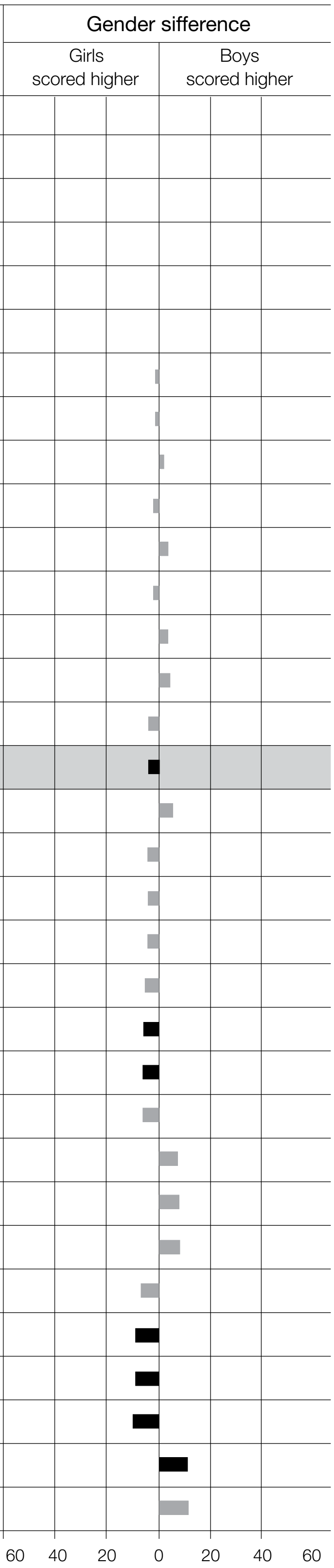
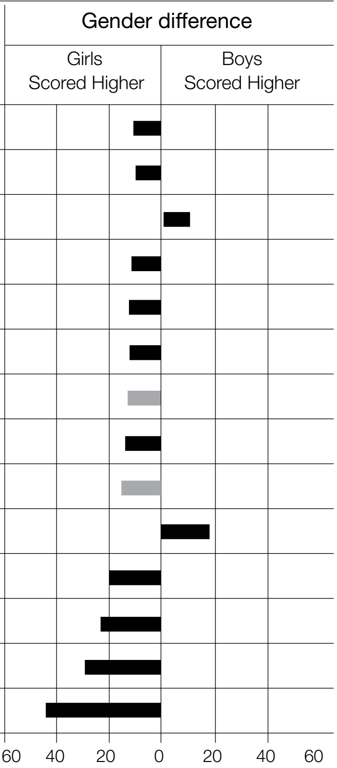


Figure 9 (contd.): Average Problem-solving Achievement, by Gender

Countries	Girls	Boys			Difference
	Percent of students	Average scale score	Percent of students	Average scale score	(Absolute value)
Slovenia	50 (0.9)	507 (3.1)	50 (0.9)	497 (3.0)	11 (3.2)
Singapore	49 (0.8)	594 (3.6)	51 (0.8)	583 (4.9)	11 (3.4)
Ghana	45 (0.9)	301 (4.2)	55 (0.9)	312 (4.0)	11 (2.3)
Moldova, Rep. of	51 (0.8)	456 (4.2)	49 (0.8)	445 (4.5)	11 (4.2)
Chinese Taipei	48 (1.0)	579 (4.4)	52 (1.0)	567 (4.5)	12 (3.8)
New Zealand	52 (1.7)	519 (5.5)	48 (1.7)	507 (6.9)	12 (6.0)
Saudi Arabia	43 (2.3)	365 (6.7)	57 (2.3)	352 (4.9)	13 (8.4)
2 Macedonia, Rep. of	49 (0.9)	435 (4.8)	51 (0.9)	422 (4.5)	14 (3.7)
Iran, Islamic Rep. of	40 (4.1)	424 (4.7)	60 (4.1)	410 (4.9)	15 (7.8)
Tunisia	53 (0.7)	382 (3.1)	47 (0.7)	399 (3.4)	17 (2.9)
Palestinian Nat'l Auth.	55 (2.4)	413 (4.3)	45 (2.4)	395 (5.1)	19 (6.8)
Cyprus	49 (0.6)	453 (2.6)	51 (0.6)	430 (2.7)	22 (3.1)
Jordan	49 (1.7)	464 (4.0)	51 (1.7)	436 (4.5)	28 (5.8)
Bahrain	50 (0.4)	439 (3.2)	50 (0.4)	369 (2.6)	42 (4.0)



■ Gender difference statistically significant ▨ Gender difference not statistically significant

Notes:

† Met guidelines for sample participation rates only after replacement schools were included.

‡ Nearly satisfied guidelines for sample participation rates only after replacement schools were included.

‡ Did not satisfy guidelines for sample participation rates.

1 National Desired Population does not cover all of International Desired Population.

2 National Desired Population covers less than 90% of International Desired Population.

^ Korea tested the same cohort of students as other countries, but later, in 2003, at the beginning of the next school year.

( ) Standard errors appear in parenthesis. Because results are rounded to the nearest whole number, some totals may appear inconsistent.

shown by a bar indicating the amount of difference and the direction of the difference (whether it favored girls or boys). A darkened bar indicates that the gender difference is statistically significant.

On average, as is evident in the figure, there was a small difference favoring girls across countries. Girls had significantly higher achievement than boys in almost one third of the countries (16 countries). In contrast, boys had significantly higher achievement than girls in three countries.

Figure 10 shows the comparisons of the gender difference in mathematics, science, and problem-solving achievement. If boys had a higher average scale score than girls, the difference is in the column labeled boys. If girls had a higher average scale score than boys, the difference is in the column labeled girls. An asterisk (*) indicates that the difference was statistically significant. Only seven countries had consistently significant differences in all three areas—three with boys outperforming girls and four with girls outperforming boys. The countries where boys had higher achievement than girls in problem solving, mathematics, and science were Belgium (Flemish), Chile, Ghana, and Tunisia. Girls had higher achievement than boys in all three areas in Bahrain, Jordan, Macedonia, and Moldova.

Figure 10: Comparisons of Gender Differences in Problem Solving, Mathematics, and Science Achievement

Countries	Relative differences in average scale scores					
	Problem solving		Mathematics		Science	
	Girls	Boys	Girls	Boys	Girls	Boys
Armenia	8		10 *		13 *	
Australia		5		13		20 *
Bahrain	42 *		33 *		29 *	
Belgium (Flemish)		10		11 *		24 *
Botswana	5		3			2
Bulgaria				1		16 *
Chile		10 *		15 *		29 *
Chinese Taipei	12 *		7			1
Cyprus	22 *		16 *		4	
Egypt	1		1		1	
England	3					12 *
Estonia	4		2		3	
Ghana		11 *		17 *		35 *
Hong Kong SAR	3		2			9 *
Hungary		7		7 *		26 *
Indonesia			1			11 *
International Avg.	**5 ***		**1**			**6 ***
Iran, Islamic Rep. of	15		9		1	
Israel		7		8		20 *
Italy	1			6 *		10 *
Japan	1			3		9 *
Jordan	28 *		27 *		27 *	
Korea, Rep. of		4		5		12 *
Latvia	9 *		6			7 *
Lebanon	5			10 *		3
Lithuania	6 *		5			6*
Macedonia, Rep. of	14 *		9 *		8 *	
Malaysia			8			10 *
Moldova, Rep. of	11 *		10 *		8 *	
Morocco		7		12 *		11 *
Netherlands		3		7		15 *
New Zealand	12 *		3			9
Norway	6 *		3			8 *
Palestinian Nat'l Auth.	19 *		8		13 *	
Philippines	10 *		13 *		7	
Romania	5		4			9 *
Russian Federation			3			11 *
Saudi Arabia	13			10	16 *	
Scotland	7		5			12 *
Serbia	5		7 *			6 *
Singapore	11 *		10 *			3
Slovak Republic		2				18 *
Slovenia	11 *		3			7 *
South Africa		3		3		2
Sweden	9 *			1		8 *
Tunisia		17 *		24 *		24 *
United States				6 *		16 *

Note: * Statistically significant

Conclusion

Even though achievement in problem solving across mathematics and science is strongly related to achievement in mathematics and to achievement in science as measured in TIMSS 2003, there were interesting relative differences in performance for the participating countries. For example, 15 countries had significantly higher relative performance in problem solving than in mathematics compared to 20 countries with higher relative performance in mathematics than in problem solving. This finding indicates that curricular and instructional emphasis on problem solving can be a factor in determining overall performance in mathematics. Similarly, 12 somewhat different countries had higher relative performance in problem solving than in science compared to 20 countries with relatively higher science achievement. Gender differences also tended to vary, with only 10 countries having consistently significant differences in problem solving, mathematics, and science. In five of these countries, boys had higher achievement than girls in all three areas; in the remaining five, girls consistently outperformed boys.

References

Case, R. (1985). *Intellectual development: Birth to adulthood.* Orlando, FL: Academic Press.

Gonzalez, E. J., Galia, J., & Li, I. (2004). Scaling methods and procedures for the TIMSS 2003 mathematics and science scales. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report.* Chestnut Hill, MA: Boston College.

Johnson, E. G., Mislevy, R. J., & Thomas, N. (1994). Scaling procedures. In E. G. Johnson & J. E. Carlson (Eds.), *The NAEP technical report* (pp. 241–256) (Report No. 23-TR20). Washington, DC: National Center for Education Statistics.

Martin, M. O. (2005). *TIMSS 2003: User guide for the international database.* Chestnut Hill, MA: Boston College.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international science report.* Chestnut Hill, MA: Boston College.

Mayer, R. E. (1985). Implications of cognitive psychology for instruction in mathematical problem solving. In E. A. Silver (Ed.), *Teaching and learning mathematical problem solving: Multiple research perspectives* (pp. 123–138). Hillsdale, NJ: Lawrence Erlbaum.

Mayer, R. E. (1992). *Thinking, problem solving, cognition* (2nd ed.). New York, NY: Freeman.

Mullis, I. V. S., Martin, M. O., & Foy P. (2005). *IEA's TIMSS 2003 international report on achievement in the mathematics cognitive domains: Findings from a developmental project.* Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report.* Chestnut Hill, MA: Boston College.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Resnick, L. B., & Ford, W. W. (1981). *The psychology of learning mathematics for instruction.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Speedie, S. M., Treffinger, D. J., & Feldhusen, J. F. (1971). Evaluation of components of the Purdue Creative Thinking Program: A longitudinal study. *Psychological Reports, 29*, 395–398.

Wicklegren, W. (1974). *How to solve problems.* San Fransisco, CA: W. H. Freeman.

Consistency between attitudes and practices relating to mathematics in the United States and South Korea[1]

Melissa McNaught
University of Missouri
Columbia, Missouri, USA

Seoung Joun Won
Yonsei University
Seoul, Korea

Abstract

This study drew on information from the 2003 Trends in International Mathematics and Science Study (TIMSS) database to investigate differences and similarities between the United States and South Korea in terms of the relationships between teachers' views of (or attitudes toward) mathematics, their classroom practices, and students' mathematics achievement. Teachers' views and classroom practices affect students' achievement, but more research is needed to determine the role these variables play. Views regarding mathematics education fall into two camps: those that comply with recommendations put forth in the Standards documents of the National Council of Teachers of Mathematics (NCTM) and those relating to more traditional stances. The influence of the Standards extends beyond the borders of the United States. One example is the Seventh Curriculum of South Korea. According to this present study, the views of United States mathematics teachers are more consistent with those of the NCTM recommendations than are the views of their South Korean counterparts, who tend to hold views that are more traditional in nature. The results also indicate that, compared to teachers in Korea, United States mathematics teachers are both individually and as a body more consistent in their views. In addition, the findings suggest that United States teachers show more alignment between their views and their classroom practices than do South Korean teachers.

Introduction

In today's information society, mathematics education is often considered the basis of global competitiveness, prompting many nations to investigate the validity of their educational curricula in mathematics. With the goal of providing effective instruction, researchers have endeavored, over the past several decades, to determine which variables contribute to student achievement in mathematics.

Although various studies show teachers' attitudes toward mathematics and mathematics education are critical in determining how they teach, many researchers report both consistencies and inconsistencies between teachers' espoused views of mathematics education and their classroom practices (Brown & Borko, 1992; Raymond, 1997; Thompson, 1984, 1992). Teachers' attitudes and classroom practice also seem to be significant factors affecting students' achievement, but we do not know much about how or to what extent these variables influence students' achievement (Staub & Stern, 2002).

The International Association for the Evaluation of Educational Achievement (IEA) has conducted a number of multinational studies on educational achievement. The Trends in International Mathematics and Science Study (TIMSS) 2003 database was designed to facilitate secondary analyses aimed at improving mathematics and science education. This database comprises student achievement data in mathematics and science as well as student, teacher, school, and curricula background data for the 48 countries that participated in TIMSS 2003 at the Grade 8 level. Using the TIMSS 2003 database, this study investigates the relationships between teachers' attitudes toward mathematics, their classroom practices, and students' mathematics achievement, and compares differences and similarities in these

[1] We gratefully acknowledge Douglas Grouws, whose comments and suggestions strengthened our analysis and discussion.

relationships between the United States and Korea.

The following questions were considered:

1. To what extent does the relationship between teachers' attitudes and their practices differ between the United States and South Korea?
2. What relationships exist among the different types of teachers' attitudes and practices in the United States and South Korea?
3. Which teachers' attitudes and practices relate to student achievement?

Background

A number of studies have examined the relationships between teacher factors and student achievement; however, the findings from this literature are not conclusive. For instance, Betts and Morell (1999) investigated whether teacher education level and teacher experience affected student academic achievement. The authors reported no relationship between teachers' education levels and students' achievement at the college level. Larson (2000), however, found that teachers' education level affected the performance of elementary school students. With regard to mathematics achievement in the United States, Grouws, Smith, and Sztajn (2004) reported that teacher's undergraduate major at Grade 8, but not at Grade 4, appeared to influence student performance on the National Assessment of Educational Progress (NAEP).

Many studies indicate that teacher experience or preparation has an effect on students' achievement (Fetler, 2001; Grouws et al., 2004; Ngwudike, 2000). Because education involves interaction between teachers and students, it seems likely that teachers' practice in the classroom is one of the most important factors influencing student achievement in mathematics. However, Shymansky, Yore, and Anderson (2000) reported that interactive constructive teaching strategies do not make a significant difference in science achievement. Moreover, the relationship of teachers' practices in classrooms and students' achievement in mathematics remains unclear.

Views regarding mathematics education are divided between those that comply with recommendations put forth in the Standards documents of the National Council of Teachers of Mathematics (NCTM, 1989, 1991, 2000) and those that relate to more traditional stances. The NCTM recommends several components of instruction necessary for a high-quality mathematics education so that mathematical learning is accessible to all students. These recommendations include ideas that encourage students to become active participants in their learning. For example, students are expected to reason, conjecture, justify, and verify mathematics as opposed to mimicking procedures demonstrated by the teacher. The goal is to ensure that the process of learning mathematics is a connected "whole" rather than a process involving isolated bits of stored information.

The acceptance of these recommendations by a teacher is what we refer to in this study as the "NCTM view." The term "traditional," as used in this paper, refers to those long-standing practices underpinned by the premise that the teacher is the possessor of knowledge and is responsible for passing this knowledge on to the students through demonstration. Table 1 classifies items from the TIMSS (2003) teacher questionnaire according to whether the statements agree primarily with the NCTM recommendations or with the traditional views.

With each of these two categories, educators have come to associate certain instructional practices and specific attitudes. For example, if a teacher views mathematics education as mostly a matter of teaching

Table 1: NCTM View versus Traditional View of Mathematics Learning

NCTM view of mathematics learning	Traditional view of mathematics learning
• More than one representation should be used in teaching a mathematics topic • Solving mathematics problems often involves hypothesizing, estimating, testing, and modifying findings • There are different ways to solve most mathematical problems • Modeling real-world problems is essential to teaching mathematics	• Mathematics should be learned as a set of algorithms or rules that cover all possibilities • Solving mathematics problems often involves

algorithms and memorization, we logically would expect routine drill activities during instruction. However, we could expect a teacher who embodies NCTM views of mathematics to engage in classroom practices such as asking students to explain their answers and to work on problems for which there is not an immediate obvious solution. However, research has shown that this is not always the case (Raymond, 1997). In this study, the term we give to the degree to which teacher views and the expected practices align is "consistency."

Method

Participants

Data were collected from the TIMSS database for students and teachers in two countries: the United States and South Korea. All students who participated in TIMSS 2003 were at the Grade 8 level. The sample design used in TIMSS 2003 was a two-stage sample design. In both countries, schools were selected at the first stage using probability-proportional-to-size sampling. At the second stage, one or two classes were randomly sampled in each school (Martin, 2005). The data gathered for analysis in this study contain survey results from 377 mathematics teachers and the achievement scores of 8,912 students in the United States, and survey results from 149 mathematics teachers and the achievement scores of 5,309 students in South Korea.

Measures

Item-related data were collected from two teacher questionnaires, both designed to gather information about teachers' attitudes toward and their classroom practice in relation to the teaching and learning of mathematics.

The first questionnaire, "Attitudes Toward Mathematics," solicited information about teachers' attitudes regarding the nature of mathematics and how the subject should be taught. It included seven observed variables:

1. More than one representation should be used in teaching a mathematics topic
2. Mathematics should be learned as sets of algorithms or rules that cover all possibilities
3. Solving mathematics problems often involves hypothesizing, estimating, testing, and modifying findings
4. Learning mathematics mainly involves memorizing
5. There are different ways to solve most mathematical problems
6. Few new discoveries in mathematics are being made
7. Modeling real-world problems is essential to teaching mathematics.

As a means of measuring the observed variables, teachers were asked, "To what extent do you agree or disagree?" The response format was 1 = agree a lot, 2 = agree, 3 = disagree, and 4 = disagree a lot.

The second questionnaire, "Teaching Mathematics to the TIMSS Class," required teachers to state how often they asked students to do various content-related activities in mathematics. The questionnaire consisted of nine observed variables:

1. Practice adding, subtracting, multiplying, and dividing without using a calculator
2. Work on fractions and decimals
3. Work on problems for which there is no immediately obvious method of solution
4. Interpret data in tables, charts, or graphs
5. Write equations and functions to represent relationships
6. Work together in small groups
7. Relate what they are learning in mathematics to their daily lives
8. Explain their answer
9. Decide on their own procedures for solving complex problems.

These items were assessed with the question, "How often do you usually ask them [the students] to do the following?" The response format was 1 = every or almost every lesson, 2 = about half the lessons, 3 = some lessons, and 4 = never. Table 2 shows the names used for each of these variables.

A matrix sampling was used to assess students' mathematics achievement. Here, the entire mathematics assessment was divided into 14-item blocks, which were then distributed across 12 booklets. The booklets were rotated among the students. Each student was given a booklet containing both mathematics and science items. Because of the complex assessment design, the mathematics results were summarized using item response theory. In this study, to improve reliability, the national mean of students' achievement

Table 2: Variables and Items

Variable	Item
Attitudes toward mathematics	
1. REPRE	More than one representation should be used in teaching a mathematics topic
2. ALGO	Mathematics should be learned as sets of algorithms or rules that cover all possibilities
3. HYPO	Solving mathematics problems often involves hypothesizing, estimating, testing, and modifying findings
4. MEMO	Learning mathematics mainly involves memorizing
5. DIFFW	There are different ways to solve most mathematical problems
6. FEWDIS	Few new discoveries are being made in mathematics
7. MODEL	Modeling real-world problems is essential to teaching mathematics
Teaching mathematics to the TIMSS class	
8. NOCAL	Practice adding, subtracting, multiplying, and dividing without using a calculator
9. WOFRA	Work on fractions and decimals
10. NOOME	Work on problems for which there is no immediately obvious method of solution
11. INDA	Interpret data in tables, charts, or graphs
12. RERE	Write equations and functions to represent relationships
13. SMGP	Work together in small groups
14. DALI	Students relate what they are learning in mathematics to their daily lives
15. EXAN	Students explain their answer
16. DEPRO	Students decide on their own procedures for solving complex problems

is used, that is, the average of five estimates of TIMSS 2003.

Analysis

We used SPSS Version 13.0 to conduct an exploratory factor analysis, the correlation matrix, and a linear regression analysis. Factor analysis was performed to compare factors and the loadings of measured indicator variables on them. The ratio of sample size to number of variables was 23.5 for the United States and 9.3 for South Korea. These ratios satisfied the generally suggested criteria of factor analysis (Kim & Muller, 1978). The extraction method involved a principal component analysis. We chose the VARIMAX rotation option to minimize the number of variables loading highly on a factor. We also used simple linear regression to explore relationships between teachers' attitudes toward mathematics and their classroom practices and students' achievement.

Results

The purpose of the factor analysis was to verify the existence of factors constituting the NCTM recommendations and the traditional attitudes of each country. Five factors were extracted from the United States data. These explained 56% of the variation in the questions about teachers' attitudes toward mathematics and their classroom practices. Table 3 shows the factor loadings for each variable along with the percentage of variance it explains. The first factor is the NCTM recommendations, the second factor is non-routine problem practice, the third is student-centered practice, the fourth is traditional views, and the fifth is practice without a calculator. These factors and the loadings of measured variables on them confirmed what we expected. Table 4 presents the results of the exploratory factor analysis of the data from South Korea. The seven factors extracted explained 65% of the variance among the 16 associated survey items.

Although the Korean classroom-practice factors were somewhat similar to those evident in the United States factor analysis, they differed in number and in how the measured indicator variables loaded on to them. Two variables, ALGO and HYPO, loaded heavily on Factor 4. The variable ALGO, "Mathematics should be learned as sets of algorithms or rules that

Table 3: VARIMAX Rotated Factor Loading for the United States

	Component				
Variable	1	2	3	4	5
REPRE	0.709	0.091	0.044	-0.016	-0.043
ALGO	0.0	-0.148	-0.068	0.711	0.155
HYPO	0.737	0.084	0.069	-0.119	-0.013
MEMO	-0.1	0.042	-0.033	0.719	0.093
DIFFW	0.647	-0.054	0.069	-0.279	0.044
FEWDIS	-0.182	-0.005	-0.061	0.563	-0.173
MODEL	0.705	-0.017	0.25	0.027	-0.058
NOCAL	-0.074	-0.013	0.071	0.029	0.816
WOFRA	0.013	0.206	0.091	0.015	0.814
NOOME	0.144	0.65	0.121	-0.169	-0.017
INDA	0.019	0.751	0.15	-0.004	0.086
RERE	-0.048	0.774	-0.031	0.097	0.112
SMGP	0.024	0.388	0.32	-0.147	-0.356
DALI	0.175	0.065	0.726	0.078	0.074
EXAN	0.031	0.029	0.779	-0.258	0.017
DEPRO	0.172	0.189	0.677	-0.029	0.061
Percentage variance explained	19.26	12.07	9.36	7.69	7.32

Table 4: VARIMAX Rotated Factor Loading for South Korea

	Component						
Variable	1	2	3	4	5	6	7
REPRE	-0.156	0.111	-0.142	0.44	-0.122	0.575	0.215
ALGO	0.003	0.046	0.104	0.721	0.134	-0.162	0.279
HYPO	0.055	-0.041	-0.027	0.693	0.097	0.171	-0.177
MEMO	0.089	-0.077	0.093	0.28	0.66	0.037	-0.06
DIFFW	0.088	-0.043	0.093	0.044	-0.214	0.734	-0.193
FEWDIS	-0.096	-0.077	-0.108	0.026	0.686	-0.12	0.042
MODEL	0.074	0.09	0.037	0.394	-0.496	0.295	0.166
NOCAL	0.841	-0.044	-0.035	0.019	-0.193	-0.065	0.078
WOFRA	0.827	0.161	0.003	0.027	0.147	0.124	0.065
NOOME	0.081	0.111	0.291	-0.285	0.284	0.525	0.36
INDA	0.11	0.818	-0.031	-0.022	-0.016	0.225	0.023
RERE	0.036	0.784	0.02	0.011	-0.17	-0.152	-0.09
SMGP	0.14	-0.042	-0.03	0.076	-0.084	-0.014	0.867
DALI	-0.141	0.481	0.482	0.12	-0.015	-0.022	0.325
EXAN	0.311	-0.024	0.703	0.045	-0.164	-0.057	-0.27
DEPRO	-0.189	-0.001	0.835	-0.02	0.061	0.147	0.1
Percentage variance explained	14.3	10.38	9.96	9.13	8.32	7.17	6.53

cover all possibilities," is one of the traditional views of mathematics education. The variable HYPO, "Solving mathematics problems often involves hypothesizing, estimating, testing, and modifying findings," is consistent with the NCTM recommendations. Thus, the underlying structure of the teachers' attitudes and practice variables for the United States and for South Korea exhibits different patterns.

Tables 5 and 6 present the correlation matrix and descriptive statistics for the United States and South Korea, respectively. Of the correlation coefficients for the United States, only one was statistically significant

Table 5: Correlation Matrix for the United States

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Attitudes toward mathematics																
1. REPRE																
2. ALGO	-.077															
3. HYPO	.347(**)	-.057														
4. MEMO	-.102(*)	.299(**)	-.150(**)													
5. DIFFW	.231(**)	-.172(**)	.388(**)	-.193(**)												
6. FEWDIS	-.115(*)	.193(**)	-.125(*)	.159(**)	-.183(**)											
7. MODEL	.329(**)	-.086	.343(**)	-.101(*)	.347(**)	-.109(*)										
Teaching mathematics to the TIMSS class																
8. NOCAL	-.006	.101(*)	-.047	.101(*)	-.068	.018	-.035									
9. WOFRA	-.004	.07	.009	.106(*)	.046	-.036	.005	.474(**)								
10. NOOME	.133(**)	-.07	.163(**)	-.175(**)	.105(*)	-.062	.083	-.025	.147(**)							
11. INDA	.100(*)	-.105(*)	.106(*)	-.016	-.025	-.074	.049	.022	.181(**)	.367(**)						
12. RERE	.011	-.066	.01	.141(**)	-.023	-.024	.001	.064	.180(**)	.307(**)	.436(**)					
13. SMGP	.109(*)	-.264(**)	.079	-.06	.067	-.110(*)	.096(*)	-.107(*)	-.064	.212(**)	.212(**)	.194(**)				
14. DALI	.135(**)	-.022	.188(**)	-.031	.126(*)	-.117(*)	.270(**)	.08	.152(**)	.111(*)	.220(**)	.08	.183(**)			
15. EXAN	.130(*)	-.163(**)	.155(**)	-.200(**)	.166(**)	-.171(**)	.167(**)	.043	0.06	.155(**)	.181(**)	.082	.160(**)	.403(**)		
16. DEPRO	.151(**)	-.054	.211(**)	-.086	.189(**)	-.135(**)	.203(**)	.049	.120(*)	.331(**)	.202(**)	.128(*)	.207(**)	.301(**)	.469(**)	
N	328	327	329	328	327	327	329	324	317	328	322	319	323	322	325	319
M	1.366	2.397	1.510	3.012	1.324	2.865	1.370	2.317	2.422	2.722	2.745	2.457	2.390	2.087	1.723	2.191
SD	.519	.723	.552	.663	.494	.687	.526	1.005	.757	.737	.532	.694	.843	.808	.811	.841
Mathematics achievement																
17.SMS	.032	.019	.04	.029	-.009	-.077	-.001	-.021	.059	.031	-.004	.112(*)	-.011	-.048	-.005	-.024

Note: (*) Correlation is significant at the 0.05 level. (**) Correlation is significant at the 0.01 level.

Table 6: Correlation Matrix for South Korea

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Attitudes toward mathematics																
1. REPRE																
2. ALGO	.218(**)															
3. HYPO	.218(**)	.275(**)														
4. MEMO	-.028	.181(*)	.065													
5. DIFFW	.278(**)	-.017	.086	-.027												
6. FEWDIS	-.044	.028	.008	.248(**)	-.191(*)											
7. MODEL	.360(**)	.093	.138	-.081	.261(**)	-.193(*)										
Teaching mathematics to the TIMSS class																
8. NOCAL	-.093	-.015	-.01	-.083	.071	-.076	.13									
9. WOFRA	-.018	-.001	.084	.07	.055	-.078	.089	.510(**)								
10. NOOME	.138	.007	-.014	-.001	.164(*)	-.021	.041	-.026	.146(*)							
11. INDA .195(*)	.002	.04	-.13	.104	-.150(*)	.095	.058	.195(*)	.188(*)							
12. RERE .01	.067	-.063	-.111	.016	-.148(*)	.148(*)	-.002	.069	.039	.402(**)						
13. SMGP	.154(*)	.212(**)	-.005	-.032	-.024	-.048	.146(*)	.115	.094	.204(**)	.068	-.011				
14. DALI	.129	.133	-.021	.033	.055	-.03	.205(**)	-.056	.009	.141	.305(**)	.208(**)	.159(*)			
15. EXAN	-.074	-.014	.019	-.02	.071	-.136	.075	.195(*)	.171(*)	.101	.057	.052	-.093	.101		
16. DEPRO	.015	.074	-.006	.029	.141	.003	.051	-.11	-.007	.260(**)	.005	-.021	.025	.367(**)	.353(**)	
N	138	138	137	137	137	137	138	137	137	137	134	135	137	136	137	137
M	1.536	2.05	2.153	3.015	1.452	2.693	1.667	2.036	2.576	2.518	2.821	2.452	2.825	2.353	1.555	2.248
SD	.555	.558	.617	.499	.499	.692	.6081	1.087	.811	.687	.532	.6547	.756	.746	.804	.755
Mathematics achievement																
17.SMS	.104	.017	.054	.049	.012	.277(**)	-.035	.094	-.148(*)	.004	-.1	.063	-.048	.03	.044	.054

Note: (*) Correlation is significant at the 0.05 level. (**) Correlation is significant at the 0.01 level.

at the .05 level of confidence. This was the correlation coefficient of .112 between RERE ("Write equations and functions to represent relationships") and student achievement. For South Korea, the correlation coefficient of .277 between FEWDIS ("Few new discoveries in mathematics are being made") and student achievement was significant at the .01 level. The correlation coefficient of -.148 between WOFRA ("Work on fractions and decimals") and student achievement was significant at the .05 level. The remaining correlation coefficients for South Korea were not significant at the .05 level. The correlation coefficient between NOCAL ("Practice adding, subtracting, multiplying, and dividing without using a calculator") and WOFRA ("Work on fractions and decimals") was the largest for both countries: for the United States, the coefficient was .47 and for South Korea it was .51.

Interestingly, the patterns of correlations among certain variables in the two countries were quite different. In the United States, correlations between variables DALI, EXAN, and DEPRO and variable HYPO were significantly positive, whereas in South Korea, the correlations for these variables were negative. Thus, in the United States, there was a positive relationship between teachers' student-centered practice (such as encouraging students to explain their answer based on their own procedures) and students' ability to solve mathematics problems involving hypothesizing, estimating, testing, and modifying findings. In South Korea, however, the relationship was in the opposite direction. The reverse of the situation was evident for other variables. In the United States, correlations between variable FEWDIS and variables SMGP, DALI, EXAN, and DEPRO were significantly negative, whereas in South Korea correlations between variable FEWDIS and variables INDA and RERE were significantly negative. In practical terms, this means that, in the United States, an increase in thinking that few new discoveries are being made in mathematics was associated with an increase in student-centered practice, while in South Korea the opposite relationship was evident.

Table 7 presents the regression slope coefficients of each variable studied. Here, we can see that one factor in the United States and another factor in South Korea had significant relationships with student achievement at the .05 and .01 levels of significance respectively. In the United States, the factor RERE is significantly associated with student achievement, while the factor FEWDIS, relates significantly to student achievement in South Korea. These findings suggest that significant variables related to achievement are those least likely to be associated with a certain attitude toward mathematics education. In the study, few variables appeared as significant factors impacting on achievement.

Discussion

When we separated the attitudes of United States and South Korean teachers' attitudes toward mathematics and practice into two positions—the NCTM recommendations and traditional views—we found that the United States teachers held more consistent views about mathematics education than their South Korean counterparts held. We also found the relationship between teachers' espoused views and classroom practices was more consistent for the United States teachers than for the South Korean teachers.

One possible explanation for these findings is the role collaboration plays in determining instructional recommendations. In the United States, much time and collaboration went into writing the NCTM recommendations, and teachers work in an environment in which they can openly discuss and investigate different views of instruction. In South Korea, the government leads the revision of curriculum structures, which means that most teachers have limited opportunity to sufficiently consider alternate views of and practices in mathematics education. Another possible explanation for the inconsistency among Korean teachers is that the Seventh Curriculum, while strongly influenced by the NCTM recommendations, has not yet taken root in South Korea because of the short amount of time since its release.

The different correlations among several important variables evident in this study reflect a salient gap between the classroom cultures relating to mathematics in the two countries. These relationships are consistent with the findings of research conducted by Bang (2000). For instance, in the United States, it seems that teachers who have views of mathematics like those presented in Lakatos (1976) (e.g., mathematics is a kind of pseudo-science because discoveries in this field rest on hypothesizing, estimating, testing, and modifying findings) are those teachers most likely to

Table 7: Regression Coefficients

	Student achievement									
	The United States				South Korea					
Variables	B	*SE*	BETA	*t*	Sig.	B	*SE*	BETA	*t*	Sig.
Attitudes toward mathematics										
1. REPRE	4.854	8.404	.032	.578	.564	16.140	13.216	.104	1.221	.224
2. ALGO	2.080	6.054	.019	.344	.731	2.567	13.236	.017	.194	.847
3. HYPO	5.734	7.872	.040	.728	.467	7.627	12.028	.054	.634	.527
4. MEMO	3.432	6.567	.029	.523	.602	8.363	14.813	.049	.565	.573
5. DIFFW	-1.485	8.847	-.009	-.168	.867	2.044	14.784	.012	.138	.890
6. FEWDIS	-8.817	6.353	-.077	-1.388	.166	34.326	10.257	.277	3.347	.001
7. MODEL	-0.100	8.278	-.001	-.012	.990	-4.895	12.137	-.035	-.403	.687
Teaching mathematics to the TIMSS class										
8. NOCAL	-1.632	4.365	-.021	-.374	.709	7.441	6.789	.094	1.096	.275
9. WOFRA	6.106	5.831	.059	1.047	.296	-15.719	9.043	-.148	-1.738	.084
10. NOOME	3.297	5.915	.031	.557	.578	.463	10.792	.004	.043	.966
11. INDA	-0.583	8.267	-.004	-.070	.944	-16.157	14.026	-.100	-1.152	.251
12. RERE	12.774	6.368	.112	2.006	.046	8.208	11.295	.063	.727	.469
13. SMGP	-1.004	5.242	-.011	-.191	.848	-5.506	9.793	-.048	-.562	.575
14. DALI	-4.655	5.440	-.048	-.856	.393	3.480	9.979	.030	.349	.728
15. EXAN	-0.487	5.405	-.005	-.090	.928	4.682	9.218	.044	.508	.612
16. DEPRO	-2.236	5.272	-.024	-.424	.672	6.193	9.805	.054	.632	.529

engage in student-centered instruction focusing on non-routine activities. However, in South Korea, the reverse seems to hold, although the correlation relevant to this consideration was not statistically significant.

One of the most interesting findings in this research was the positive relationship in South Korea between the variable "few new discoveries in mathematics" and student achievement, and the fact that in the United States, the negative relationship between the two variables is not statistically significant. To better explain the relationship between teacher attitudes and practice variables and student achievement, we will need to analyze more relevant variables, such as individual backgrounds.

The reliability of the data used in this research has some limitations because the teacher questionnaire was a self-report. However, Burstein, McDonnell, Van Winkle, Ormseth, Mirocha, and Guitton (1995) found teachers' self-reports were consistent with data drawn from observations of the teachers in the classroom. Mayer (1999) concluded that a self-reported composite of classroom practices was both reliable and valid, although individual indicators were less trustworthy. In this study, teachers' attitudes and classroom practice variables were composites, so they were assumed to be more reliable indicators than those found in some studies.

We did not consider other possible predictors of achievement, such as teachers' and students' backgrounds and school characteristics. We need to explore these relationships in future studies. In particular, we need to take different approaches, such as considering the impact of factors from multilevels. Importantly, different cultural contexts will need to be reflected in a comparative study.

References

Bang, J. (2000). When changes do not make changes: Insights from Korean and U.S. elementary mathematics classrooms. *Education of Primary School Mathematics, 4*(2), 111–125.

Betts, J. R., & Morell, D. (1999). The determinants of undergraduate grade point average: The relative importance of family background, high school resources, and peer group effects. *Journal of Human Resources, 34*, 268–293.

Brown, C. A., & Borko, H. (1992). Becoming a mathematics teacher. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 209–239). New York: Macmillan.

Burstein, L., McDonnell, L. M., Van Winkle, J., Ormseth, T., Mirocha, J., & Guitton, G. (1995). *Validating national curriculum indicators.* Santa Monica, CA: RAND.

Fetler, M. (2001). Student achievement test scores, dropout rates, and teacher characteristics. *Teacher Education Quarterly, 28*(1), 151–168.

Grouws, D., Smith, M., & Sztajn, P. (2004). The preparation and teaching practices of United States mathematics teachers: Grades 4 and 8. In P. Kloosterman & F. Lester, Jr. (Eds.), *Results and interpretations of the 1990 through 2000 mathematics assessments of the National Assessment of Educational Progress* (pp. 221–267). Reston, VA: National Council of Teachers of Mathematics.

Kim, J., & Muller, C. W. (1978). *Factor analysis: Statistical methods and practical issues.* Thousand Oaks, CA: Sage Publications.

Lakatos, I. (1976). *Proofs and refutations: The logic of mathematical discovery* (J. Worrall & E. Zahar, Eds.). Cambridge: Cambridge University Press.

Larson, J. C. (2000). *The role of teacher background and preparation in students' algebra success* (research report). Rockville, MD: Montgomery County Public Schools.

Martin, M. O. (2005). *TIMSS 2003 user guide for the international database.* Chestnut Hill, MA: Boston College.

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

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

Ngwudike, B. C. (2000). *Third International Mathematics and Science Study (TIMSS): Reforming teacher preparation programs to improve student achievement.* Washington, DC: United States Department of Education.

Raymond, A. M. (1997). Inconsistency between a beginning elementary school teacher's mathematics beliefs and teaching practice. *Journal for Research in Mathematics Education, 28*, 550–576.

Shymansky, J. A., Yore, L. D., & Anderson, J. O. (2000). *A study of changes in students' science attitudes, awareness and achievement across three years as a function of the level of implementation of interactive-constructivist teaching strategies promoted in a local systemic reform effort.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Staub, F. C., & Stern, E. (2002). The nature of teachers' pedagogical content beliefs matters for students' achievement gains: Quasi-experimental evidence from elementary mathematics. *Journal of Educational Psychology, 94*, 344–355.

Thompson, A. (1984). The relationship of teachers' conceptions of mathematics and mathematics teaching to instructional practice. *Educational Studies in Mathematics, 15*, 105–128.

Thompson, A. (1992). Teachers' beliefs and conceptions: A synthesis of research. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 127–146). New York: Macmillan.

Effects of public preschool expenditures on the test scores of fourth-graders: Evidence from TIMSS[1]

Jane Waldfogel and Fuhua Zhai
Columbia University
New York, USA

Abstract

The study presented in this paper examined the effects of public preschool expenditures on the mathematics and science scores of fourth-graders; child, family, and school characteristics, other relevant social expenditures, and country and year effects were held constant. The study draws on data from the 1995 and 2003 cycles of the Trends in International Mathematics and Science Study (TIMSS) for seven Organisation for Economic Co-operation and Development (OECD) countries—Australia, Japan, the Netherlands, New Zealand, Norway, the United Kingdom and the United States. This study also explored whether preschool expenditures matter more for children at risk of poor school achievement, as indexed by having low levels of resources in the home or by coming from an immigrant family or a family that does not always speak the test language. Our results indicated small but significant positive effects of public preschool expenditures on the mathematics and science scores of fourth-graders. We found that an increase in preschool expenditures of $US100 per child lifted children's mathematics scores by between .07 and .13 of a standard deviation, and their science scores by between .03 and .07 of a standard deviation. We also found some evidence that children from low-resource homes may gain more from increased public preschool expenditures than might other children, but that children of immigrants may gain less (perhaps because they are less likely to attend such programs). Thus, this study provides new cross-national evidence that increasing public preschool expenditures would raise children's mathematics and science achievement, but mixed evidence as to the role of such expenditures in helping to close gaps in achievement between less and more advantaged students.

Introduction

Early childhood education and care (ECEC) has become an important public policy issue in many countries. In particular, some form of preschool education for children in the one or two years prior to school entry has become normative, as policymakers, researchers, and parents increasingly recognize the value of early education for later learning. By 2001, the number of children in preschool the year prior to school entry was near universal in European countries such as Belgium, France, Norway, Sweden, and the United Kingdom, and approached two-thirds in the United States (Organisation for Economic Co-operation and Development/OECD, 2001).

However, countries differ in their approach to whether families or government should bear the cost and responsibility of providing ECEC services. Many European countries consider ECEC a public responsibility and have been moving to universal and free or heavily subsidized ECEC programs for all children regardless of family income or the employment status of parents, while the United States continues to rely mainly on family members, employers, and private ECEC programs (Kamerman & Waldfogel, 2005; Smolensky & Gootman, 2003; Waldfogel, 2006a, 2006b). Although overall public investment in ECEC is still relatively low in most countries, with the average expenditure on preschool programs per capita as a percentage of that for primary school programs ranging from 12% to 17% in OECD countries since 1985, spending is substantially lower in the United States, where the comparable figures are 0.2–0.5% (OECD, 2005a, 2005b).

Do such cross-country differences in preschool expenditures have consequences for children's school

[1] We gratefully acknowledge funding support from the Russell Sage Foundation and the John D. and Catherine T. MacArthur Foundation. Please address correspondence to Jane Waldfogel, Professor of Social Work and Public Affairs, Columbia University School of Social Work, 1255 Amsterdam Avenue, New York, NY 10027. jw205@columbia.edu An updated version of this paper will appear in a special issue of the journal *Educational Research and Evaluation*.

achievement? Although studies in individual countries such as the United States have shown short- and long-term positive effects of publicly financed ECEC programs such as Head Start and pre-kindergarten, there is little cross-national evidence on the effects of public preschool expenditures. Ideally, we would like to know if, across countries, preschool expenditures improve school achievement for children on average, as well as whether such expenditures provide a disproportionate boost to children at risk of poor school achievement, when other differences across countries are held constant.

Our study provides some preliminary evidence on these questions by examining whether public preschool expenditures are associated with mathematics and science scores for fourth-graders, and by holding constant child, family, and school characteristics as well as country and year effects. Our analysis used two waves of micro-data from the Trends in International Mathematics and Science Study (TIMSS) in seven OECD countries, supplemented with OECD data on public expenditures in those countries over time on preschool, primary school, and other key social programs for families and children. We also examined whether preschool expenditures mattered more for children at risk of poor school achievement, as indexed by having low levels of resources in the home or coming from an immigrant family or a family that did not always speak the test language.

Background and conceptual framework

A critical issue for researchers to explore in relation to justification of public expenditures on ECEC is the effects of such expenditures on school readiness and school achievement. Economic theory suggests that improvements in the availability and quality of ECED programs arising out of public preschool expenditures should be reflected in greater school readiness and school achievement (Heckman & Krueger, 2003). Public expenditures may allow provision of care to more children and improve the quality of care by making it possible to recruit and retain more highly qualified staff, reduce child/staff ratios, equip sufficient and quality facilities, and maintain an effective support and monitoring system (OECD, 2001). However, additional expenditures may not increase access to care if parents simply substitute publicly provided care for care they would have purchased anyway, and may not improve the quality of care if the publicly provided care is not of better quality than the care children were placed in previously (Blau, 2001).

If public preschool expenditures increase access to care, and possibly the quality of that care, then there are several reasons to expect that children from disadvantaged families might benefit more than would other children from expansions in public preschool expenditures. First, disadvantaged children might be more likely than more advantaged children to receive no care or inferior quality child-care in the absence of public spending. For instance, in the United States, large disparities in preschool enrollment rates exist between lower-income and higher-income children, and between Hispanic children and non-Hispanic children (Bainbridge, Meyers, Tanaka, & Waldfogel, 2005; Meyers, Rosenbaum, Ruhm, & Waldfogel, 2004). A second reason why disadvantaged children could benefit more from public preschool expansions is that they might gain more than other children from a given level of provision. Many studies of preschoolers have found that the benefits are larger for children who come from more disadvantaged backgrounds than for children who come from more advantaged backgrounds (Magnuson & Waldfogel, 2005).

Studies in individual countries, and in particular, the United States, show short- and long-term positive effects of publicly financed ECEC programs, such as Head Start (Currie 2001; Currie & Thomas, 1995; Garces, Thomas, & Currie, 2002; Puma, Bell, Cook, Heid, & Lopez, 2005) and pre-kindergarten (Barnett, Lamy, & Jung, 2005; Gormley & Gayer, 2005; Gormley, Gayer, Phillips, & Dawson, 2005; Magnuson, Meyers, Ruhm, & Waldfogel, 2004; Magnuson, Ruhm, & Waldfogel, in press). However, there is little cross-national evidence on the effects of public preschool expenditures. We were unable to locate any prior studies that examined the effects of public preschool expenditures on children's school achievement across countries and over time.

Cross-national studies are potentially useful because they allow for an examination of the effects of the policy variation that exists across countries. However, a challenge to causal inference is that there might be unobserved variables that are correlated with individual or national indicators and that affect child outcomes. Our study aimed to address this challenge by using multiple waves of data and including country and year fixed effects. Nevertheless, we recognize that with only two waves of data, our power to detect effects is limited, and therefore we intend to include further waves of data (as these become available) in future work.

Our approach draws on two previous bodies of literature. First, we build on studies, such as that by Hampden-Thompson and Johnston (2006), which demonstrate that some individual-level variables, including both school factors (e.g., curriculum and teacher qualifications) and non-school factors (e.g., students' socioeconomic and immigrant status), are key predictors of children's educational achievement. The inclusion of such individual-level controls is essential, particularly given the differences in school and non-school characteristics of children across countries. Second, to inform our analysis of the role of policies, we drew on two cross-national studies that examined the effects of parental leave policies, and other child and family policies, on infant health (and other outcomes). These exemplary studies by Ruhm (2000) and Tanaka (2005) used multiple waves of data across countries over time and included country and year fixed effects, as well as controls for other relevant social expenditures. The inclusion of the country and year fixed effects is key, since it means that the analyses provide estimates of the effects of changes in policies within countries over time. In the case of parental leave, these studies indicate what the effect is on infant health when a country extends its period of paid parental leave, and when they (the studies) hold constant fixed characteristics of the country as well as secular trends in infant health across countries over time. We utilized a similar approach here, although, as noted earlier, our analysis was limited by our having only two waves of data at the time.

Thus, our study expands previous literature on the effects of public preschool expenditures on student achievement in two ways. First, in utilizing a similar methodology to the one Ruhm (2000) and Tanaka (2005) used in their work on parental leave policies, our study provides the first cross-national evidence on whether public preschool expenditures are associated with higher mathematics and science scores for fourth-graders, when child, family, and school characteristics, other relevant social expenditures, and country and year effects are held constant. Second, we provide evidence on whether preschool expenditures matter more for children from disadvantaged backgrounds. We measured disadvantaged background in terms of the child belonging to an immigrant family or a family that did not always speak the test language, or to a family that had low levels of resources in the home.

Data

We used data from two waves of TIMSS—1995 and 2003. TIMSS collects educational achievement data, as well as extensive background information on child, family, teacher, and school factors related to the learning and teaching of mathematics and science, for children who are primarily in Grade 4 and Grade 8. We used data for fourth-graders, so that we could estimate the effects of preschool experiences as close to school entry as possible. We focused our analysis on countries that were present in both waves of the TIMSS data and for which we had complete data on public expenditures, our main independent variable of interest. Our sample included seven countries: Australia, Japan, the Netherlands, New Zealand, Norway, the United Kingdom, and the United States.

Our outcome variables were the mathematics and science test scores of children in Grade 4 in TIMSS 1995 and TIMSS 2003. The mathematics scores in the 1995 wave had a mean of 542, with a standard deviation of 91. In the 2003 wave, the mean was 512, with a standard deviation of 81. To make the test scores comparable between waves, we standardized each wave to have a mean of 100 and a standard deviation of 10.

With regard to the independent variables, we needed to take account of the fact that the TIMSS surveys are administered at school rather than at household level, and so they lack some important socio-demographic information about the child's parents and family background. For example, there is no information about factors such as parents' level of education and employment or family structure and income, all of which have important effects on children's academic achievement (see, for example, Brooks-Gunn, 1995; Brooks-Gunn & Markman, 2005; Leventhal & Brooks-Gunn, 2002; McLanahan & Casper, 1995). However, the TIMSS surveys do include some information gathered from the child about his or her home environment. In this study, we used information that we considered would be particularly likely to reflect the socioeconomic status of the family and the family's attitudes toward and support for education. We accordingly included variables for the immigration status of parents, whether the child always spoke the test language at home, the number of books in the home, and controls for whether the child had a calculator, computer, study desk, or dictionary at home.

To maximize the sample size so as to maintain the power of regression results, we used several strategies to keep observations with missing values in the analysis. For category variables, we created the category "missing" to flag those observations with missing values. For continuous variables, we used the means of the non-missing observations to impute the missing values, and created a dummy variable to note whether the values of observations were missing or imputed. In both cases, when estimating the regression models, we always included the categories that indicated missing observations. However, for the sake of simplicity, we do not report the coefficients on these missing variable dummies. Table 1 shows the percentage of cases for which there is missing data on any item.

Our final analysis sample included 62,294 observations, with 28,437 from TIMSS 1995 and 33,857 from TIMSS 2003. As evident in Table 1, the distribution of children participating in TIMSS from our sample of seven countries was fairly constant across the two years, although Australia dropped from being 23% of the sample in 1995 to 13% in 2003. On average, children in the sample were about 10 years old and evenly distributed by gender. Their average family size was five people. About 16% of the children in 1995 and 22% in 2003 had parents born in a foreign country, and about 10% of the children did not always speak the test language at home. Most of the children spent some time doing jobs at home and reading books for enjoyment. The majority of children had more than 25 books at home, and had a calculator, computer, study desk, and dictionary at home.

TIMSS also includes extensive data about teacher and school characteristics that are likely to matter for students' achievement in mathematics and science. As shown in Table 1, about half of the teachers were in the middle age group (30 to 49) and the majority of them were female. Most of the teachers had bachelor and higher degrees, and they had, on average, 15 years of teaching experience. The average class size was 26. Most of the schools in the sample were in urban areas. The majority of schools had less than 5% of their students absent on a typical day. Less than 20% of the schools had 50% of more of their students belonging to disadvantaged families.

We extracted our data on public expenditures on preschool and primary school from the OECD online database (OECD, 2005a). We converted the data on total public expenditures on preschool and primary school to expenditures per child by dividing the total expenditures on each item by the number of children in the relevant age group (which we obtained by dividing the total number of children of ages 0 to 14 by three). Thus, our measures capture how much the government spends per child in that age group, not how much the government spends per enrolled child. This is the correct measure given that we wanted to gauge how many children the public expenditures reach and how generous the expenditures are per enrolled child. (This distinction is of little importance for primary school, but is important for preschool because the share of children served by public programs is not 100% and varies considerably across countries.)

Data on other public social expenditures per capita, including spending on family cash benefits, health, and other social spending (this consists of programs such as old age benefits, survivors, incapacity-related benefits, employment, unemployment, and housing) are from OECD health data 2005 (OECD, 2005b). The OECD provides these figures on a per capita basis and we used them in this form. It would not be correct to standardize on a per-child basis, as these expenditures are intended to reach adults as well as children. All the figures for public expenditures on preschool and primary school education, as well as for other social programs, are in $US2,000 adjusted by purchasing power parity (PPP) so that they are comparable across countries and years.

We assigned to each child the average value of the expenditure variables in his or her country during his or her preschool years, and primary school years. We defined the preschool years for children in Grade 4 in 1995 as 1985 to 1991, and their primary school years as 1991 to 1995. We defined the preschool years for children in Grade 4 in the 2003 wave as 1993 to 1999, and their primary school years as 1999 to 2003. Each country's average expenditures on preschools, primary schools, and other social programs (family cash benefits, health, and other) during the child's

Table 1: Descriptive Statistics for Child, Family, Teacher, and School Characteristics

Variables	1995 (*N* = 28,437)		2003 (*N* = 33,857)	
Participating countries				
Australia	0.23	(0.42)	0.13	(0.33)
Japan	0.15	(0.36)	0.13	(0.34)
Netherlands	0.09	(0.28)	0.09	(0.28)
New Zealand	0.09	(0.28)	0.13	(0.33)
Norway	0.08	(0.27)	0.13	(0.33)
United Kingdom	0.11	(0.31)	0.11	(0.31)
United States	0.26	(0.44)	0.29	(0.45)
Child and family characteristics				
Child's age				
Months (imputed)	122.20	(5.47)	121.58	(5.39)
Missing	0.02	(0.13)	0.02	(0.14)
Girl				
Yes	0.50	(0.50)	0.50	(0.50)
No	0.50	(0.50)	0.50	(0.50)
Missing	0.00	(0.06)	0.00	(0.02)
Number of family members				
Members (imputed)	4.89	(1.52)	4.76	(1.34)
Missing	0.19	(0.39)	0.04	(0.20)
Child of immigrants				
Yes	0.16	(0.36)	0.22	(0.42)
No	0.63	(0.48)	0.69	(0.46)
Missing	0.22	(0.41)	0.09	(0.28)
Always speaks test language at home				
Yes	0.68	(0.47)	0.89	(0.31)
No	0.10	(0.31)	0.09	(0.29)
Missing	0.22	(0.41)	0.02	(0.14)
Child does jobs at home				
No time	0.15	(0.35)	0.16	(0.37)
Less than 1 hour	0.48	(0.50)	0.45	(0.50)
1–2 hours	0.22	(0.41)	0.20	(0.40)
More than 2 hours	0.10	(0.30)	0.14	(0.35)
Missing	0.06	(0.24)	0.05	(0.21)
Child reads books for enjoyment				
No time	0.19	(0.39)	0.20	(0.40)
Less than 1 hour	0.43	(0.50)	0.42	(0.49)
1–2 hours	0.21	(0.41)	0.20	(0.40)
More than 2 hours	0.10	(0.30)	0.14	(0.35)
Missing	0.06	(0.24)	0.05	(0.21)
Number of books at home				
0–25 books	0.14	(0.35)	0.29	(0.46)
More than 25 books	0.66	(0.47)	0.68	(0.47)
Missing	0.19	(0.39)	0.02	(0.15)
Has a calculator at home				
Yes	0.74	(0.44)	0.90	(0.30)
No	0.08	(0.28)	0.06	(0.24)
Missing	0.18	(0.38)	0.04	(0.19)

Table 1 (contd.): Descriptive Statistics for Child, Family, Teacher, and School Characteristics

Variables	1995 (*N* = 28,437)		2003 (*N* = 33,857)	
Has a computer at home				
Yes	0.52	(0.50)	0.86	(0.35)
No	0.30	(0.46)	0.11	(0.31)
Missing	0.18	(0.38)	0.03	(0.17)
Has a study desk at home				
Yes	0.71	(0.45)	0.82	(0.39)
No	0.11	(0.31)	0.16	(0.36)
Missing	0.18	(0.38)	0.03	(0.16)
Has a dictionary at home				
Yes	0.71	(0.45)	0.84	(0.37)
No	0.11	(0.31)	0.14	(0.34)
Missing	0.18	(0.38)	0.02	(0.15)
Teacher and school characteristics				
Teacher's age				
Young (< 30)	0.15	(0.36)	0.19	(0.39)
Middle age (30–49)	0.60	(0.49)	0.49	(0.50)
Old age (50+)	0.15	(0.36)	0.24	(0.43)
Missing	0.10	(0.30)	0.08	(0.28)
Female teacher				
Yes	0.63	(0.48)	0.70	(0.46)
No	0.27	(0.44)	0.22	(0.41)
Missing	0.10	(0.29)	0.08	(0.27)
Teacher's education				
Secondary and lower	0.33	(0.47)	0.22	(0.41)
BA/equivalent	0.29	(0.45)	0.50	(0.50)
MA/PhD	0.12	(0.32)	0.19	(0.39)
Missing	0.26	(0.44)	0.09	(0.28)
Teacher's years of teaching				
Years (imputed)	15.64	(8.54)	14.82	(10.12)
Missing	0.10	(0.31)	0.10	(0.30)
Teacher's class size				
Number of children (imputed)	26.97	(6.23)	26.01	(5.78)
Missing	0.16	(0.37)	0.29	(0.45)
Rural area of school				
Yes	0.19	(0.39)	0.23	(0.42)
No	0.70	(0.46)	0.67	(0.47)
Missing	0.11	(0.32)	0.10	(0.30)
Percentage of students absent from school				
Less than 5%	0.71	(0.45)	0.74	(0.44)
More than 5%	0.15	(0.36)	0.18	(0.38)
Missing	0.13	(0.34)	0.08	(0.28)
Percentage of students from disadvantaged families at school				
0–10%	0.17	(0.37)	0.31	(0.46)
11–50%	0.15	(0.35)	0.28	(0.45)
> 50%	0.05	(0.23)	0.18	(0.39)
Missing	0.63	(0.48)	0.23	(0.42)

Note: Means with standard deviations in parentheses.

preschool and primary school years were assigned to each child accordingly.

Table 2 shows that Norway had the highest expenditures per child on preschool programs in both waves, while Japan and Australia ranked as the lowest spending countries. The United States ranked third among the seven countries in both years, with the second highest increase in the level of spending from 1995 to 2003. Norway raised its preschool expenditures dramatically during this period, while Australia, in contrast, reduced its spending. All countries spent substantially more per child on primary school expenditures than on preschool expenditures. Norway again was the highest spender on primary school expenditures per child, followed by the United States. Norway also had the highest expenditure per capita on family cash benefits, health, and other social programs during both waves; the United States generally stood in the middle of these countries.

Method

In determining if higher public preschool expenditures are associated with better student achievement (while holding constant other factors that vary across countries, as well as across students), we wanted to control for observed country differences and also country-level factors that we could not observe but knew might be correlated with higher spending and with better outcomes. To address the problem of unobserved heterogeneity across countries, we used ordinary least squares (OLS) regression analysis with country and year fixed effects, as well as controls for school and non-school factors. The inclusion of country and year fixed effects was important, as it meant that our estimates of the effect of preschool expenditures reflected the effect of changes in these expenditures within countries over time. Because other expenditures, such as primary education and social expenditures, can also affect the educational achievement of children in Grade 4, our analysis also controlled for public expenditures on primary education, and other social programs.

The basic model used in this study therefore was:

$$S_{ict} + \beta_0 + \gamma_c + \gamma_1 + \beta_i X_{ict} + \beta_p E_{ct} + \varepsilon_{ict}$$

where S_{ict} represents the test scores of individual child i from country c at time t; γ_c is the country fixed effect; γ_1 stands for the year fixed effect; X_{ict} represents a vector of child, parent, family, teacher, and school characteristics related to individual child i in country c at time t; E_{ct} indicates public expenditures, including expenditures on preschools, primary schools, and other social programs, in country c at time t; and ε is a random error term.

Our estimation approach involved a series of increasingly controlled models, which allowed us to see first the raw differences in school achievement across countries and then how those differences changed with the addition of controls for key sets of predictor variables. Our first model predicted children's mathematics and science scores in Grade 4 as a function of country fixed effects and control for the earlier wave of data, 1995. We included country dummies for Australia, Japan, the Netherlands, New Zealand, Norway, and the United Kingdom, which meant our base case for this model was a fourth-grader from the United States. Model 2 added controls for child age, gender, and family size. Model 3 added controls for whether the child's parent was an immigrant and whether the child always spoke the test language at home. Model 4 added controls for the child doing jobs and reading books at home, as well as controls for home resources, such as having fewer than 25 books in the home and not having a calculator, computer, study desk, or dictionary. Model 5 added teacher and school characteristics, including the teacher's age, gender, education, years of teaching, class size, urban location, percentage of students absent on a typical day, and percentage of students from disadvantaged families.

Our next step was to estimate five additional models exploring the role of preschool and other policies. In Model 6, we added a control for public expenditures on preschool per child as well as for primary education expenditures per child. These controls captured the main variable of interest in this study—public preschool expenditures—as well as primary school expenditures, which are important in their own right and which might correlate with preschool expenditures. As a robustness check, we added, in Model 7, detailed controls for three other types of social expenditures (family cash benefits, health, and other social programs) that occur during children's preschool years, while in Model 8 we added detailed controls for expenditures on these three other categories of programs during the children's primary school years to date. As a further robustness check, Models 9 and 10 controlled for measures of total other expenditures during the preschool years and primary school years.

Table 2: Public Education and Social Expenditures by Country and Year

Public education and social expenditures during preschool years (PPP $US2,000)

	Preschool expenditures (per child)			Public social expenditures (per capita)								
				Family cash benefits			Health			Other social expenditures		
	Wave 1995	Wave 2003	Differences	Wave 1995	Wave 2003	Differences	Wave 1995	Wave 2003	Differences	Wave 1995	Wave 2003	Differences
Australia	207	189	-18	215	563	348	813	1,252	439	1,094	2,033	939
Japan	156	351	195	65	104	39	761	1,256	495	1,019	1,854	835
Netherlands	1,249	1,481	232	280	294	14	841	1,308	467	3,175	4,045	870
New Zealand	294	539	245	337	407	70	717	1,009	292	1,677	2,029	352
Norway	2,047	3,415	1,368	447	904	457	1,068	1,752	684	1,398	3,894	2,496
United Kingdom	525	856	331	313	480	167	740	1,193	453	1,942	3,055	1,113
United States	914	1,398	484	101	155	54	921	1,701	780	1,692	2,436	744

Public Education and Social Expenditures during Primary School Years (PPP 2000 U.S. Dollars)

	Preschool expenditures (per child)			Public social expenditures (per capita)								
				Family cash benefits			Health			Other social expenditures		
	Wave 1995	Wave 2003	Differences	Wave 1995	Wave 2003	Differences	Wave 1995	Wave 2003	Differences	Wave 1995	Wave 2003	Differences
Australia	4,519	6,663	2,144	412	743	331	1,030	1,634	604	1,699	2,418	719
Japan	4,800	6,828	2,028	86	141	55	1,074	1,561	487	1,501	2,463	962
Netherlands	4,483	6,060	1,577	298	314	16	1,215	1,516	301	4,009	4,180	171
New Zealand	3,498	5,231	1,733	339	469	130	894	1,244	350	1,975	2,177	202
Norway	8,831	14,305	5,474	752	1,120	368	1,451	2,347	896	3,440	4,850	1,410
United Kingdom	4,959	4,734	-225	418	554	136	1,040	1,505	465	2,712	3,462	750
United States	6,673	8,487	1,814	154	132	-22	1,447	2,022	575	2,224	2,732	508

To examine whether the disadvantaged children benefited more from preschool expenditures than did their more advantaged peers, we added interactions between disadvantaged background factors and public expenditures to the models. These factors included variables related to whether the child's parents were immigrants, if the child did not always speak the test language at home, whether the family had fewer than 25 books in the home, and whether the child did not have a calculator, computer, study desk, or dictionary at home. We added the interactions between the disadvantaged background factors and per child expenditures on preschools and primary schools but did not include controls for other social expenditures in the models. In additional analyses (results not shown but available upon request), we added the interactions between these disadvantaged background factors and all public expenditures, including expenditures on preschools, primary schools, family cash benefits, health, and other social programs.

Results

Table 3 presents the results for the children's mathematics scores. The model that we begin with here simply estimates raw differences across countries, as compared to the base case of a fourth-grader in the United States in 2003. The results showed that fourth-graders from Japan and the Netherlands out-scored fourth-graders from the United States by about 6.7 and 4.5 points, respectively. Because the standard deviation of the test score outcome variable was 10, these coefficients translated into differences of .67 and .45 of a standard deviation, respectively. Fourth-graders in the other four countries had mean scores ranging from .07 standard deviations above the United States mean to .58 standard deviations below the United States mean.

Our next model (Model 2) added controls for child age, gender, and family size. Our third model (Model 3) added controls for parent immigrant status and language at home, while Model 4 added controls for other child and home characteristics. These controls affected mathematics scores in the direction we expected, with children of immigrants, children who did not always speak the test language at home, and children with fewer books or lacking other resources in the home attaining lower scores in mathematics. However, adding these controls to the models did not substantially alter the country positions. In Model 4, fourth-graders from the Netherlands, for instance, had an advantage of .36 of a standard deviation, as compared to .45 in Model 1. The addition of teacher and school characteristics in Model 5 had a somewhat larger effect on the country coefficients. The Netherlands' advantage, for instance, reduced to 0.26 of a standard deviation. Like the child and family characteristics, the teacher and school characteristics worked as expected. For instance, children gained higher mathematics scores when they had teachers with higher levels of teaching experience, when they were in smaller classes, and when they were in schools that had absentee rates and percentages of disadvantaged students that were lower than those of the other schools in the samples.

Model 6 added the control for public preschool expenditures per preschool-age child. The coefficient was small but statistically significant, and effectively meant that a $100-per-child increase in public preschool expenditures led to a gain of .07 of a standard deviation in the fourth-graders' mathematics scores. The model also controlled for public primary school expenditures per school-age child, but the coefficient on this variable was negative, indicating that when countries increase their primary school expenditures by $100 per child, mathematics scores fall by an average of .03 of a standard deviation. This latter result is somewhat puzzling, but it is consistent with findings in the literature on school expenditures that show higher expenditures are not always associated with improved school achievement (Hanushek, 2006). One possible reason is that such expenditures may be endogenous (i.e., school spending goes up when achievement is lagging or when more special-needs students are enrolled).

Adding the controls for preschool and primary school expenditures in Model 6 did alter some of the country coefficients. Most dramatically, the coefficient for the Netherlands switched to .51 of a standard deviation below the United States average, a finding which suggests that preschool expenditures play a role in explaining the Netherlands' superior position in the earlier models. New Zealand and the United Kingdom also had more negative means relative to the United States than they had in the previous models, suggesting that, for them also, higher preschool expenditures confer advantages relative to the situation in the United States.

Table 3: OLS Regression of Mathematics Scores without Interactions

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Wave of data (wave 2003 omitted)										
Wave 1995	-0.588***	-0.059	-0.209***	-0.572***	-0.828***	-3.675***	-5.711***	-3.752***	-4.923***	-5.033***
	(0.076)	(0.078)	(0.080)	(0.079)	(0.096)	(0.143)	(0.366)	(0.791)	(0.249)	(0.249)
Participating countries (the US is the omitted base case)										
Australia	0.737***	0.740***	0.508***	-0.211**	-0.792***	-0.013	-1.658***	5.476***	1.046***	1.821***
	(0.116)	(0.114)	(0.114)	(0.107)	(0.126)	(0.219)	(0.384)	(0.830)	(0.279)	(0.306)
Japan	6.913***	8.206***	7.396***	5.815***	3.929***	5.905***	8.430***	10.947***	6.205***	6.974***
	(0.123)	(0.139)	(0.142)	(0.140)	(0.173)	(0.266)	(0.331)	(0.472)	(0.270)	(0.298)
Netherlands	4.495***	4.536***	4.342***	3.597***	2.623***	-5.134***	-0.140	11.073***	-3.143***	-4.005***
	(0.146)	(0.144)	(0.142)	(0.134)	(0.168)	(0.277)	(0.537)	(1.109)	(0.427)	(0.449)
New Zealand	-2.406***	-2.292***	-2.156***	-2.639***	-3.307***	-7.161***	-8.805***	0.000	-5.986***	-4.502***
	(0.135)	(0.134)	(0.133)	(0.124)	(0.152)	(0.230)	(0.483)	(0.000)	(0.300)	(0.384)
Norway	-5.760***	-5.594***	-6.168***	-6.020***	-6.553***	-6.002***	-18.496***	0.000	-7.456***	-12.705***
	(0.136)	(0.138)	(0.137)	(0.131)	(0.164)	(0.368)	(1.100)	(0.000)	(0.438)	(0.957)
United Kingdom	0.291**	-0.153	-0.461***	-0.822***	-0.934***	-5.414***	-4.851***	5.690***	-3.710***	-3.947***
	(0.135)	(0.133)	(0.132)	(0.124)	(0.138)	(0.203)	(0.431)	(0.465)	(0.344)	(0.346)
Child and family characteristics										
Child's age (months)		2.959***	2.673***	2.039***	1.948***	1.671***	1.631***	1.633***	1.644***	1.637***
		(0.137)	(0.135)	(0.125)	(0.123)	(0.122)	(0.122)	(0.122)	(0.123)	(0.123)
Child's age (squared)		-0.012***	-0.011***	-0.008***	-0.008***	-0.007***	-0.006***	-0.006***	-0.006***	-0.006***
		(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Girl		-0.561***	-0.584***	-1.122***	-1.087***	-1.089***	-1.088***	-1.089***	-1.087***	-1.086***
		(0.074)	(0.073)	(0.069)	(0.068)	(0.067)	(0.067)	(0.067)	(0.067)	(0.067)
Number of family members		-0.789***	-0.672***	-0.523***	-0.477***	-0.459***	-0.453***	-0.452***	-0.458***	-0.458***
		(0.026)	(0.026)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)
Child of immigrants			-0.350***	-0.157*	-0.125	-0.081	-0.115	-0.131	-0.093	-0.074
			(0.102)	(0.095)	(0.095)	(0.094)	(0.094)	(0.094)	(0.094)	(0.094)
Does not always speak test language at home			-4.641***	-3.169***	-2.826***	-2.814***	-2.847***	-2.840***	-2.813***	-2.838***
			(0.133)	(0.125)	(0.124)	(0.123)	(0.122)	(0.122)	(0.123)	(0.123)
Child does jobs at home (no time omitted)										
Less than 1 hour				1.494***	1.449***	1.456***	1.455***	1.452***	1.452***	1.456***
				(0.101)	(0.100)	(0.099)	(0.099)	(0.099)	(0.099)	(0.099)
1–2 hours				0.349***	0.360***	0.356***	0.365***	0.363***	0.348***	0.354***
				(0.117)	(0.115)	(0.114)	(0.114)	(0.114)	(0.114)	(0.114)
More than 2 hours				-2.524***	-2.360***	-2.317***	-2.309***	-2.309***	-2.322***	-2.321***
				(0.135)	(0.133)	(0.131)	(0.131)	(0.131)	(0.131)	(0.131)

Table 3 (contd.): OLS Regression of Mathematics Scores without Interactions

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Child reads books for enjoyment (no time omitted)										
Less than 1 hour				1.781***	1.717***	1.806***	1.766***	1.764***	1.798***	1.791***
				(0.095)	(0.093)	(0.092)	(0.092)	(0.092)	(0.092)	(0.092)
1–2 hours				2.941***	2.877***	2.987***	2.940***	2.937***	2.982***	2.974***
				(0.111)	(0.109)	(0.108)	(0.108)	(0.108)	(0.108)	(0.108)
More than 2 hours				2.334***	2.340***	2.478***	2.424***	2.423***	2.471***	2.457***
				(0.129)	(0.127)	(0.126)	(0.125)	(0.125)	(0.126)	(0.126)
Fewer than 25 books at home				-4.015***	-3.650***	-3.523***	-3.509***	-3.504***	-3.524***	-3.528***
				(0.088)	(0.087)	(0.087)	(0.086)	(0.086)	(0.087)	(0.087)
No calculator at home				-3.590***	-3.349***	-3.511***	-3.552***	-3.555***	-3.537***	-3.530***
				(0.138)	(0.136)	(0.135)	(0.135)	(0.135)	(0.135)	(0.135)
No computer at home				-1.610***	-1.459***	-1.664***	-1.521***	-1.512***	-1.623***	-1.611***
				(0.093)	(0.092)	(0.092)	(0.093)	(0.093)	(0.092)	(0.092)
No study desk at home				-1.659***	-1.416***	-1.305***	-1.340***	-1.347***	-1.296***	-1.297***
				(0.106)	(0.105)	(0.104)	(0.103)	(0.104)	(0.103)	(0.103)
No dictionary at home				-2.106***	-1.956***	-2.102***	-2.091***	-2.098***	-2.137***	-2.111***
				(0.109)	(0.108)	(0.108)	(0.108)	(0.108)	(0.108)	(0.108)
Teacher and school characteristics										
Teacher's age (younger than 30 omitted)										
Old age (50+)					-0.288*	-0.245	-0.290*	-0.295*	-0.262	-0.265
					(0.166)	(0.164)	(0.164)	(0.164)	(0.164)	(0.164)
Male teacher					0.055	-0.027	-0.065	-0.058	-0.020	-0.044
					(0.082)	(0.081)	(0.081)	(0.081)	(0.081)	(0.081)
Teacher's education (secondary school and lower omitted)										
BA/equivalent					-0.104	-0.107	-0.114	-0.090	-0.084	-0.152
					(0.104)	(0.103)	(0.103)	(0.104)	(0.103)	(0.103)
MA/PhD					0.426***	0.382***	0.275**	0.286**	0.381***	0.319**
					(0.134)	(0.133)	(0.133)	(0.133)	(0.133)	(0.133)
Years of teaching					0.026***	0.032***	0.034***	0.034***	0.033***	0.032***
					(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Teacher's class size					-0.050***	-0.045***	-0.043***	-0.043***	-0.041***	-0.042***
					(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
School in urban area					0.265***	0.392***	0.515***	0.518***	0.382***	0.422***
					(0.087)	(0.086)	(0.087)	(0.087)	(0.086)	(0.086)
More than 5% of students absent from school					-1.403***	-1.253***	-1.242***	-1.241***	-1.247***	-1.248***
					(0.098)	(0.097)	(0.097)	(0.097)	(0.097)	(0.097)

Table 3 (contd.): OLS Regression of Mathematics Scores without Interactions

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Percentage of students from										
disadvantaged families at school										
(0–10% omitted)										
11–50%					-2.050***	-1.870***	-2.024***	-2.033***	-1.873***	-1.905***
					(0.102)	(0.101)	(0.102)	(0.102)	(0.101)	(0.101)
More than 50%					-4.272***	-4.599***	-4.839***	-4.848***	-4.600***	-4.651***
					(0.132)	(0.131)	(0.134)	(0.134)	(0.131)	(0.132)
National average education and social										
expenditures (per head PPP US dollars)										
Preschool education expenditures						0.007***	0.013***	0.011***	0.008***	0.009***
						(0.000)	(0.001)	(0.001)	(0.000)	(0.000)
Primary education expenditures						-0.003***	-0.003***	-0.003***	-0.003***	-0.003***
						(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Expenditures on family cash benefits							0.014***	0.008**		
during preschool years							(0.001)	(0.004)		
Expenditures on health							-0.000	-0.010***		
during preschool years							(0.001)	(0.002)		
Other social expenditures							-0.006***	-0.003***		
during preschool years							(0.000)	(0.001)		
Expenditures on family cash benefits								-0.007*		
during primary school years								(0.004)		
Expenditures on health								0.021***		
during primary school years								(0.001)		
Other social expenditures								-0.005***		
during primary school years								(0.001)		
Total social expenditures									-0.001***	-0.003***
during preschool years									(0.000)	(0.000)
Total social expenditures										0.002***
during primary school years										(0.000)
Constant	99.604***	-81.640***	-63.441***	-26.469***	-18.606**	14.848*	25.341***	14.792*	18.246**	15.298**
	(0.079)	(8.463)	(8.337)	(7.760)	(7.659)	(7.638)	(7.726)	(7.850)	(7.655)	(7.668)
Observations	62294	62294	62294	62294	62294	62294	62294	62294	62294	62294
R-squared	0.12	0.15	0.18	0.29	0.31	0.33	0.33	0.33	0.33	0.33

Notes: Standard errors in parentheses.
* Marginally significant at 10%; ** significant at 5%; *** significant at 1%.

As a robustness check, we estimated four additional models. We added detailed controls for other social expenditures during the preschool years (Model 7), detailed controls for other social expenditures during both the preschool and primary school years (Model 8), detailed controls for total other social expenditures during the preschool years (Model 9), and detailed controls for total other expenditures during the preschool and primary school years. These expenditure variables were fairly highly correlated, so it may not have been appropriate to control for all of them in the same model. However, if we had omitted these controls, our estimates of the effects of preschool expenditures may have been biased. Therefore, we decided not to emphasize results from one model but instead to consider how the preschool coefficient changed across models. In the models with detailed controls for other expenditures, the effect of a $100 per child increase in preschool expenditures increased, to .13 of a standard deviation in Model 7 and .11 of a standard deviation in Model 8. When we added controls for total other expenditures during the preschool years or preschool and primary school years, the $100 per child increase in preschool fell again, to .08 (Model 9) and .09 (Model 10) of a standard deviation.

Table 4 presents the results for children's science test scores. Rather than discuss these results in detail, we focus our attention on the effect of public preschool expenditures, our main independent variable of interest. As with the mathematics scores, we found that public preschool expenditures had small but statistically significant effects on fourth-graders' science scores. These effects were considerably smaller than they were for the mathematics scores. They ranged from a coefficient of .003 (Model 6) to a coefficient of .007 (Model 7), implying that a $100 per child increase in expenditures is associated with an increase of between .03 and .07 of a standard deviation in science test scores.

We now turn to the question of whether the effects of preschool expenditures are larger for more disadvantaged children. Here, we considered seven distinct measures of disadvantage: being a child of immigrant parent(s), not always speaking the test language at home, having fewer than 25 books in the home, not having a calculator, not having a computer, not having a study desk, and not having a dictionary. We began with our simplest preschool expenditure model from Tables 3 and 4 (Model 6) and tested for differential effects of preschool expenditures for these groups by adding interaction terms to the model.[2] Thus, when testing for larger (or smaller) effects for children of immigrants, we included (in addition to the main effect for being a child of immigrants and the control for public preschool expenditures per child) an interaction term for being a child of immigrants and the level of public preschool expenditures per child. The coefficient on the interaction term told us whether the effect of public preschool expenditures differed significantly between that for children of immigrants and that for other children. A similar logic applied when we used interaction terms for our other measures of disadvantage.

The results for the mathematics scores appear in the first three columns of Table 5. The table provides mixed evidence for the hypothesis that public preschool expenditures benefit disadvantaged children more than other children. With regard to our first indicator of disadvantage (being a child of immigrants), we can see a negative interaction effect, indicating that public preschool expenditures had less of an effect for the children of immigrants than for the other children. The reason may be because children of immigrants were less likely to have participated in preschool programs (because they were not resident in the country during their preschool years or because they did not attend such programs). However, for our second indicator, we can see from the table a positive interaction between not always speaking the test language and preschool expenditures in the science test score model, providing some support for the idea that this group derived greater benefit from preschool spending. Across the other five indicators, all of which have to do with resources in the home, it is apparent that of 10 possible interaction terms between low resources and preschool expenditures, five are positive and statistically significant or marginally significant, and the other five are not significant. These results lend some support to the idea that public preschool spending may confer greater advantages to children from homes with relatively low levels of resources.

[2] We also estimated interaction models based on Models 7 to 10, where we interacted the group indicator (e.g., child of immigrant) with each of the expenditure variables. In results not shown but available on request, the effects of the interaction between the group indicators and the public preschool expenditures were basically the same as in the results presented in Table 5. Because the results became quite complex with so many interaction terms, we show only the results from Model 6.

Table 4: OLS Regression of Science Scores without Interactions

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Wave of data (wave 2003 omitted)										
Wave 1995	-0.467***	0.144*	-0.080	-0.348***	-0.443***	-2.909***	-3.309***	-0.916	-4.477***	-4.514***
	(0.080)	(0.081)	(0.082)	(0.082)	(0.099)	(0.149)	(0.381)	(0.822)	(0.258)	(0.259)
Participating countries (the US is the omitted base case)										
Australia	1.005***	0.914***	0.585***	-0.110	-0.868***	-1.808***	-2.082***	0.745	-0.477	-0.213
	(0.121)	(0.119)	(0.117)	(0.111)	(0.130)	(0.228)	(0.399)	(0.862)	(0.290)	(0.318)
Japan	2.479***	3.877***	2.613***	1.277***	-0.771***	-1.055***	1.587***	4.179***	-0.677**	-0.416
	(0.128)	(0.144)	(0.147)	(0.144)	(0.179)	(0.276)	(0.344)	(0.491)	(0.281)	(0.309)
Netherlands	0.466***	0.401***	0.060	-0.606***	-1.619***	-6.353***	0.461	10.908***	-3.850***	-4.143***
	(0.152)	(0.150)	(0.147)	(0.139)	(0.174)	(0.288)	(0.558)	(1.152)	(0.444)	(0.467)
New Zealand	-1.000***	-0.864***	-0.820***	-1.314***	-2.338***	-6.009***	-5.041***	0.000	-4.532***	-4.028***
	(0.141)	(0.139)	(0.137)	(0.128)	(0.157)	(0.239)	(0.502)	(0.000)	(0.311)	(0.399)
Norway	-5.624***	-5.530***	-6.308***	-6.087***	-6.534***	-3.486***	-13.043***	0.000	-5.313***	-7.096***
	(0.142)	(0.144)	(0.141)	(0.135)	(0.169)	(0.383)	(1.143)	(0.000)	(0.455)	(0.995)
United Kingdom	1.058***	0.526***	0.002	-0.361***	-0.574***	-4.224***	-1.389***	6.287***	-2.082***	-2.162***
	(0.141)	(0.139)	(0.136)	(0.128)	(0.143)	(0.211)	(0.448)	(0.483)	(0.358)	(0.360)
Child and family characteristics										
Child's age (months)		2.746***	2.385***	1.765***	1.648***	1.489***	1.441***	1.445***	1.455***	1.453***
		(0.142)	(0.139)	(0.129)	(0.128)	(0.127)	(0.127)	(0.127)	(0.127)	(0.127)
Child's age (squared)		-0.011***	-0.009***	-0.007***	-0.006***	-0.006***	-0.006***	-0.006***	-0.006***	-0.006***
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Girl		-0.706***	-0.743***	-1.301***	-1.266***	-1.259***	-1.262***	-1.264***	-1.257***	-1.257***
		(0.077)	(0.075)	(0.071)	(0.070)	(0.070)	(0.070)	(0.070)	(0.070)	(0.070)
Number of family members		-1.022***	-0.857***	-0.711***	-0.662***	-0.651***	-0.641***	-0.640***	-0.650***	-0.650***
		(0.027)	(0.027)	(0.025)	(0.025)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)
Child of immigrants			-0.867***	-0.664***	-0.564***	-0.531***	-0.572***	-0.606***	-0.546***	-0.540***
			(0.105)	(0.099)	(0.098)	(0.097)	(0.097)	(0.098)	(0.097)	(0.097)
Does not always speak test language			-6.113***	-4.654***	-4.268***	-4.279***	-4.298***	-4.284***	-4.278***	-4.286***
at home			(0.137)	(0.129)	(0.128)	(0.127)	(0.127)	(0.127)	(0.127)	(0.127)
Child does jobs at home (no time omitted)										
Less than 1 hour				1.201***	1.146***	1.140***	1.139***	1.134***	1.134***	1.136***
				(0.105)	(0.103)	(0.103)	(0.103)	(0.102)	(0.103)	(0.103)
1–2 hours				0.028	0.029	0.017	0.038	0.034	0.007	0.009
				(0.121)	(0.119)	(0.119)	(0.119)	(0.119)	(0.119)	(0.119)
More than 2 hours				-2.556***	-2.389***	-2.386***	-2.352***	-2.353***	-2.391***	-2.391***
				(0.139)	(0.137)	(0.137)	(0.137)	(0.136)	(0.137)	(0.137)

Table 4 (contd.): OLS Regression of Science Scores without Interactions

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Child reads books for enjoyment (no time omitted)										
Less than 1 hour				1.699***	1.632***	1.707***	1.657***	1.653***	1.696***	1.694***
				(0.098)	(0.096)	(0.096)	(0.096)	(0.096)	(0.096)	(0.096)
1–2 hours				3.188***	3.129***	3.224***	3.171***	3.165***	3.218***	3.215***
				(0.114)	(0.113)	(0.112)	(0.112)	(0.112)	(0.112)	(0.112)
More than 2 hours				3.249***	3.255***	3.355***	3.300***	3.297***	3.347***	3.342***
				(0.133)	(0.131)	(0.131)	(0.130)	(0.130)	(0.130)	(0.131)
Fewer than 25 books at home				-4.245***	-3.871***	-3.791***	-3.774***	-3.763***	-3.792***	-3.793***
				(0.091)	(0.090)	(0.090)	(0.090)	(0.090)	(0.090)	(0.090)
No calculator at home				-3.619***	-3.377***	-3.529***	-3.624***	-3.631***	-3.563***	-3.560***
				(0.142)	(0.141)	(0.140)	(0.140)	(0.140)	(0.140)	(0.140)
No computer at home				-1.730***	-1.554***	-1.735***	-1.539***	-1.518***	-1.683***	-1.679***
				(0.096)	(0.095)	(0.095)	(0.096)	(0.096)	(0.096)	(0.096)
No study desk at home				-1.450***	-1.211***	-1.157***	-1.208***	-1.223***	-1.146***	-1.146***
				(0.109)	(0.108)	(0.108)	(0.108)	(0.108)	(0.108)	(0.108)
No dictionary at home				-2.071***	-1.940***	-1.948***	-1.940***	-1.955***	-1.993***	-1.984***
				(0.113)	(0.112)	(0.112)	(0.112)	(0.112)	(0.112)	(0.112)
Teacher and school characteristics										
Teacher's age (younger than 30 omitted)										
Middle age (30–49)					0.214*	0.287**	0.230**	0.228**	0.274**	0.270**
					(0.114)	(0.113)	(0.113)	(0.113)	(0.113)	(0.113)
Old age (50+)					-0.132	-0.078	-0.111	-0.123	-0.098	-0.099
					(0.172)	(0.171)	(0.171)	(0.171)	(0.171)	(0.171)
Male teacher					-0.166**	-0.194**	-0.249***	-0.234***	-0.185**	-0.194**
					(0.085)	(0.084)	(0.084)	(0.084)	(0.084)	(0.084)
Teacher's education (secondary school and lower omitted)										
BA/equivalent					-0.192*	-0.187*	-0.087	-0.033	-0.157	-0.181*
					(0.107)	(0.107)	(0.107)	(0.108)	(0.107)	(0.107)
MA/PhD					0.286**	0.238*	0.172	0.196	0.237*	0.216
					(0.138)	(0.138)	(0.138)	(0.138)	(0.138)	(0.138)
Years of teaching					0.023***	0.025***	0.027***	0.028***	0.026***	0.026***
					(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Teacher's class size					-0.048***	-0.042***	-0.037***	-0.038***	-0.037***	-0.037***
					(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
School in urban area					-0.157*	-0.038	0.094	0.100	-0.051	-0.038
					(0.090)	(0.089)	(0.090)	(0.090)	(0.089)	(0.090)

Table 4 (contd.): OLS Regression of Science Scores without Interactions

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
More than 5% of students absent					-1.502***	-1.392***	-1.372***	-1.371***	-1.385***	-1.385***
at school					(0.101)	(0.101)	(0.101)	(0.101)	(0.101)	(0.101)
Percentage of students from disadvantaged families at school (0–10% omitted)										
11–50%					-1.856***	-1.748***	-1.936***	-1.956***	-1.753***	-1.764***
					(0.105)	(0.105)	(0.106)	(0.106)	(0.105)	(0.105)
More than 50%					-4.298***	-4.488***	-4.823***	-4.843***	-4.488***	-4.506***
					(0.137)	(0.137)	(0.139)	(0.139)	(0.136)	(0.137)
National average education and social expenditures (per head PPP US dollars)										
Preschool education expenditures						0.003***	0.007***	0.006***	0.004***	0.005***
						(0.000)	(0.001)	(0.001)	(0.000)	(0.000)
Primary education expenditures						-0.002***	-0.002***	-0.002***	-0.002***	-0.002***
						(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Expenditures on family cash benefits during							0.011***	0.004		
preschool years							(0.001)	(0.004)		
Expenditures on health during							0.004***	-0.001		
preschool years							(0.001)	(0.002)		
Other social expenditures during							-0.006***	-0.003***		
preschool years							(0.000)	(0.001)		
Expenditures on family cash benefits during								0.001		
primary school years								(0.005)		
Expenditures on health during								0.015***		
primary school years								(0.001)		
Other social expenditures during								-0.007***		
primary school years								(0.001)		
Total social expenditures during									-0.002***	-0.002***
preschool years									(0.000)	(0.000)
Total social expenditures during										0.001**
primary school years										(0.000)
Constant	100.236***	-67.095***	-44.083***	-8.005	1.827	23.972***	28.169***	21.732***	28.243***	27.241***
	(0.082)	(8.794)	(8.591)	(8.023)	(7.918)	(7.940)	(8.029)	(8.157)	(7.958)	(7.973)
Observations	62294	62294	62294	62294	62294	62294	62294	62294	62294	62294
R-squared	0.05	0.09	0.13	0.24	0.27	0.27	0.28	0.28	0.27	0.27

Notes: Standard errors in parentheses.
* Marginally significant at 10%; ** significant at 5%; *** significant at 1%.

Table 5: OLS Regression with Interactions between Disadvantaged Factors and Education Expenditures

		Mathematics scores			Science scores		
			Preschool expenditures	Primary school expenditures		Preschool expenditures	Primary school expenditures
Child of immigrants	Variable	-0.644*	0.007***	-0.003***	-0.130	0.003***	-0.002***
		(0.339)	(0.000)	(0.000)	(0.352)	(0.000)	(0.000)
	Interaction	-	-0.002***	0.000***	-	-0.001***	0.000*
			(0.000)	(0.000)		(0.000)	(0.000)
Does not always speak test language at at home	Variable	-1.908***	0.006***	-0.003***	-2.964***	0.002***	-0.002***
		(0.433)	(0.000)	(0.000)	(0.450)	(0.000)	(0.000)
	Interaction	-	0.000	-0.000	-	0.001***	-0.000***
			(0.000)	(0.000)		(0.000)	(0.000)
Fewer than 25 books at home	Variable	-3.712***	0.006***	-0.003***	-4.020***	0.003***	-0.002***
		(0.309)	(0.000)	(0.000)	(0.321)	(0.000)	(0.000)
	Interaction	-	0.000	0.000	-	0.000	-0.000
			(0.000)	(0.000)		(0.000)	(0.000)
No calculator at home	Variable	-2.911***	0.007***	-0.003***	-1.387***	0.003***	-0.002***
		(0.505)	(0.000)	(0.000)	(0.524)	(0.000)	(0.000)
	Interaction	-	0.002***	-0.000***	-	0.003***	-0.001***
			(0.000)	(0.000)		(0.000)	(0.000)
No computer at home	Variable	-1.389***	0.007***	-0.003***	-1.545***	0.003***	-0.002***
		(0.349)	(0.000)	(0.000)	(0.363)	(0.000)	(0.000)
	Interaction	-	0.000	-0.000	-	0.000*	-0.000
			(0.000)	(0.000)		(0.000)	(0.000)
No study desk at home	Variable	-0.703*	0.007***	-0.003***	-1.102***	0.003***	-0.002***
		(0.401)	(0.000)	(0.000)	(0.417)	(0.000)	(0.000)
	Interaction	-	-0.000	-0.000	-	-0.000	0.000
			(0.000)	(0.000)		(0.000)	(0.000)
No dictionary at home	Variable	-1.593***	0.007***	-0.003***	-1.399***	0.003***	-0.002***
		(0.374)	(0.000)	(0.000)	(0.389)	(0.000)	(0.000)
	Interaction	-	0.001***	-0.000**	-	0.001***	-0.000**
			(0.000)	(0.000)		(0.000)	(0.000)

Notes: Standard errors in parentheses.
* Marginally significant at 10%; ** significant at 5%; *** significant at 1%.

Conclusions

This study examined the effects of public preschool expenditures on the mathematics and science scores of fourth-graders, with child, family, and school characteristics, other relevant social expenditures, and country and year effects held constant. We used data from the 1995 and 2003 Trends in International Mathematics and Science Study (TIMSS) for children from seven OECD countries—Australia, Japan, the Netherlands, New Zealand, Norway, the United Kingdom, and the United States. We also explored whether preschool expenditures matter more for children at risk of poor school achievement, as indexed by having low levels of resources in the home or coming from an immigrant family or a family that did not always speak the test language.

Our results showed small but significant positive effects of public preschool expenditures on the mathematics and science scores of fourth-graders. We found that an increase in preschool expenditures of $100 per child lifted children's mathematics scores by .07 to .13 of a standard deviation, and raised their science scores by .03 to .07 of a standard deviation. These estimates were somewhat sensitive to how we controlled for other social expenditures, but in each model remained positive and statistically significant.

We also found some evidence that children from low-resource homes may gain more from increased public preschool expenditures than may other children, but that children of immigrants might gain less (perhaps because they are less likely to attend such programs). Thus, this study provides new evidence that increasing public preschool expenditures raises children's mathematics and science achievement but mixed evidence as to the role of such expenditures in helping to close gaps in achievement between less and more advantaged students.

A key concern in this study was controlling for unobserved heterogeneity across countries that may correlate with both public spending and student outcomes. We attempted to address this heterogeneity by using multiple waves of data and including country fixed effects as well as year effects. However, as noted earlier, our analyses were limited in that we had only two waves of data per country. Thus, we cannot be certain that we controlled for all the factors that vary across countries and that may matter for student achievement as well for as public spending. A further limitation is that TIMSS contains few controls for key family background variables such as parental education or for the quality of preschool or school programs that children attend. Clearly, there are many factors consequential for student achievement that we were not able to control for in these data.

Despite these shortcomings, it is striking how consistent our results are across the models in pointing to a small but significant positive association between public preschool expenditures and higher student mathematics and science scores in Grade 4. These results suggest that public preschool expenditures may play a role in raising children's mathematics and science achievement. The exact magnitude of these effects, how they occur, and whether and how they vary across different groups of children, are all excellent topics for further research.

References

Bainbridge, J., Meyers, M., Tanaka, S., & Waldfogel, J. (2005). Who gets an early education? Family income and the gaps in enrollment of 3–5 year olds from 1968–2000. *Social Science Quarterly, 86*(4), 724–745.

Barnett, W. S., Lamy, C., & Jung, K. (2005). *The effects of state pre-kindergarten programs on young children's school readiness in five states.* Available from the National Institute for Early Education Research at www.nieer.org

Blau, D. M. (2001). *The child care problem: An economic analysis.* New York: Russell Sage Foundation Press.

Brooks-Gunn, J. (1995). Strategies for altering the outcomes of poor children and their families. In P. L. Chase-Lansdale & J. Brooks-Gunn (Eds.), *Escape from poverty: What makes a difference for children?* (pp. 87–117). New York: Cambridge University Press.

Brooks-Gunn, J., & Markman, L. B. (2005). The contribution of parenting to ethnic and racial gaps in school readiness. *The Future of Children, 15*(1), 139–168.

Currie, J. (2001). Early childhood intervention programs: What do we know? *Journal of Economic Perspectives, 15*, 213–238.

Currie, J., & Thomas, D. (1995). Does Head Start make a difference? *American Economic Review, 85*(3), 341–364.

Garces, E., Thomas, D., & Currie, J. (2002). Long-term effects of Head Start. *The American Economic Review, 92*(4), 999–1012.

Gormley, W., & Gayer, T. (2005). Promoting school readiness in Oklahoma: An evaluation of Tulsa's pre-K program. *Journal of Human Resources, 40*, 533–558.

Gormley, W., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology, 41*(6), 872–884.

Hampden-Thompson, G., & Johnston, J. S. (2006). *Variation in the relationship between non-school factors and student achievement on international assessments* (Statistics in Brief, NCES 2006-014, April 2006). Washington, DC: Institute of Education Sciences, National Center for Education Statistics.

Hanushek, E. (2006). Class size and resources. In E. Hanushek & F. Welch (Eds.), *Handbook of the economics of education.* Amsterdam: Elsevier.

Heckman, J. J., & Krueger, A. B. (2003). *Inequality in America: What role for human capital policies?* Cambridge, MA: MIT Press.

Kamerman, S. B., & Waldfogel, J. (2005). Market and non-market institutions in early childhood education and care. In R. Nelson (Ed.), *Market and non-market institutions.* New York: Russell Sage Foundation Press.

Leventhal, T., & Brooks-Gunn, J. (2002). Poverty and child development. *The International Encyclopedia of the Social and Behavioral Sciences, 3*, Article 78, 11889–11893.

Magnuson, K., Meyers, M., Ruhm, C., & Waldfogel, J. (2004). Inequality in preschool education and school readiness. *American Educational Research Journal, 41*(1), 115–157.

Magnuson, K., Ruhm, C., & Waldfogel, J. (in press). Does prekindergarten improve school preparation and performance? *Economics of Education Review.*

Magnuson, K., & Waldfogel, J. (2005). Child care, early education, and racial/ethnic test score gaps at the beginning of school. *The Future of Children, 15*(1), 169–196.

McLanahan, S., & Casper, L. (1995). Growing diversity and inequality in the American family. In R. Farley (Ed.), *State of the union: America in the 1990s* (pp. 6–13). New York: Russell Sage Foundation Press.

Meyers, M., Rosenbaum, D., Ruhm, C., & Waldfogel, J. (2004). Inequality in early childhood education and care: What do we know? In K. Neckerman (Ed.), *Social inequality.* New York: Russell Sage Foundation Press.

Organisation for Economic Co-operation and Development (OECD). (2001). *Starting strong: Early childhood education and care.* Paris, France: OECD Publications.

Organisation for Economic Co-operation and Development (OECD). (2005a). *OECD education database.* Available online at http://www1.oecd.org/scripts/cde/members/linkpage.html

Organisation for Economic Co-operation and Development (OECD). (2005b). *OECD health data 2005: Statistics and indicators for 30 countries* (Version 06/06/2005). Paris: Author.

Puma, M., Bell, S., Cook, R., Heid, C., & Lopez, M. (2005). *Head Start impact study: First year findings.* Washington, DC: United States Department of Health and Human Services, Administration for Children and Families.

Ruhm, C. J. (2000). Parental leave and child health. *Journal of Health Economics, 19*, 931–961.

Smolensky, E., & Gootman, J. A. (Eds.). (2003). *Working families and growing kids: Caring for children and adolescents* (Committee on Family and Work Policies, National Research Council, Institute of Medicine of the National Academies). Washington, DC: The National Academies Press.

Tanaka, S. (2005). Parental leave and child health across OECD countries. *The Economic Journal, 115*(February), F7–F28.

Waldfogel, J. (2006a). Early childhood policy: A comparative perspective. In K. McCartney & D. Phillips (Eds.), *The handbook of early childhood development* (pp. 576–594). Malden, MA: Blackwell Publishers.

Waldfogel, J. (2006b). *What children need.* Cambridge, MA: Harvard University Press.

TIMSS versus PISA: The case of pure and applied mathematics[1]

Liv Sissel Grønmo and Rolf Vegar Olsen
University of Oslo
Norway

Abstract

The framework, aims, and populations of the IEA Trends in International Mathematics and Science Study (TIMSS) (Grade 8 level) and of the OECD Programme for International Students Assessment (PISA) differ. This consideration provides greater opportunity for researchers to gain more knowledge and insight into the education systems of different countries than one study alone can offer. Several countries participated in both studies in 2003. Comparison of the achievement score ranks of the participating countries in the two studies shows that a group of countries, particularly some Nordic and English-speaking ones, performed relatively better in PISA than in TIMSS. However, the East-European countries performed relatively much better in TIMSS. This study presents an analysis of the mathematical coverage in the two studies in order to understand these shifts in rank. The findings of our analyses are twofold. First, the assessment frameworks are formulated from largely different perspectives on mathematics. While PISA describes in detail the contexts and phenomena where mathematical competence may be of importance, TIMSS gives a very fine-grained definition of some important aspects of mathematics from within the discipline itself. Second, the items in PISA emphasize the use of numerical information in the form of tables and graphs taken from real-world contexts, while items in TIMSS give much more attention to pure mathematics, including formal aspects of algebra and geometry. We also considered country characteristic profiles across major categories in TIMSS and PISA for five selected countries. Based on these results, we discuss the relationship between pure and applied mathematics in school, and conclude that, to do well in applied mathematics, students need to have a basis in elementary knowledge and skills in pure mathematics. For some countries, like Norway, what seems most problematic is that students lack elementary knowledge and skills in a topic such as Number.

Introduction

TIMSS and PISA are regarded as highly similar types of study. They are large-scale surveys with a very similar methodological basis. For example, they

- Are sample-based studies of clearly defined populations
- Apply the same type of instruments (e.g., student questionnaire and cognitive booklets)
- Process the data with similar psychometrical methods
- Are governed by a consensus-driven process from initial ideas to final instruments
- Enforce rigorous quality control, for example, of translation or adaptation of the test material
- Have cyclic designs with a focus on measuring trends.

Furthermore, both studies include measurements of highly related constructs, including mathematical and scientific competency, and student and school background characteristics and attitudes. However, we use terms like "similar" and "same" rather than "equal," identical," or "equivalent." There are important differences between the studies, and we highlight some of these in this paper.

The problem of this apparent similarity between the studies was particularly evident in December 2004. Both TIMSS and PISA published the results from their studies (conducted in 2003) within a few days of this month (Martin, Mullis, Gonzalez, & Chrostowski, 2004; Mullis, Martin, Gonzalez, & Chrostowski, 2004; OECD, 2004). Both studies were followed up with intensity in the media for a period of time, with media reports regularly referring to both with statements like "An international study of mathematics achievement shows that students in our country ...". Moreover, most countries referred to both studies as involving students at the same educational level—(lower) secondary school. It is important to communicate that the key terms "international study,"

[1] We would like to thank our colleague Inger Throndsen for her contribution to the analysis presented in this paper. In particular, she helped us to categorize TIMSS items into the PIRLS framework.

"mathematics," and "students" are not synonymous when referring to PISA and TIMSS respectively. In this paper, we compare and contrast these studies, and particularly discuss how the two studies have differently operationalized the keyword "mathematics." Although the case of Norway is central in the paper (Grønmo, Kjærnsli, & Lie, 2004; Kjærnsli, Lie, Olsen, Roe, & Turmo, 2004), the overall relevance and purpose of what is stated here goes beyond a national context.

Relevance of the study

There are several reasons why a comparison such as ours is relevant:

- It emphasizes how the monitoring of educational systems[2] by international comparative studies may be done from different perspectives, leading to different results and possibly different interpretations.
- The two studies represent two partly overlapping and partly diverging views on mathematics education. Furthermore, by comparing the two studies, one might also unravel or discover more tacit, or rarely spoken of, aspects of the design of the two studies.
- Several countries have participated in both studies and will be participating in future cycles of the studies. Of the 32 educational systems that took part in PISA 2000, 28 also participated in either TIMSS 1995 or TIMSS 1999, or both. In 2003, 22 educational systems took part in both studies. For researchers in mathematics education, the fact that many countries participated in both studies may yield new opportunities for studying international differences and similarities in mathematics education.

Others have also recognized the importance of analyzing the relationships between these studies. The Organisation for Economic Co-operation and Development (OECD) has commissioned a thematic report to compare PISA and TIMSS (not yet published), and in the United States context, the National Center for Education Statistics is also endeavouring to compare TIMSS and PISA, and furthermore to relate the studies to National Assessment for Educational Progress/NAEP studies (Hutchison & Schagen, in press; National Center for Education Statistics/NCES, 2005; Neidorf, Binkley, Gattis, & Nohara, 2006; Nohara, 2001).

In this paper, we present the results of comparisons of frameworks and items in TIMSS and PISA. We also present the achievement profiles across different content domains for five countries selected to represent the international diversity in achievement profiles. These results give us a background for discussing differences in mathematics education in different countries that is beyond the data offered by just one of the studies. A main topic in the discussion is the perceived tension between elementary mathematical knowledge and skills versus problem solving in mathematics, including a discussion of the relationship between pure and applied mathematics.

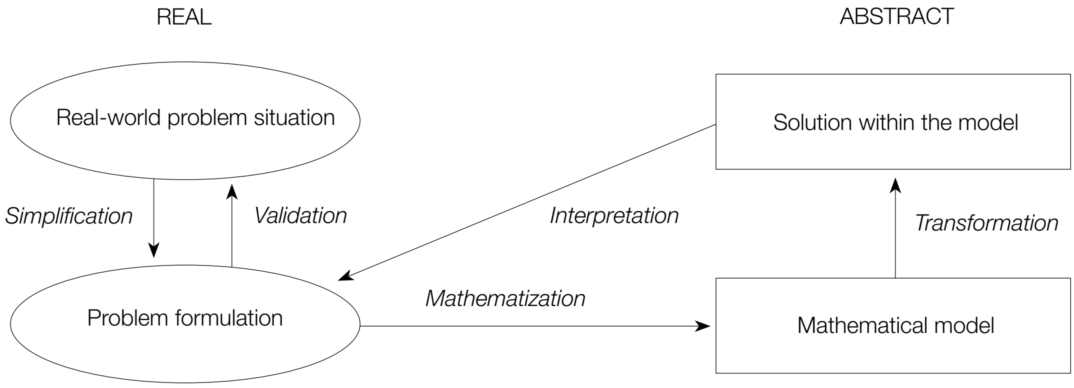
What constitutes school mathematics?

Since the Second World War, there has been ongoing discussion about what should constitute mathematics in school. This discussion has aligned, to a large extent, with the considerable effort and use of resources to develop education for all citizens in western societies (Ernest, 1991; Skovsmose, 1994). Central to this discussion has been the relationship between pure and applied mathematics in the school curriculum. There have been several "back to the basics" movements, particularly in the United States context. The discussion has been so heated that the label "mathematics wars" has been frequently used (Schoenfeld, 2004). We argue that the different operationalizations of mathematics in TIMSS and PISA reflect, to a large degree, this discussion about the relationship between mastery of basic facts and procedures in mathematics, on the one hand, and the ability to relate mathematics to real-world phenomena on the other.

Figure 1 presents a model of how a problem situation in the real world transforms into a mathematical problem, and how a solution subsequently is found and validated against the original real context in which the problem originated. The model is taken from a very influential United States policy document on standards in mathematics (National Council of Teachers of Mathematics, 1989). PISA has also adopted this model, in a slightly different form (OECD, 2003, p. 38). The right-hand side of Figure 1 represents the mathematical world—an abstract world with well-defined symbols and rules. The left-

[2] The participating educational systems generally are countries. However, some autonomous states or regions within some countries are also included. In the following, we use the term "country" to include all participating systems, both formal national states and autonomous regions with independent participation status in the projects.

Figure 1: The Mathematization Cycle



Source: National Council of Teachers of Mathematics (1989).

hand side represents the real concrete world; what we may call daily life. Working with pure mathematics, as numbers or as algebra, out of any context, means working only on the right-hand side of Figure 1. In applied mathematics, the starting point is supposed to be a problem from the real world, which first has to be simplified, and then mathematized into a model representing the problem. In most cases, school mathematics rarely starts with a real problem. Instead, the problem presented to the students has generally already been simplified.

It is obvious from the mathematization cycle in Figure 1 that a premise for any type of applied mathematics is that the students have the necessary knowledge in pure mathematics to find a correct mathematical solution. The students have to be able to orientate themselves in a pure mathematical world. Thus, applied mathematics can be seen as more complex than pure mathematics, if we agree with the premise that the same mathematics is involved in the two cases. Given the increasing focus that various countries place on applied mathematics in compulsory school, it seems that the importance of understanding that mathematics is an exact and well-defined science, and that one needs to be able to orientate in the world of pure mathematics, has been neglected to some extent. This neglect has met with criticism from some researchers. For example, Gardiner (2004) points out that even if mathematical literacy—the ability to use mathematics to solve daily life problems—is a main goal for school mathematics, this form of literacy should not be seen as an alternative to having basic knowledge and skills in pure mathematics.

The different ways that TIMSS and PISA operationalize mathematics reflect different opinions of what school mathematics is or what it should be. PISA tests students in the type of applied mathematics they may need in their daily life as citizens in a modern society; an aim for PISA is to embed all items in a context as close to a real-life problem as possible. Most of the items in TIMSS, however, are pure mathematical items with no context, or items with the simple and artificial contexts that are almost a signature of traditional mathematical problems in school. TIMSS therefore, to a considerable extent, gives information about students' knowledge in pure mathematics, while PISA displays students' knowledge in applied mathematics.

Our discussion of the results of our analysis of data from TIMSS and PISA relates to these considerations regarding the relationship between pure and applied mathematics, a relationship that differs markedly from how school curricula in various countries represent it. Our discussion does not involve analysis of these countries' formal mathematical curriculum—that is, the intended curriculum. The TIMSS framework (Mullis et al., 2001) involves a model containing three levels of the curriculum—the intended, the implemented, and the attained. Our discussion involves analyses of the attained curriculum in particular, that is, what students have achieved in terms of mathematical knowledge. We also use this model as an indicator of the focus of instruction in school, that is, of the implemented curriculum.

Comparison of the general features of TIMSS and PISA

TIMSS and PISA have many overlapping features in their design and administration. However, we need to address some important differences in order to produce meaningful comparisons between the two studies' definitions and their operationalizations of students' achievement in mathematics.

First, it is important to note that the two studies rest on two largely different visions of how to monitor educational outcomes. TIMSS is designed to specifically target what is perceived to be the best common core of mathematics curricula across the world. In other words, TIMSS aims to measure education systems according to the stated aims for those systems. Much effort therefore is expended on reviewing the items and frameworks in each country, to formulate a framework that represents this "best common core." The principle followed in PISA is rather different. Instead of basing the framework on the curriculum in the participating countries, PISA challenges leading scholars in mathematics education to define mathematics according to what students will need in their present life and in their future life. The label used, therefore, is mathematical literacy in order to distance the concept from what most people think of when they hear the term mathematics. For most people, mathematics refers to either a school subject and/or an academic discipline. PISA uses the term mathematical literacy to emphasize that the study aims to measure how well students can cope with situations where mathematical skills and knowledge are important.

However, in the public domain, it is convenient to use the term mathematics, and so the distinction between TIMSS and PISA is often not present in discussions of the results by journalists, policymakers, and even academic scholars. While it is reasonable to state that TIMSS measures how well an education system has implemented the mathematics curriculum, it is reasonable to claim that PISA is a better measure of how well the curriculum serves the general needs of students in their current and future lives as citizens. The policy implications we can draw from the two studies are therefore rather different in principle. While we can use TIMSS to study the implementation of the core curriculum, we can use PISA to evaluate if the curriculum is well suited to the task of equipping citizens with the mathematical competencies they will most likely need.

The visible effects of these two different perspectives is that mathematics, as defined, operationalized, and reported in TIMSS, relates much more to understanding fundamental concepts and procedures in mathematics than is the case in PISA. We can see this difference, for instance, in the very fine-grained and detailed framework, particularly for the content dimension. TIMSS has five content domains in mathematics, one of which is "Number." Each content domain is partitioned by a set of more specific objectives, for example, "Integers" (one of four objectives within "Number"). Finally, each objective is operationalized in terms of more specific statements, for example, "Compare and order integers" (one of five statements defining the objective "Integers") (Mullis et al., 2001, pp. 12–13). Thus, the defining statements of mathematics achievement describe very specific knowledge and skills. PISA defines mathematics with less rigor on the content dimension. Instead of defining different content domains within mathematics, PISA describes four classes of phenomena in which mathematical knowledge and reasoning is of great importance. These are labeled "Overarching Ideas." These four phenomenological aspects are not subdivided, but are instead defined by a coherent introductory text giving reference to this class of phenomena generally. In addition, the framework refers to a number of specific situations and many specific examples of items that could be used to assess each of the categories (OECD, 2003).

To summarize:

- The "operational distance" from framework to the specific items is narrower in TIMSS than in PISA.
- The content sub-domains in TIMSS are conceptually more isolated than are the overarching ideas in PISA.
- The TIMSS content sub-domains clearly demarcate a line between included and excluded content, while the PISA overarching ideas do not draw such a line.
- The TIMSS framework can be captured by the slogan, "What school mathematics is," while the PISA framework may be considered as "What school mathematics should be."
- The PISA framework gives more attention to the process dimension of using mathematics than does the TIMSS framework.

The specific focus of this article is on the content dimensions of the two studies (what type of mathematics each involves). We therefore do not discuss in detail the differences between the cognitive dimensions in the two studies. However, to provide a complete comparison of the two studies, we also need to study the differences along this process or cognitive dimension. Our first-hand experience of reviewing, marking, and analyzing mathematics items in the two studies showed us that the cognitive complexity of the PISA items is higher than that of TIMSS. Confirmation of this premise comes from the findings of a panel reviewing the items on behalf of the United States National Center for Education Statistics (NCES, 2005). Here, it was evident that a majority of the PISA items would be classified "Reasoning" in the TIMSS 2003 framework.

We have reason to believe that the cognitive dimension does not adequately reflect the variation between the profiles of the participating countries. This supposition is supported by the international study on scaling the TIMSS mathematics items across the sub-domains of the cognitive domain conducted by Mullis, Martin, and Foy (2005). They found that the variation within countries across the sub-domains of the cognitive domain was lower overall than the variation across the sub-domains of the content domain. The reasons for this finding might be several, but we believe that one of the main reasons is that the process or cognitive dimension is particularly hard to divide into distinctly different sub-classes. Persons confronted with the task of classifying a specific item along both the content and the cognitive domain sub-categories would most likely do so with a higher degree of agreement within the content domain, and a higher degree of disagreement in the cognitive domain. The reason why is probably because an item addresses different cognitive domains for different students. For example, the intention behind an item testing factual knowledge is to separate students who know this fact from those who do not. Nevertheless, students who do not know this fact might be able to reach the correct solution by using reasoning skills.

Another obstacle in creating meaningful scales across the sub-domains of the cognitive dimension relates to the inevitable conceptual hierarchy. Even if an item targets higher order cognitive skills, it inevitably also, to a large degree, includes important elements of lower order cognitive skills. The degree of hierarchy in the cognitive domain thus makes assessing which one of several mutually exclusive sub-categories an item belongs to very difficult. In our account of countries' profiles across different domains in mathematics, we therefore focus mainly on the profiles across the content domains.

In addition to measurement of different constructs, we consider that a number of other important differences between the two studies deserve brief mention. For a fuller account of these and other differences, see Olsen (2005).

- PISA targets students who are generally two years older than the Grade 8 TIMSS students. Furthermore, TIMSS has a grade-based population, while PISA has an age-based population. Thus, in TIMSS, the ages of the students may vary within and across countries, while in PISA, grades may vary within and across the countries.
- TIMSS samples classes within schools, while PISA samples students within schools. The data from each study consequently reflects different hierarchical structures.
- Through the questionnaires, the studies collect data on background variables used to supplement, augment, and explain the achievement measures. Because TIMSS aims to measure the implementation of the curriculum, the questionnaire gives heavier emphasis than does PISA to issues relating directly to what happens in the teaching of the subject matter. PISA, however, puts more effort than does TIMSS into measuring students' social backgrounds.
- The compositions of the participating countries in the two studies differ. The range of countries from different developmental levels is wider in TIMSS than in PISA, a situation that makes the international comparative backgrounds of the two studies very different.

Method

The particular aim of this study was to describe and account for some of the apparent non-consistent performance evident in some countries between the TIMSS Grade 8 mathematics assessment of 2003 and the PISA assessment of mathematical literacy conducted in 2003. Some countries seemed to perform relatively better in one of the studies. It is not possible to combine the measures of achievement in the two studies at the student or school level in any meaningful way; we cannot link the two measures in any formal sense. What we can do instead is link the two studies at the country or system level. In the analysis presented in this paper, we include only those countries/education systems with a reported average achievement score in both studies. The analysis accordingly includes 22 countries and regions. Although the ranks in the studies are only ordinal measures of achievement, they comprise the only "scale" common to both studies. We therefore use the rank orders from 1 to 22 for these countries in both studies, independently. The difference in rank provides a rough indicator of relative success in one of the studies as compared to the other.

Having done that, our next aim was to identify the possible sources internal to the studies themselves of the shifts. By "internal sources," we mean characteristic features of the items used to measure mathematics achievement in the two studies. More specifically, we mapped all PISA mathematics literacy items into the mathematics content domain of the TIMSS framework for the Grade 8 population. We decided that we should each conduct two independent classifications, after which we worked together to resolve disagreements and so reach one final classification.

In addition to classifying the PISA items according to the TIMSS content domains, we used a set of study-independent item descriptors as external criteria for comparison of the two studies. We then classified all the mathematics items in both studies according to a set of five-item descriptors, presented in Table 1. Olsen and Grønmo (2006) used these (and some other) indicators in a cluster analysis of countries' relative achievement profiles in mathematics in PISA. They concluded that it is possible to use the indicators to develop a meaningful description of the relative strengths and weaknesses in the different country clusters of achievements in mathematics in PISA. Because the present study's final aim is to evaluate countries' relative success in PISA versus TIMSS, these descriptors may give us valuable clues as to why some countries seem to perform better in one of the two studies. We discuss any found differences between the studies in light of the shifts in ranks, and in light of the relative achievement profiles of some selected countries across different types of content in the two studies.

Results

Country ranks in the two studies

Although the ranks do not allow study of the distances between the countries' achievements, it is the only viable "scale" to use for this comparison, given there is no formal link between the scales.[3] Table 2 presents the ranks (and the scores) for each of the education systems participating in both PISA and TIMSS. The "difference rank"—PISA rank minus TIMSS rank—appears in the right-most column. The table is sorted from high to low differences, which means that the countries in the upper part of the table performed relatively better in PISA than in TIMSS, while the countries in the lower part of the table performed better in TIMSS than in PISA.

Table 1: Description of Some Broad (External) Descriptors Used to Classify Items

Variable name	Description
Item format	Classification of items into the three main formats: selected response (SR), short constructed response (SCR), and extended selected response (ESR). The difference between the two constructed response formats is that the former are items where the response is either one word or one number, while the latter requires a description, an explanation, an argument, etc.
Algebraic expressions	Classifies the items that include the use of explicit algebraic expressions.
Calculations	Classifies the items that require exact calculations.
Graphics	Classifies the items that include the use of graphical representations of quantities.

[3] To judge the appropriateness of using this rank, we calculated the leaps in the international scores for each step in the rank order. On average, a shift of one position in the ranks for each of the two studies represents six points on both of the internationally centered scales (both scales with an international mean of 500 and a standard deviation of 100). Some large deviations from this mean were evident in the differences between some of the very low and the very high ranks. The jump between rank 20 and 21 for the two studies, that is, the jump down to Indonesia, represented an extreme value because the two low-performing countries (Indonesia and Tunisia) are extreme cases on this list.

Table 2: Ranks (from 1 to 22) for the Countries Participating in both TIMSS and PISA

	PISA score	PISA rank	TIMSS score	TIMSS rank	DIFF rank*
Scotland	524	8	498	15	7
New Zealand	523	10	494	16	6
Norway	495	14	461	20	6
Spain (Basque province)	502	12	487	17	5
Belgium (Flemish)	553	1	537	5	4
Australia	524	8	505	12	4
Sweden	509	11	499	14	3
Netherlands	538	4	536	6	2
Canada (Ontario)	530	7	521	8	1
Hong Kong SAR	550	2	586	2	0
Indonesia	360	21	411	21	0
Tunisia	359	22	410	22	0
Canada (Quebec)	537	5	543	4	-1
Italy	466	19	484	18	-1
Serbia	437	20	477	19	-1
Korea, Rep. of	542	3	589	1	-2
Japan	534	6	570	3	-3
USA	483	16	504	13	-3
Slovak Republic	498	13	508	9	-4
Latvia	483	16	508	9	-7
Hungary	490	15	529	7	-8
Russian Federation	468	18	508	9	-9

Note: * Gives the difference between the two ranks; positive numbers correspond to a higher rank in PISA.

Overall, there was a strong correspondence in the country rankings of the two studies ($r = 0.76$). The high-achieving countries were more or less the same in both studies. However, there was also a systematic tendency in how the ranks of the countries differed between the two studies. Of course, the number of countries is small, and to generalize this pattern to include even more countries is not entirely possible. Still, it is interesting to note that some of the English-speaking countries, the two Scandinavian countries, and the Dutch/Flemish-speaking education systems, in addition to the Basque province in Spain, ranked higher in PISA than in TIMSS. What is even more striking is that the East-European countries performed more strongly in TIMSS than in PISA.

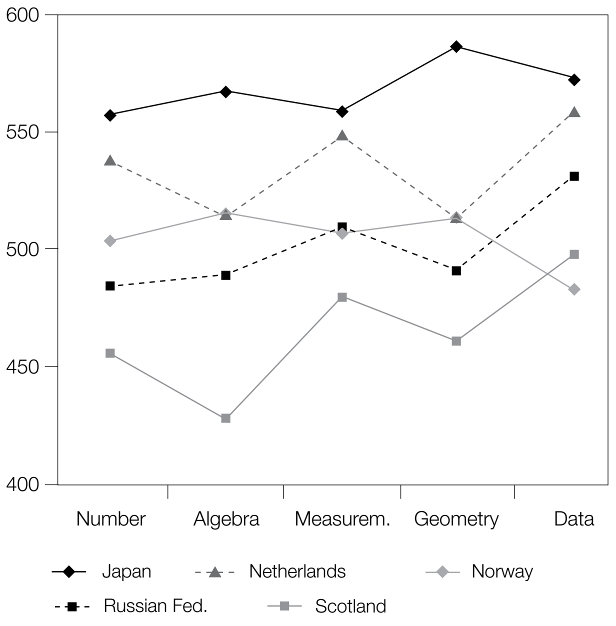
Country profiles across content

Figures 2 and 3 show the characteristic profiles of each country across the major content categories in TIMSS (Grade 8) and PISA for five selected countries. We selected these five countries because each represents a group of countries with similar achievement profiles in mathematics as documented in the above-mentioned cluster analyses. Based on the outcomes of these analyses, we considered it reasonable to highlight some particularly stable groups of countries representing different profiles in mathematics education in school. We may also talk about a continental European group, although this group is not equally well documented in these studies. In Figures 2 and 3, each of these five groups is represented by the specific country mentioned in parenthesis below. In short, we used three criteria to select these five countries:

1. The selected countries must represent an international variation in domain-specific achievement patterns. The five selected countries each represented one of five well-established clusters of countries with distinctly different achievement profiles across the mathematics items in both PISA and TIMSS (Grønmo, Bergem, Kjærnsli, Lie, & Turmo, 2004; Olsen, 2006; Zabulionis, 2001). The clusters, and their associated country, were an English-speaking group (Scotland); an East-European group (Russia); an East-Asian group (Japan); a Nordic group closely related to the English-speaking group (Norway); and a continental European group (the Netherlands).[4]
2. The selected countries must have participated in both TIMSS 2003 and PISA 2003.
3. The five selected countries must be comparable in terms of average age (of students) in TIMSS (Grade 8). We included this criterion because one possible confounding variable when interpreting data from TIMSS is the age of the participating students.

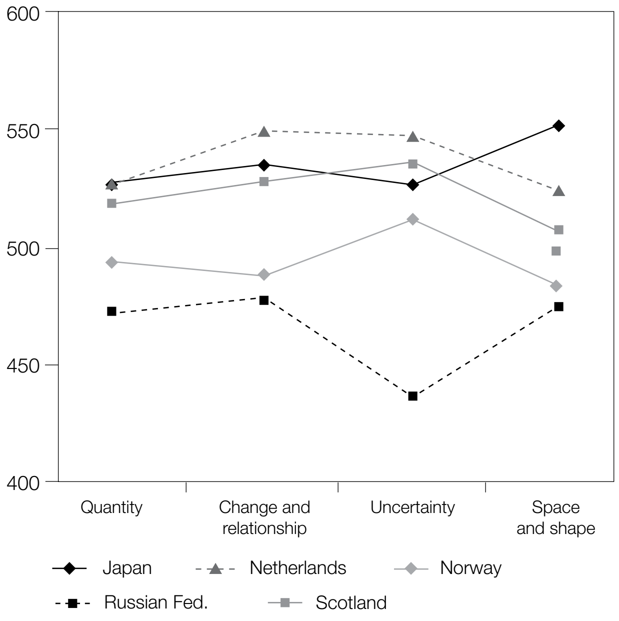
Comparison of the overall tendencies in Figures 2 and 3 reveals many notable features First, the between-country variation is larger in TIMSS than it is in PISA. The data from TIMSS included in Figure 2 span the achievement scales from about 420 to 590, while the corresponding PISA data in Figure 3 range from 440 to 550. Second, the variation across the content domains for each country is larger in TIMSS than is the variation between the overarching ideas in PISA. For the TIMSS data in Figure 2, the range for the achievement scores for each country averages 45 points, while in Figure 3 the same measure is 30 points. Third, the two figures illustrate the overall patterns in the shift of ranks between TIMSS and PISA as presented in Table 2: Scottish 15-year-olds performed much better in the PISA assessment than they did in the TIMSS assessment. Russia had the opposite shift, with students scoring much better in TIMSS. A fourth overall observation is the consistently high performance of Japan and the Netherlands on both assessments across all the mathematical content. It is particularly interesting to study in more detail the Dutch mathematics performance relative to a country

Figure 2: Achievement in Major Content Areas in TIMSS (Grade 8)



Note: The first axis crosses the second axis at the overall international mean value of 467.

Figure 3: Achievement in Major Content Areas in PISA



Note: The first axis crosses the second axis at the overall international mean value of 500.

[4] A specific problem with this last group, in terms of the analyses in this paper, is that most of the predominantly German-speaking countries did not participate in TIMSS 2003.

like Norway. The profiles of both countries are almost identical across the content domains in TIMSS (Figure 2), and there are many similarities in the profiles across the overarching ideas in PISA (Figure 3).

Furthermore, we can note particular country-specific elements in the profiles. Japan, besides scoring very highly on all subscales of content in both studies, performed extraordinarily well in Geometry and in Space and Shape. The Netherlands performed very well in all domains. However, it did not perform as well in Algebra and Geometry, the two content domains that include the most formal and abstract mathematics. Norway achieved a very low level of performance overall as compared to the performance overall of the other countries in TIMSS. However, Norway's performance in PISA was much better relative to that of the other countries, but it was still toward the lower end of the range. Specifically, the Norwegian performance was very weak in Algebra, and much stronger in Data and Measurement in TIMSS and in Uncertainty in PISA. The figures show the Russian profile has a characteristic minimum for items within Data in TIMSS and Uncertainty in PISA, and the profile is higher for Algebra and Geometry in TIMSS.

Comparison of the items in the two studies

It is not easy to compare the instruments used in TIMSS and PISA. First, what should be the comparative criteria? As we discussed in the introduction, the PISA items were developed according to content categories reflecting what is labeled as overarching ideas. As briefly noted above, these labels and the whole framework of PISA suggest that the concept of mathematical literacy should be operationalized through more realistically contextualized mathematical problem-solving tasks. In general, this means that many problems include different types of mathematics. We could say that while TIMSS items have "high fidelities," the PISA items have broad "bandwidths" in terms of content coverage. We could therefore classify several PISA items according to more than one TIMSS content domain. One hypothetical PISA item could be, for instance, a realistic problem relating to the overarching idea Space and Shape. Items in this category may typically involve geometrical objects with measurements, and the task may require students to perform calculations with decimal numbers. In classifying this item into the TIMSS framework, we would therefore have to define it predominantly as a Geometry, Measurement, or Number problem. For most items, this approach was easy, but for others it was extremely difficult, and in a few cases even impossible.

Table 3 summarizes the results of this classification work. The table reports the final classifications after agreement between the two of us. The table also summarizes the classifications according to the external criteria referred to in Table 1. Table 3 supports the previous brief comparison of the frameworks. PISA differs from TIMSS mainly in its inclusion of more items relating to the reading, interpretation, and evaluation of data, in one form or another. According to the second set of comparisons established under the external criteria, the main difference is that more items in PISA than in TIMSS concern graphical representations of data, which is a natural consequence of a test that seeks to relate more strongly to realistic or authentic contexts. In addition, mathematics in the "algebraic mood" is "weighted down" in PISA as compared to TIMSS. However, as indicated by the external descriptor "Algebraic Expressions," the difference is not primarily a product of the use of formal algebraic expressions. The stronger emphasis on algebra in TIMSS relates to less stringent notions of algebra as expressed in the concept pre-algebra. Thus, the focus is on items that typically set up tasks that require students to find general patterns in sequences of numbers or figures. Finally, Table 3 shows that PISA included more open-ended response formats than did TIMSS.

Also evident from Table 3 is the fact that we found five of the PISA items impossible to classify according to the TIMSS framework. These items relate to what could be labeled discrete mathematics; three related to combinatorics. If we did force these items into the TIMSS framework, then Number would be the most suitable content domain to use.

Each of us independently placed 73% of the PISA items into the same content domain. We noted some recurring differences in interpretation. There were two main sources of disagreement. Of our total 21 disagreements, six related to Geometry versus Measurement. Many of the problems in PISA relating to Geometry include measurements of lengths, areas, or volumes. In reaching consensus for these six items, we classified four as Measurement. We also disagreed

Table 3: Relative Distribution of Items across Content Descriptors

	Item content descriptor	TIMSS (*N* = 194)	PISA (*N* = 84)
TIMSS content domains	Number	29	25
	Algebra	24	8
	Measurement	16	10
	Geometry	16	18
	Data	14	35
	Unclassified		5
External criteria	Item-format (SR / SCR / ECR)	66 / 31 / 3	33 / 42 / 25
	Algebraic expressions	16	10
	Calculations	42	37
	Graphics	9	21
	Tables	11	13

Note: The three figures given for item format are Selected Response (SR), Short Constructed Response (SCR) and Extended Constructed Response (ECR).

on whether six other items should be classified as Number or Data. These items typically involved the use of tabular or graphical representation of quantities. Eventually, we classified five of these items as Data.

Table 4 shows in more detail the match between the PISA content categories—the overarching ideas—and the TIMSS content domains. There is an overall agreement, with a rough one-to-one correspondence as follows:

- The majority of the Space and Shape items in PISA relate to Geometry in TIMSS;
- The majority of the Quantity items in PISA relate to Number in TIMSS; and
- The majority of the Uncertainty items in PISA relate to Data in TIMSS.

In addition to presenting an overall simple correspondence between some of the overarching ideas in PISA and the content domains in TIMSS, Table 4 shows notable deviations from this pattern. First, the overarching idea labeled Change and Relationship in PISA spreads across several TIMSS content domains. Phenomena organized under this heading might be described mathematically through several forms of representation. In other words, even if the phenomena might reasonably be placed in one class, as in PISA, the mathematics involved might vary across the specific representations of those phenomena. The items might relate to simple numerical descriptions of special cases of change, or they might be described as more general relationships, either in algebraic forms or with tables and graphs. It is also interesting to note that the overarching idea Space and Shape in PISA includes several items that match the descriptions of the TIMSS content domain Measurement, a relevant issue that we return to in the following discussion.

Conclusion and discussion

Country-specific patterns of achievement

Figures 2 and 3 (above) presented specific profiles of achievement for five countries across the content domains of TIMSS and the overarching ideas in PISA. The main criterion in the selection of these five countries was that each should represent clusters of countries that were shown in previous studies to be stable groupings in both TIMSS 1995 and PISA 2003.

In a follow-up of the cluster analysis of the PISA 2003 mathematics achievement profiles, Olsen and Grønmo (2006) studied in more detail the characteristic features of each of the profiles. The most prominent findings were that the English-Nordic group had particularly high achievement for items setting realistic and authentic tasks within realistic stimulus materials. However, these countries performed relatively lower on items requiring exact calculations and use of algebraic expressions, while the group of Central East-European countries had largely the opposite profile, scoring relatively better on purer mathematical tasks that required students to calculate and/or relate to algebraic expressions. Although the profiles for the cluster of Central West-European countries and the cluster of East-Asian countries also was very distinct, they were not equally strongly related to the set of criteria describing the items.

Table 4: Correspondence of Classification Categories to PISA 2003 Mathematics Items, Classified According to the TIMSS Framework

			Content domain						
			Number	Algebra	Measurement	Geometry	Data	Unclassified	Total
Overarching idea PISA	Space & Shape	Count	1	0	5	14	0	0	20
		% within overarching idea	5%	0%	25%	70%	0%	0%	100%
	Change & Relationship	Count	3	6	3	0	10	0	22
		% within overarching idea	14%	27%	14%	0%	45%	0%	100%
	Quantity	Count	15	1	0	0	2	4	22
		% within overarching idea	68%	5%	0%	0%	9%	18%	100%
	Uncertainty	Count	2	0	0	1	17	0	20
		% within overarching idea	10%	0%	0%	5%	85%	0%	100%
Total		Count	21	7	8	15	29	4	84
		% within overarching idea	25%	8%	10%	18%	35%	5%	100%

The main overall conclusion in respect of our present study is that what we found is consistent with our previous work. The English and Nordic countries (as represented by Scotland and Norway, respectively) have profiles across the content categories that are consistent with the description given above. It is also evident from our detailed comparisons of items in PISA and TIMSS that the main difference between PISA and TIMSS is the stronger emphasis in PISA on realistic contexts. In other words, the differences between PISA and TIMSS items relate considerably to the relative strengths and weaknesses evident in the English-Nordic and East-European profiles, with the effect that the English-Nordic countries seem to be more successful in mathematics (as it is defined in PISA) than they are in TIMSS, and vice versa. East-Asian countries and the other clusters of countries did not have such articulated differences along the content domains. Consistent with this, there was little difference between the two sets of ranks for many of these countries.

Operational sharpness of the content domains

Our initial analysis of the TIMSS and PISA frameworks suggested that TIMSS has an operational definition of the content dimension that is far finer in grain than is the definition for PISA. The data in Figures 2 and 3 support this premise. There we could see that the overall variation for specific countries across the five content domains in TIMSS was larger than the corresponding variation across the four overarching ideas in PISA. It is also reasonable to think that the larger between-country variation seen in TIMSS than in PISA relates to this increased operational sharpness.

This sharpness of definition has two consequences. First, data from comparative studies like PISA and TIMSS should highlight differences between countries. Such differences, either between diverging education systems or closely related education systems, are useful starting points for educational research (Olsen, 2005). Thus, from this perspective, enlarging the differences between countries increases the potential for meaningful comparison. However, from a measurement perspective, increasing the differential performance across the items or across the content domains poses a threat for meaningful comparison because the increase can indicate potential international measurement errors due to content-by-country interactions (Wolfe, 1999). Put simply, replacing specific items or redefining the relative weights of each content domain is likely to have a larger effect on the total score in mathematics in TIMSS than in PISA. Thus, we could say that TIMSS gives us a measure of mathematics achievement that is sharper than that for PISA, but which has the possible side-effect of a measure that is less stable.

Our independent categorizations of the PISA items from 2003, using the framework of TIMSS 2003, produced an agreement level of 73%. We believe the main reason behind the non-perfect classification agreement is because the relevant items truly are crossovers of these domains. As such, we think it is preferable to interpret this index of classification agreement as a measure of the degree to which it is possible to link PISA items to the TIMSS framework, rather than as a measure of marker reliability. Furthermore, our disagreement seemed fairly systematic, thereby indicating that the TIMSS 2003 framework (Mullis et al., 2001) may in some cases be a little unclear. The topic Measurement and its relation to the topic Geometry seem particularly problematic; likewise, Number versus Data. In discussing these disagreements with each other, we decided that the framework allows good arguments for both types of classification. This is equally true for those items on which we happened to agree.

With these considerations in mind, it is interesting to note that the new framework for TIMSS 2007 has deleted the category Measurements. Some of the items originally classified as Measurement in 2003 are being used as link items in the 2007 cycle of TIMSS. These items have thus been reclassified, and several of the items previously categorized as Measurement have been categorized as Geometry. The reasons given for this change in the TIMSS framework has been somewhat technical thus far. One point made is that trend studies cannot have too many categories, as these limit ability to report measures of trends of an appropriate quality. Our analysis adds substantial arguments for why it is wise to exclude the Measurement category. One such argument is that the quality of the reporting scales for the remaining four content domains increases because of the increased specificity in the operational definitions of the remaining content domains.

The case of pure and applied mathematics in school

Many countries have as a goal that, on leaving compulsory school, all students have at hand a type of competence we can term mathematical literacy; in other words, they are well prepared to solve daily life problems using mathematics and can be active citizens in a modern society. PISA seems well suited to allowing us to determine if students in a country have met this goal. TIMSS complements this information. If we look back at the model of the mathematization cycle in Figure 1, we can see that PISA measures the mastery of all the processes involved in solving mathematical problems originating from a real-world context. TIMSS, however, gives a measure of the mastery of the mathematical processes (see the right-hand side of the figure). Taken together, TIMSS and PISA allow us to identify more specifically how a country might increase the mathematical literacy of its students. If a country or group of countries achieves better in PISA than in TIMSS, it may be because students have problems with competence in pure mathematics in general, or in specific topics in mathematics. If the opposite is the case, that is, a country is achieving better in TIMSS than in PISA, it may be because the students are not experiencing the full mathematical cycle for applied mathematics as illustrated in Figure 1.

We have documented that a country like Russia, representing the East-European profile, performs better in TIMSS than in PISA. From the simple proposition suggested above, the interpretation would be that most of the East-European countries give little attention to the left-hand side of the mathematization cycle. The general message that this example therefore serves to communicate is that "back to basics" is not a complete solution if the aim is to foster students with mathematical literacy.

Japan, representing the East-Asian profile, attained a high level of achievement in both studies, but more so in TIMSS than in PISA. This situation indicates that the mathematics curricula for schools in East-Asian countries focus to a considerable extent on pure mathematics in all topics, but simultaneously give some attention to the full cycle of applied mathematics. The Netherlands, like Japan, is among the high-achieving countries in both studies. Nevertheless, the TIMSS achievement data revealed some distinct differences between their achievement levels in different topics. While Japan and the Netherlands achieved equally well in the topics Number, Measurement, and Data in TIMSS, there were clear differences between these countries in their achievement levels in Algebra and Geometry. This finding tells us that even high-achieving countries may not be identical when it comes to what they focus on in the curriculum. Algebra and Geometry seem to be much more in focus in Japan than in the Netherlands. And when it comes to achieving well in mathematical literacy, as tested in PISA, the Netherlands is doing just as well as Japan. We take all of this as an indication that the "basics" of most importance for daily life mathematics are the fundamental concepts of numbers and operations with numbers, more so than a "basic" expressed as formal insight into geometry and algebra.

The shape of the graphs in this paper is more or less the same for student achievement in specific mathematics TIMSS topics in Norway and the Netherlands. As such, the mathematics curricula for schools in both countries have many similarities. The difference, however, is that, unlike Norway, the Netherlands is a high-achieving country in general in TIMSS and even more so in PISA. The achievement of Norwegian students was lower than the achievement of the students in all other Nordic countries in PISA, and was even lower in TIMSS at both Grades 4 and 8. A comparison of the Norwegian students' achievement in TIMSS with what the Norwegian mathematics curriculum focuses on suggests that the most problematic topic in TIMSS is Number. That the achievement in Algebra is even lower is easy to explain, since this topic is generally a very small part of what has been taught in compulsory school over the last decade. Number, however, is an extensive part of the curriculum all through compulsory school. Norwegian students' lack of elementary knowledge and skills in Number was even more pronounced for the Grade 4 students who participated in TIMSS, but we have not presented this result in this paper.

Because Number is an extensive part of the curriculum in compulsory schools in Norway, as it is in most other countries, the point we want to emphasize is one we made earlier in this paper. One consequence of a growing focus on applied mathematics is that problems arise if too little attention is given to pure mathematics, especially pure mathematics involving elementary knowledge and skills in a topic like Number. The Norwegian results in PISA and TIMSS seem to confirm that this may be the case in Norwegian schools. Earlier analysis of the kinds of items in PISA that Norwegian students performed well on and not well on also underlines this thinking as a reasonable interpretation of the results. While Norwegian students performed relatively well on PISA items (that is, close to the knowledge and abilities that students need to have in their daily lives), they performed poorly on items requiring any type of exact calculations (Olsen & Grønmo, 2006). Basic skills in elementary mathematics seem to be necessary conditions for doing well in applied mathematics, as tested in PISA. The TIMSS data also support this supposition: the countries that did well on items in problem solving also gained high scores on the more elementary items (Mullis et al., 2004).

Our analysis and comparisons between TIMSS and PISA support the notion that if our students are to do well in the mathematics that they need to know to navigate daily life, they need a basis of knowledge and skills in pure mathematics, especially elementary knowledge and skills in Number. This consideration points to the importance of school mathematics curricula not seeing mathematical literacy as an alternative to pure mathematics. A reasonably high level of competence in pure mathematics seems to be necessary for any type of applied mathematics, as we pointed out in our discussion of Figure 1. However, if countries give too little attention to the full cycle of applied mathematics, their students are unlikely to develop the type of competence known as mathematics literacy.

References

Bjørkquist, O. (2001). Matematisk problemløsing [Mathematics problem-solving]. (H. Strømsnes, trans.). In B. Grevholm (Ed.), *Matematikk for skolen* [*Mathematics for schools*] (pp. 51–70). Bergen: Fagbokforlaget.

Ernest, P. (1991). *The philosophy of mathematics education.* London: Falmer Press.

Gardiner, A. (2004, June). *What is mathematical literacy?* Paper presented at the ICME 10, Copenhagen.

Grønmo, L. S., Bergem, O. K., Kjærnsli, M., Lie, S., & Turmo, A. (2004). *Hva i all verden har skjedd i realfagene? Norske elevers prestasjoner i matematikk og naturfag i TIMSS 2003* [*What on earth has happened to science and mathematics? Norwegian students' performance in TIMSS 2003*]. Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo [Department of Teacher Education and School Development, University of Oslo.].

Grønmo, L. S., Kjærnsli, M., & Lie, S. (2004). Looking for cultural and geographical factors in patterns of response to TIMSS items. In C. Papanastasiou (Ed.), *Proceedings of the IRC-2004 TIMSS* (Vol. 1, pp. 99–112). Nicosia: Cyprus University Press.

Hutchison, D., & Schagen, I. (in press). Comparisons between PISA and TIMSS: Are we the man with two watches? In T. Loveless (Ed.), *Lessons learned: What international assessments tell us about math achievement.* Washington, DC: Brookings Institution Press.

Kjærnsli, M., Lie, S., Olsen, R. V., Roe, A., & Turmo, A. (2004). *Rett spor eller ville veier? Norske elevers prestasjoner i matematikk, naturfag og lesing i PISA 2003* [*Are we on the right track? Norwegian students' performance in mathematics, science and reading in PISA 2003*] . Oslo: Universitetsforlaget [University publishers].

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international science report. Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades.* Boston, MA: Boston College.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2005). *IEA's TIMSS 2003 international report on achievement in the mathematics cognitive domains: Findings from a developmental project.* Boston, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades.* Boston, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzalez, E. J. et al. (2001). *TIMSS assessment frameworks and specifications 2003.* Boston, MA: Boston College.

National Center for Education Statistics (NCES). (2005). *Comparing NAEP, TIMSS, and PISA in mathematics and science.* Retrieved May 24, 2005, from http://nces.ed.gov/timss/pdf/naep_timss_pisa_comp.pdf

National Council of Teachers of Mathematics (NCTM). (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

Neidorf, T. S., Binkley, M., Gattis, K., & Nohara, D. (2006). *A content comparison of the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and the Programme for International Student Assessment (PISA) 2003 mathematics assessments* (NCES 2006–029). Washington, DC: National Center for Education Statistics.

Nohara, D. (2001). *A comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Students Assessment (PISA).* Working Paper No. 2001-07. Retrieved May, 2005, from http://nces.ed.gov/pubs2001/200107.pdf

Olsen, R. V. (2005). *Achievement tests from an item perspective: An exploration of single item data from the PISA and TIMSS studies, and how such data can inform us about students' knowledge and thinking in science.* Oslo: Unipub forlag.

Olsen, R. V. (2006). A Nordic profile of mathematics achievement: Myth or reality? In J. Mejding & A. Roe (Eds.), *Northern lights on PISA 2003: A reflection from the Nordic countries* (pp. 33–45). Copenhagen: Nordic Council of Ministers.

Olsen, R. V., & Grønmo, L. S. (2006). What are the characteristics of the Nordic profile in mathematical literacy? In J. Mejding & A. Roe (Eds.), *Northern lights on PISA 2003: A reflection from the Nordic countries* (pp. 47–57). Copenhagen: Nordic Council of Ministers.

Organisation for Economic Co-operation and Development (OECD). (2003). *The PISA 2003 assessment framework: Mathematics, reading, science and problem solving knowledge and skills.* Paris: OECD Publications.

Organisation for Economic Co-operation and Development (OECD).(2004). *Learning for tomorrow's world. First results from PISA 2003.* Paris: OECD Publications.

Schoenfeld, A. H. (2004). The math wars. *Educational Policy, 18*(1), 253–287.

Skovsmose, O. (1994). *Towards a philosophy of critical mathematics education.* Dordrecht: Kluwer Academic Publishers.

Wolfe, R. G. (1999). Measurement obstacles to international comparisons and the need for regional design and analysis in mathematics surveys. In G. Kaiser, E. Luna, & I. Huntley (Eds.), *International comparisons in mathematics education* (pp. 225–240). London: Falmer Press.

Zabulionis, A. (2001). Similarity of mathematics and science achievement of various nations. *Education Policy Analysis Archives, 9*(33). Retrieved from http://epaa.asu.edu/epaa/v9n33/

Achievement of Australia's junior secondary school Indigenous students: Evidence from TIMSS[1]

Sue Thomson
Australian Council for Educational Research
Melbourne, Victoria, Australia

Introduction

Australia is one of the most developed countries in the world, ranking second on the United Nations Human Development Index. Rates for infant and maternal mortality, educational enrolment, life expectancy, adult literacy rates, and GDP per capita are among the best of any highly developed nation. Australia, however, is also home to an indigenous population known as the Aboriginal and Torres Strait Islander[2] people, and they have not shared in the high state of development of other Australians. Indigenous Australians experience higher infant and maternal mortality rates, lower levels of education, higher rates of substance abuse and imprisonment, and a life expectancy rate some 17 years lower than that of other Australians. Improving educational experiences and outcomes for such a disadvantaged group is critical to improving all outcomes for the group.

Under the Australian constitution, education is a responsibility of the eight state and territory governments. In 1999, the Australian State, Territory, and Federal Ministers of Education met at the 10th Ministerial Council on Education, Employment, Training, and Youth Affairs (MCEETYA), and at this meeting made a historic commitment to improve schooling in Australia within a framework of national collaboration. A broad range of areas of common concern were identified. Among them was the view that students should have attained the skills of numeracy and English literacy such that "every student should be numerate, able to read, write, spell and communicate at an appropriate level" (MCEETYA, 1999). Recognizing that outcomes for Indigenous students have been identified for many years as problematic, the ministers further argued that "schooling should be socially just, so that ... Aboriginal and Torres Strait Islander students have equitable access to, and opportunities in, schooling so that their learning outcomes improve and, over time, match those of other students" (MCEETYA, 1999).

Over the years, Indigenous educational policy has emphasized the importance of monitoring Indigenous students' educational outcomes nationally as a means of assessing the ongoing efficacy of implemented educational policy. A considerable amount of data has been gathered on Indigenous students' literacy levels, in particular for children of primary school age (see, for example, Frigo, Corrigan, Adams, Hughes, Stephens, & Woods, 2003). These data show low baseline levels of literacy and rates of increase that are lower than the rates for other Australian students.

Results from the Organisation for Economic Co-operation and Development's (OECD) Programme for International Students Assessment (PISA) in 2000 and in 2003 indicated that Australia's Indigenous 15-year-old students performed at a lower level in all three areas of assessment—reading literacy, mathematical literacy and scientific literacy—than did non-Indigenous students. In addition, the achievement levels of Indigenous students were lower than the international means in all three assessment areas, whereas the achievement levels of non-Indigenous Australian students were well above the international means in each of the three areas (De Bortoli & Cresswell, 2004; Thomson, Cresswell, & De Bortoli, 2004).

In all cycles of TIMSS in Australia, achievement levels of Indigenous students were significantly lower than the achievement levels of other Australian students and significantly lower than the international averages in both mathematics and science (Lokan, Ford, & Greenwood, 1997; Thomson & Fleming,

[1] The author thanks Pippa McKelvie, Hannah Murnane, and Lisa De Bortoli for their contributions to this paper.

[2] It is recognized that Aboriginal people and Torres Strait Islander people are separate ethnic groups. However, in this paper I refer to them collectively as Indigenous people.

2004a, 2004b; Zammit, Routitsky, & Greenwood, 2002). This lower level of achievement is also evident in national benchmarking studies and in Australia's national longitudinal study of young people known as the *Longitudinal Surveys of Australian Youth*. A national report into Indigenous education in 2003 reiterated that considerable gaps between Indigenous and non-Indigenous outcomes remained in literacy, numeracy, student attendance, retention into senior secondary education, Year 12 certificates and completion rates in vocational education and training (VET) and in higher education. However, it was encouraging to find the gap between Indigenous and non-Indigenous students enrolling at bachelor degree or higher continuing to decrease (Commonwealth of Australia, 2005).

The role of education is to prepare children for their futures. The MCEETYA recognizes that being literate and numerate is an essential part of this preparation for all students. To continue to fail to provide Indigenous children with opportunities to access high-quality education condemns them to remain on the fringes of Australian society. The purpose of this paper, therefore, is to explore and document the educational outcomes of the Indigenous people of Australia, and to explore the factors that have the strongest influences on their levels of achievement in mathematics and science in the middle years of secondary school, that is, the period shortly before many of them will leave school.

TIMSS Australia 2002

In 2002, TIMSS collected data from students Australia-wide.[3] Within this sample, Indigenous students were deliberately over-sampled (relative to actual numbers) to permit a more detailed statistical analysis of Indigenous student achievement than would have been possible under normal sampling conditions. Teachers were asked to administer the TIMSS questionnaires and tests to all students in the selected year level, sampled or not, who identified as Aboriginal or Torres Strait Islanders, thus providing a larger sample of Indigenous students for reporting purposes. There are particular problems obtaining a "good" sample of Indigenous students, as many live in remote areas and attend small schools that have a lower probability of being selected in the main sample, and in some cases their schools are excluded either because they are just too remote, or because the language of instruction is not English. In addition, absenteeism is frequently a problem in schools with a high proportion of Indigenous students, particularly on days when there is an external assessment planned. Given these problems, however, it is seen as imperative that such data are collected, and are collected from as broad a sample of the Indigenous population as possible.

The total sample of students used in this study comprised 5,127 Australian students, sampled as the main sample for TIMSS 2002. In addition, the data set included a further 356 Indigenous students who were sampled in the same schools, bringing the total sample size to 5,483 with 562 Indigenous students. While this is still a small number from which to draw any firm conclusions, it is much larger than the 4% that is the actual proportion of Indigenous students in the secondary school population.

Background information

Table 1 provides some basic demographic background for the Indigenous and non-Indigenous students who participated in TIMSS 2002. There is little difference between the Indigenous and non-Indigenous population in terms of gender and age distributions. Australian schools have a policy of automatic promotion from grade to grade, and so ages for the subgroups should indeed be similar.

Secondary student outcomes generally are reported according to geographic location, based on a combination of population, accessibility, and remoteness. Three broad classifications were defined for TIMSS: metropolitan, provincial, and remote. Most Australians live in metropolitan areas—almost two-thirds of the sampled students did so, but only half of the Indigenous students sampled lived in the major cities. Almost one in five Indigenous students sampled for TIMSS 2002 lived in what are classified as remote areas, compared with just one in 50 non-Indigenous students.

There is a strong link between educational resources and achievement. In almost every country in the world, students from homes with extensive educational resources have higher achievement in all areas than do those students from less advantaged backgrounds. In TIMSS, the first of these resources, "books," asks students to estimate the number of

[3] For comparability across countries and across assessments, testing was conducted at the end of the school year. The countries in the Southern Hemisphere tested in late 2002, while the remaining countries were tested at the end of the 2002/2003 school year (i.e., in 2003).

books in their home. Table 1 shows the highest and lowest categories. As is evident from the table, three times the proportion of Indigenous to non-Indigenous students had very few books in the home. Almost one third of non-Indigenous students had very few books in the home, and only 15% had more than 200 books in their home. Access to technology is also an issue for Indigenous students. About three-quarters had a computer at home compared with almost all non-Indigenous students. Similar proportions can be seen in Table 1 for having a desk in the home.

The other substantial educational resource is parents. For most children, parents are their first and probably most important educators, and so the level of parental education is an important educational resource in the home. In TIMSS studies, higher levels of parental education have consistently related to higher levels of achievement in mathematics and science. While the data in Table 1 provide some indication of levels of parental education, around one-third of Indigenous students and one-quarter of non-Indigenous students responded "I don't know" to this question. Given this caveat, a little under one third of non-Indigenous students and one-fifth of Indigenous students had parents with tertiary qualifications. However, for about one in 10 Indigenous students, the highest educational level of a parent was primary school.

In summary, Australian Indigenous students have less access to resources than do their non-Indigenous counterparts, and they are more likely to live in areas of Australia that are classified as remote. The next section of this paper examines the achievement levels of Indigenous and non-Indigenous students on the TIMSS mathematics and science assessments.

Mathematics achievement

The performance of Australia's non-Indigenous students on TIMSS compared well internationally and was significantly above the international mean for Year 8[4] mathematics. However, the performance of Australia's Indigenous students was significantly lower than the performance of non-Indigenous Australian students and significantly lower than the international mean. Table 2 displays the distribution of mathematics achievement scores for a selection of countries at Year 8.

*Table 1: Demographic Background Data for Indigenous and non-Indigenous Students in Australia, TIMSS 2002**

		Indigenous	Non-Indigenous	Total
Gender	Female (%)	53 (4.1)	51 (2.3)	51 (2.3)
	Male (%)	47 (4.1)	49 (2.3)	49 (2.3)
Age (Years)		14.0 (0.13)	13.8 (0.01)	13.8 (0.01)
Geographic location				
	Metropolitan (%)	50 (8.2)	64 (4.0)	63 (3.9)
	Provincial (%)	32 (3.4)	34 (4.0)	34 (3.9)
	Remote (%)	19 (9.9)	2 (0.4)	3 (0.6)
Home educational resources				
Books	None or very few (%)	16 (2.4)	5 (0.5)	6 (0.5)
	More than 200 (%)	15 (1.9)	31 (1.4)	31 (1.4)
Computer (%)		78 (2.7)	96 (0.3)	96 (0.3)
Desk (%)		77 (3.5)	92 (0.4)	92 (0.4)
Parental education				
Finished university or higher (%)	20 (4.6)	30 (1.3)	29 (1.3)	
Post-secondary vocational/technical (%)	15 (3.4)	28 (1.0)	27 (1.0)	
Finished secondary school (%)	32 (4.8)	25 (1.2)	25 (1.1)	
Some secondary school (%)	22 (4.8)	15 (0.8)	15 (0.9)	
Primary school or did not go (%)	12 (5.0)	3 (0.4)	3 (0.4)	

Note: *Standard errors in brackets.

[4] In Australia, the designation Year rather than Grade is used to denote each year of schooling.

Table 2: Distribution of Mathematics Achievement across Selected Countries, Year 8 Students

Year 8 TIMSS 2002/03 countries	Mean scale score (*SE*)	Average age
Singapore	605 (3.6)	14.3
Non-Indigenous Australian students	*508 (4.5)*	*13.9*
Australia	**505 (4.9)**	**13.9**
United States	504 (3.3)	14.2
Scotland	498 (3.7)	13.7
England	498 (4.7)	14.3
New Zealand	494 (5.3)	14.1
International mean	**467 (0.5)**	**14.5**
Indigenous Australian students	*429 (7.6)*	*14.0*
South Africa	264 (5.5)	15.1

On average, Australian Indigenous students scored 79 points (more than three-quarters of a standard deviation) lower than non-Indigenous Australian students in mathematics. The performance level of non-Indigenous students was comparable to the performance of students in highly developed countries such as the United States, England, New Zealand, and Scotland, although the average age of Australian students was lower than the averages for each of these countries. Indigenous Year 8 students' performance was similar to the performance of students in less-developed countries such as Egypt, Tunisia, and Indonesia.[5]

Science achievement

As in mathematics, the performance of Australia's non-Indigenous students compared well internationally and was significantly above the international mean for Year 8 science. However, the performance of Australia's Indigenous students was again significantly lower than the performance of non-Indigenous Australian students and significantly lower than the international mean. Table 3 displays the distribution of science achievement scores for a selection of countries at Year 8.

On average, Australian Indigenous students scored 72 points (almost three-quarters of a standard deviation) lower than the non-Indigenous Year 8 Australian students in science. Non-Indigenous Year 8 students' performance can again be compared to students' performance in countries such as the United States and New Zealand, and again the average age of

Table 3: Distribution of Science Achievement across Selected Countries, Year 8 Students

Year 8 TIMSS 2002/03 countries	Mean scale score (*SE*)	Average age
Singapore	578 (4.3)	14.3
England	550 (4.3)	14.3
Non-Indigenous Australian students	*530 (3.7)*	*13.9*
United States	527 (3.1)	14.2
Australia	**527 (3.8)**	**13.9**
New Zealand	519 (4.9)	14.1
Scotland	512 (3.4)	13.7
International Mean	**474 (0.6)**	**14.5**
Indigenous Australian students	*458 (7.0)*	*14.0*
South Africa	244 (6.7)	15.1

[5] These results are not shown in the tables in this paper but are available in full in the international TIMSS reports.

the Australian students was lower than the average age for these countries. While the scores for Indigenous students were not as low as they were for mathematics, there was clearly still a wide gap in performance between Indigenous and non-Indigenous students in both subject areas.

Trends in mathematics and science

To enable comparisons between cycles, TIMSS used IRT methodology to place the TIMSS 2002 results on the same scales developed for TIMSS 1994. Figure 1 and Figure 2 present the means and 95% confidence intervals for Australian Indigenous and non-Indigenous Year 8 students in TIMSS 1994 and TIMSS 2002, for mathematics and science respectively.

While other countries in TIMSS showed growth, in some cases a great deal of it, between TIMSS 1994

Figure 1: Mean Achievement Levels (and 95% Confidence Intervals) in Year 8 Mathematics for Indigenous and Non-Indigenous Students

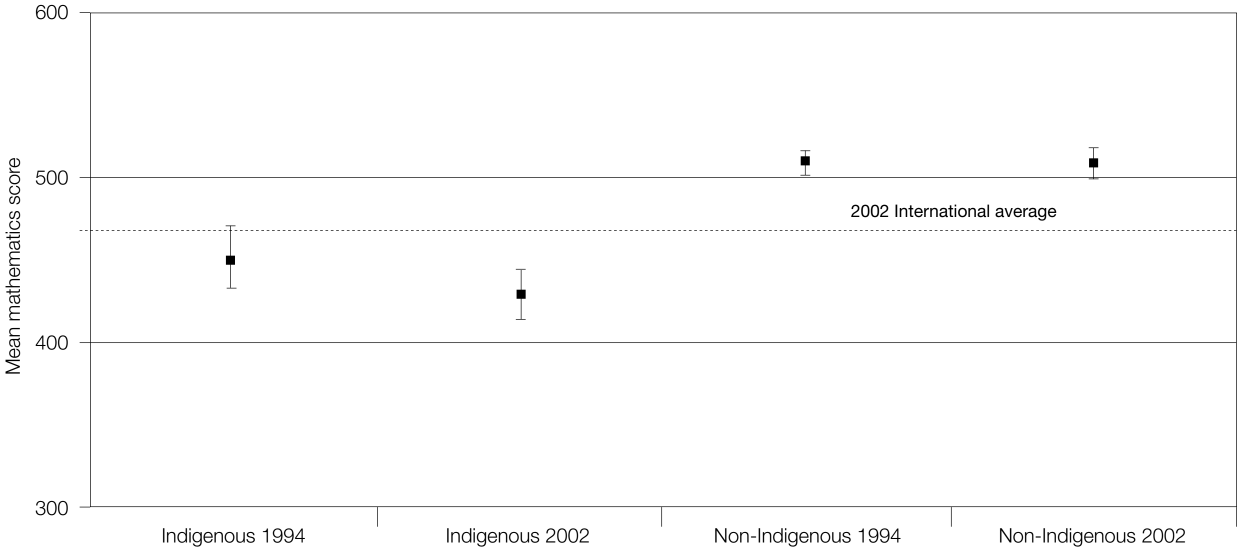
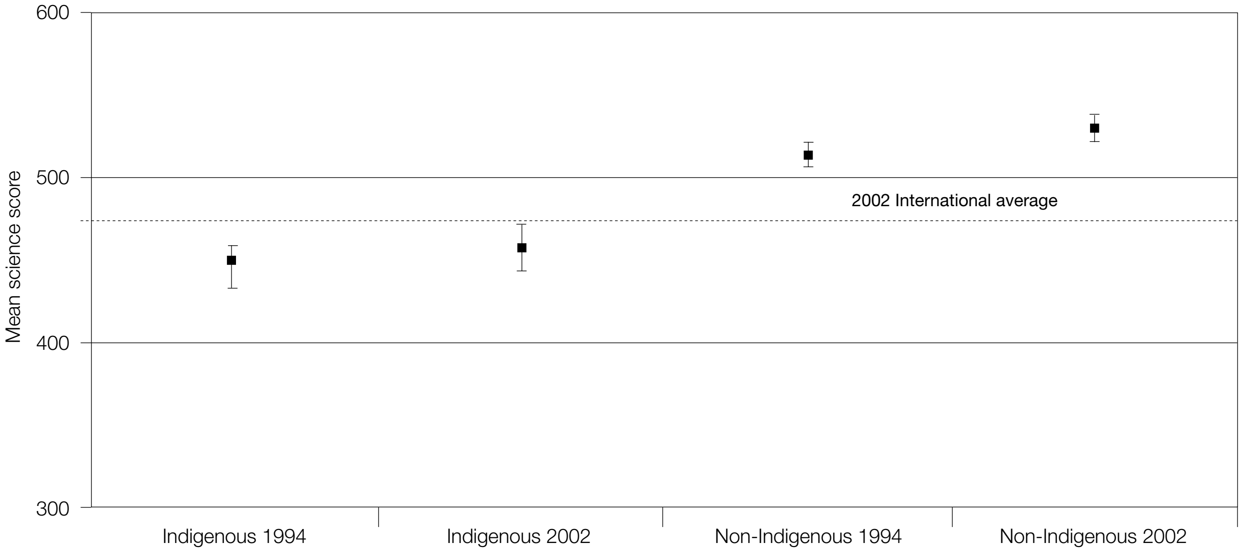


Figure 2: Mean Achievement Levels (and 95% Confidence Intervals) in Year 8 Science for Indigenous and non-Indigenous Students



and TIMSS 2002, this was not the case for Australia. Figure 1 shows that in mathematics, achievement levels remained static over this time. Also, the gap between the achievement levels of Indigenous and non-Indigenous students remained statistically about the same, although the level of performance of Indigenous students declined slightly.

Figure 2, however, shows an increase in the achievement levels of both Indigenous and non-Indigenous students. This increase was significant for non-Indigenous students, and while not so for Indigenous students, it placed them closer to the international mean.

Gender differences

Results for TIMSS 1994 found no gender differences for Australian students in either mathematics or science over the whole sample. In TIMSS 2002, gender differences for Australian students were reported for science in Year 8, in favor of males, but no gender differences overall were found for mathematics. So, in the eight years between cycles, male achievement improved significantly but the achievement level of females remained the same. Table 4 presents the results of further investigations carried out to examine gender differences for Indigenous and for non-Indigenous students. As is evident from the table, the difference in mathematics scores between Indigenous and non-Indigenous males was significant, as was the difference between Indigenous and non-Indigenous females. However, between males and females within each ethnic group (i.e., within Indigenous students), there were no significant differences. The differences in science scores were significant within and between groups. Males achieved at a significantly higher level than females within both the Indigenous and the non-Indigenous groups, while non-Indigenous males and females achieved at a significantly higher level than their Indigenous counterparts.

Performance at the international benchmarks

While the mean scores give a summary measure of students' achievement levels in a country, it is important to provide other measures that give meaningful descriptions of what performance on the scale could mean in terms of the mathematics and science that students know and can do. In order to do this, points on each of the mathematics and science scales were identified to use as international benchmarks. Selected to represent the range of performance shown by students internationally, the advanced benchmark was set at 625, the high benchmark at 550, the intermediate benchmark at 475, and the low benchmark at 400. For the purposes of the figures presented in this section of the paper, the proportion of students not yet achieving at the low international benchmark are also included.

Figure 3 show the proportion of Indigenous and non-Indigenous students achieving at each of the international mathematics benchmarks, plus the proportion of students not yet achieving at the low international benchmark, for both TIMSS 1994 and TIMSS 2002.

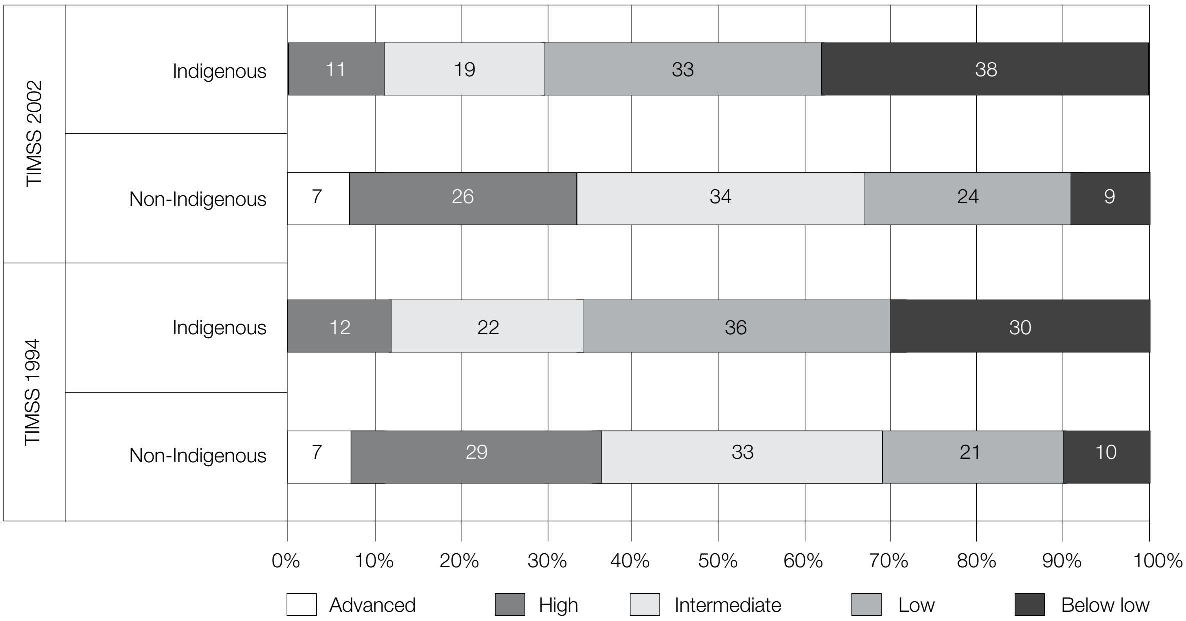
The proportion of students achieving at the highest international benchmark is of some interest, identifying as it does the highest performing students in each group. Seven percent of non-Indigenous students but less than 1% of Indigenous students achieved at this highest level. At the lowest levels, 38% (in 2002) of Indigenous students and about 10% of non-Indigenous students did not meet the requirements for the lowest international benchmark. These proportions showed no change from TIMSS 1994 for non-Indigenous students, but they worsened over the period for Indigenous students.

The Australian Performance Measuring and Reporting Taskforce (PMRT) have indicated that the minimum national benchmark for mathematics and science achievement at Year 8 level is likely to be set at the intermediate level. This means that, in 2002, one-

Table 4: Means and Standard Errors for Mathematics and Science Achievement by Indigenous Status and Gender

	Indigenous		Non-Indigenous	
	Males	*Females*	*Males*	*Females*
Mathematics achievement	445 (11.6)	415 (10.5)	514 (5.7)	502 (5.6)
Science achievement	477 (8.6)	442 (10.6)	540 (4.7)	520 (4.5)

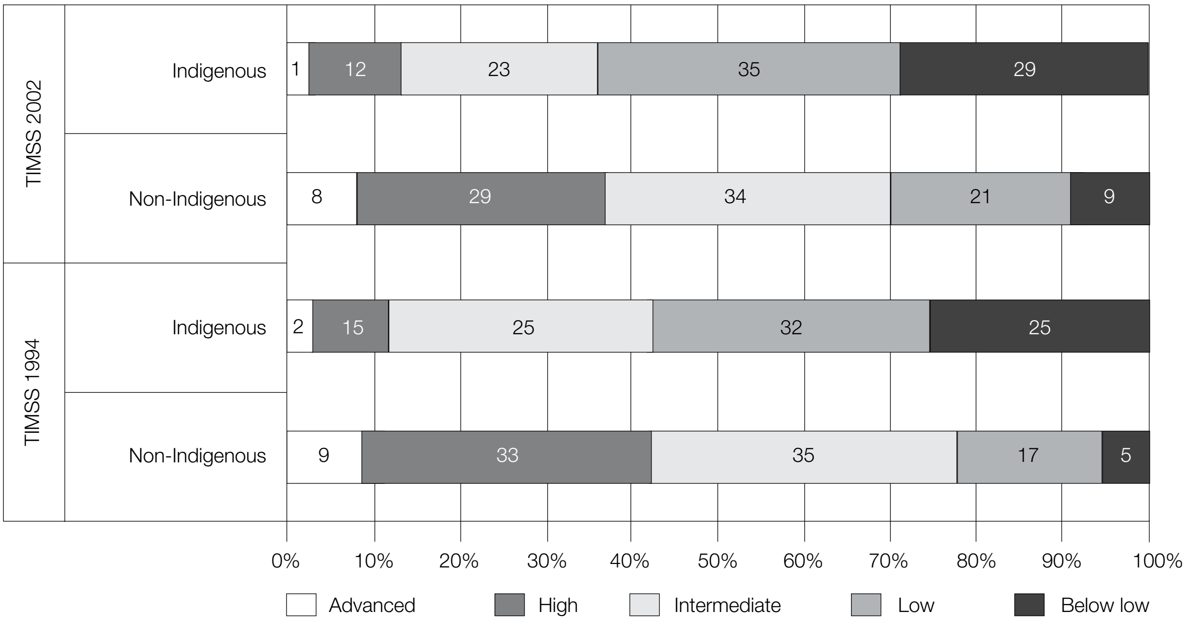
Figure 3: Performance at the Year 8 Mathematics International Benchmarks for Indigenous and Non-Indigenous students



third of non-Indigenous students would have failed to meet the national benchmark in mathematics. However, of greater concern is the fact that almost three-quarters of Indigenous students would have failed to meet the benchmark.

Figure 4 provides the same information for science achievement. Because mean science achievement increased for both Indigenous and non-Indigenous students, we could anticipate seeing increased proportions of students achieving at the highest benchmark, and a lower proportion of students failing to achieve the lowest benchmark. As Figure 4 shows,

Figure 4: Performance at the Year 8 Science International Benchmarks for Indigenous and Non-Indigenous Students



about the same proportion of students within each ethnic grouping achieved at the highest benchmark; around 2% of Indigenous students and around 9% of non-Indigenous students. Also evident is the fact that the proportion of students failing to achieve the national benchmarks, as previously described, decreased for both groups of students from TIMSS 1994 to TIMSS 2002.

Factors influencing achievement in mathematics and science

It is clear from the data presented so far in this paper that levels of achievement for Indigenous students improved little between the time of the TIMSS 1994 assessment and the TIMSS 2002 assessment. In science, there was a significant improvement in overall achievement, but in 2002 substantial proportions of students were still failing to achieve at the lowest international benchmarks. To examine in more detail some of the factors that might particularly affect achievement for Indigenous students in mathematics and science, we conducted a number of correlational analyses to examine influences on achievement for Indigenous and non-Indigenous students. These analyses produced the following findings:

- Self-confidence had a positive relationship with Indigenous achievement.
- There was a positive relationship between non-Indigenous students' enjoyment of and valuing learning in mathematics and science. These correlations were not observed for Indigenous students.
- Aspiration to higher education had a significant positive relationship with achievement. The achievement of those Indigenous students aspiring to university studies exceeded the non-Indigenous national average in mathematics, and was statistically similar to the achievement levels of non-Indigenous students.
- Those Indigenous students who spoke English infrequently in the home appeared to be at a distinct educational disadvantage. They achieved at a level substantially below that of the Indigenous students who spoke English frequently at home.
- Students who attained high scores on the home educational resources index (HERI) performed at a level that was similar to the non-Indigenous national average. However, six times as many Indigenous as non-Indigenous students fell into the low category on the home educational resources index.
- More than half of the Indigenous students sampled were attending schools that had more than one-quarter of their student population drawn from economically disadvantaged homes. Mathematics achievement was lower in these schools than in schools with fewer disadvantaged students.
- One-fifth of Indigenous students were attending schools in which the principals identified serious problems such as truancy and lateness, whereas one in 10 non-Indigenous students were attending schools with these problems. Mathematics achievement was lowest in such schools.

We next carried out a multilevel analysis in order to examine the combined influences of these and other background variables, such as gender and age, on achievement for all students, and we accounted for the clustering effect of students within schools. It was not possible to carry out multilevel analysis for the Indigenous sample, because the number of students was spread too thinly across schools. We therefore used equation modeling to find the best set of predictors for Indigenous students.

Multivariate influences on mathematics achievement

A two-level hierarchical analysis was conducted on the whole data set. We entered variables found to influence achievement into the model, and removed any variables found not to be statistically significant until the most parsimonious model was left. We used the null model to estimate the amount of between-class and within-class variance. This model indicated that 48% of the variance in student achievement in mathematics was due to differences between schools, and 52% of the variance was attributable to differences between students (within-class).

Table 5 shows the model for the multilevel analysis of mathematics achievement. Here, we can see that self-confidence had the strongest association with mathematics achievement, all other things being equal. Students in the medium category of achievement achieved an average of 33 score points more than students who reported low levels of self-confidence. Students who reported high levels of self-confidence achieved, on average, another 33 score points higher than those in the medium category. However, as is usual

Table 5: Estimates of Influences on Mathematics Achievement, All Students, Year 8

	Coefficient (*SE*)
Intercept	490 (7.4)
Student-level variables	
Self-confidence in mathematics	33 (1.4)
Indigenous	-21 (5.6)
Aspirations to higher education	10 (1.4)
Computer usage	8 (1.6)
Books in the home	6 (1.3)
Parents' highest level of education	5 (1.3)
School-level variables	
Emphasis on mathematics homework	18 (4.0)
Principal's perception of school climate	16 (3.9)
Good school and class attendance	10 (3.8)
Variance	
Explained by the model	46%
Unexplained between-schools	19%
Unexplained within-schools	36%

with data of this type, inability to control for prior mathematics achievement or for earlier self-confidence in mathematics meant it was not possible to determine whether self-confidence influenced achievement or vice versa, or indeed whether the relationship was a reciprocal one.

The model also shows that Indigenous status had the next highest relationship with mathematics achievement, with the performance of the Indigenous students being, on average, about 21 score points lower than the average score for the non-Indigenous students, all other things being equal. These factors seem to be the most influential of the student-level variables. Teachers' reports of the emphasis they placed on mathematics homework, principals' perceptions of school climate, and principals' ratings of the frequency and severity of behavioral issues in the school (good school and class attendance) were class/school-level factors that influenced achievement.

In the next step of this investigation, we carried out a regression analysis on the data for the Indigenous students only in the sample. Unfortunately, because it was not possible to use multilevel modeling, we could not identify the variance within schools and between schools. However, we can assume that the degree of variance would have been similar for Indigenous students and for all students. Far fewer factors were found to be significant for the Indigenous students in the sample than for the non-Indigenous students, and these are shown in Figure 5. The intercept, unstandardized regression coefficients (B), standard error, standardized regression coefficients (β), and significance for the significant variables are shown in Table 6. The model was statistically significant (F (5,130) = 13.978, p < .001) and explained 35% of the variation in mathematics achievement. The predictor that made the strongest unique contribution to mathematics achievement was student educational aspirations (β = .27).

On average (and when all other predictors in the model were statistically controlled):

- Students who aspired to a degree or higher had an average mathematics achievement score 50 points higher than the average score of students who aspired to a TAFE certificate or lower.
- Students who spoke English at home (always or almost always) had an average mathematics achievement score 64 points higher than the score of students who rarely spoke English at home.
- Mathematics achievement scores increased by 17 points for each additional educational possession students had in their homes.
- Students who had a high level of self-confidence when learning mathematics had an average mathematics achievement score 36 points higher than the average score of students with a low level of self-confidence.

Figure 5: Significant Influences on Indigenous Students' Mathematics Achievement

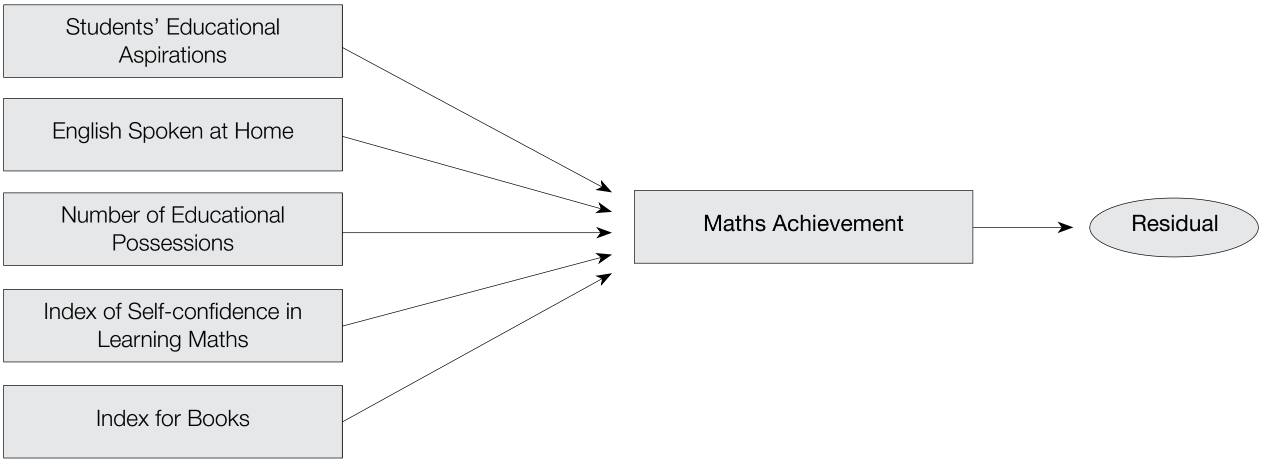

Students' Educational Aspirations
English Spoken at Home
Number of Educational Possessions
Index of Self-confidence in Learning Maths
Index for Books
Maths Achievement
Residual

Table 6: Results of Regression of Student and School Variables on Mathematics Achievement

Predictor	B	*SE*	ß	*p*
Intercept	274	22.1		<.001
English mostly spoken at home	64	18.8	0.25	<.001
Student's educational aspirations	50	14.0	0.28	<.001
Self-confidence in learning mathematics	36	14.0	0.20	<.01
Books in the home	30	13.8	0.17	<.05
Number of educational possessions	17	7.8	0.17	<.05

- Students who had a medium to a high number of books at home had an average mathematics achievement score 30 points higher than the average score of students who had a low number of books at home.

Multivariate analysis of influences on science achievement

The next step in the present investigation involved a two-level hierarchical analysis on the entire data set. As with the process for the mathematics data set, we entered variables found to influence achievement into the model. Any variables that were not statistically significant were removed until the most parsimonious model was left. We used the null model to estimate the amount of between-class and within-class variance. This model indicated that 20% of the variance in student achievement in science was due to differences between schools, and 80% of the variance was attributable to differences between students (within-class). The difference between the variance at each level for mathematics and science could be a result of how the classes were selected. Intact mathematics classes were selected, so for each class there was just one mathematics teacher. However, in most cases, these classes would separate for science lessons, meaning that most schools had three or four science teachers per mathematics class, a situation that created some issues when aggregating data to school level.

Table 7 presents the model for the multilevel analysis of science achievement. The model shows that language spoken at home had the strongest association with science achievement, all other things being equal. Students who spoke English at home all of the time or almost all of the time gained, on average, 23 score points more than students who did not speak English at home on a regular basis. Self-confidence had as strong an association with science as with mathematics. Students who had high self-confidence in science scored, on average, 20 score points more than did students with moderate levels of self-confidence, and 40 score points more than students exhibiting low levels of self-confidence. The model also shows that Indigenous status had the next highest relationship with science achievement, with Indigenous students

Table 7: Estimates of Influences on Science Achievement, All Students, Year 8

	Coefficient (*SE*)
Intercept	508 (7.6)
Student-level variables	
English spoken at home	23 (4.4)
Self-confidence in science	20 (1.7)
Indigenous	-19 (4.9)
Books in the home	13 (1.5)
Science homework	-8 (1.9)
Computer usage	8 (1.5)
Gender	-7 (2.4)
Perception of safety	9 (1.5)
Educational possessions	6 (1.5)
School-level variables	
Good school and class attendance	10 (3.8)
Percentage of disadvantaged students in school	-12 (2.6)
Variance	
Explained by the model	27%
Unexplained between-schools	7%
Unexplained within-schools	66%

attaining, on average, about 19 score points below the average score attained by non-Indigenous students, all other things being equal.

These factors seemed to be the most influential of the student-level variables. Principals' ratings of the frequency and severity of behavioral issues in their respective schools (good school and class attendance), and the proportion of students from economically disadvantaged homes in the school, were class/school-level factors that influenced achievement. Students in schools with lower levels of both severity and frequency of poor attendance, truancy, and skipping classes achieved, on average, 13 points more than did students in schools where one or more of these were a severe problem. The other significant association shows that students in schools with less than a quarter of the students from a disadvantaged background achieved, on average, 12 points more than did students in schools with over a quarter of their students from a disadvantaged background.

Again, as with the process used for the mathematics analysis, the next step of this investigation used structural equation modeling on just the data for Indigenous students in the sample. Far fewer factors were significant for these students than for the non-Indigenous students. These factors are shown in Figure 6. Table 8 presents the intercept, unstandardized regression coefficients (B), standard error, standardized

Figure 6: Significant Influences on Indigenous Students' Science Achievement

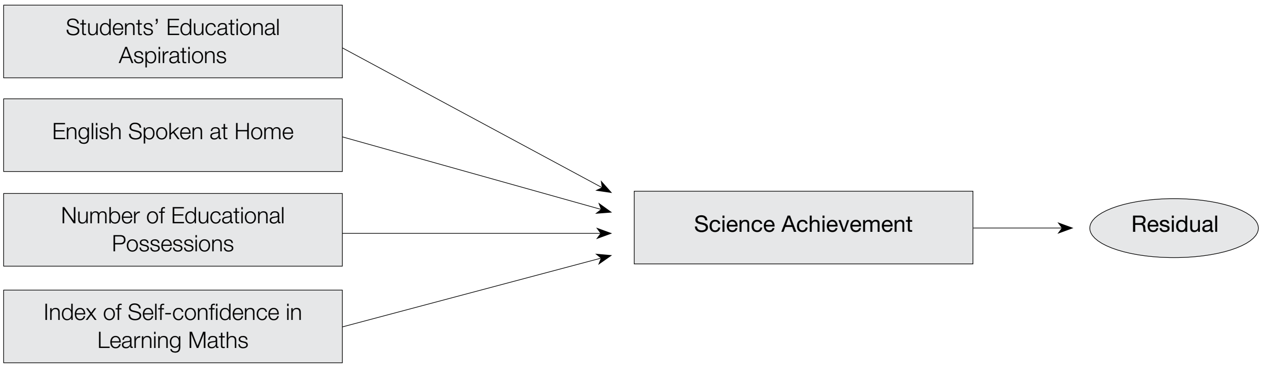

Students' Educational Aspirations
English Spoken at Home
Number of Educational Possessions
Index of Self-confidence in Learning Maths
Science Achievement
Residual

Table 8: Results of Regression of Student and School Variables on Mathematics Achievement

Predictor	B	*SE*	ß	*p*
Intercept	316	20		< .001
Student's educational aspirations	77	12.1	0.45	< .001
English mostly spoken at home	63	17.2	0.26	< .001
Books in the home	25	12.6	0.15	< .05
Number of educational possessions	15	7.1	0.15	< .05

regression coefficients (ß), and level of significance for the significant variables. The model was statistically significant ($F(4,131) = 20.973, p < .001$) and explained 39% of the variation in science achievement. The predictor that made the strongest unique contribution to science achievement was a student's educational aspirations ($\beta = .45$).

On average (and with all other predictors in the model statistically controlled):

- Students who aspired to a degree or higher had an average science achievement score 77 points above the average score of students who aspired to a TAFE (vocational) certificate or lower.
- Students who spoke English at home (always or almost always) had an average science achievement score 63 points above the average score of students who rarely spoke English at home.
- Science achievement scores increased by 15 points for each additional educational possession students had in their homes.
- Students who had a medium to high number of books at home had an average science achievement score 25 points above the average score of students who had a low number of books at home.

Summary and conclusions

The OECD has stated that "all adults—not just those aspiring to a scientific career—[should] be mathematically, scientifically and technologically literate" (OECD, 2004, p. 37). This is not the case for many of Australia's Indigenous students, and unfortunately many of these students are approaching the age at which education is not compulsory (15 in most states), and will leave school.

The aims of this paper were to examine, with reference to data from TIMSS 2002, Indigenous performance in mathematics and science, and to explore the factors that influenced this achievement. On all measures, Indigenous performance was well below the performance of non-Indigenous Australian students. There is an indication of some improvement in the area of science between TIMSS 1994 and TIMSS 2002, and this should be explored in greater depth. Overall, though, the achievement gap between Indigenous and non-Indigenous students persisted.

Four factors significantly influenced both mathematics and science achievement. All of these are largely out of the control of schools and systems. Language spoken at home is one such factor. For students who do not speak English at home, extra resources need to be provided at school level so that these students receive the support they need and have opportunities to listen and speak mathematically and scientifically at school.

We also found that the number and the type of educational possessions and books in the home related significantly to students' mathematics and science achievement. These factors reflect, to some extent, the level of support for a student's education and the socioeconomic level of the student's home. Increasing resources to schools so that students have somewhere quiet that they can study before or after school hours, providing access to computers, and encouraging students to access books and other materials over the internet, if possible, could help compensate for the effects of these variables.

The other two factors found to influence students' mathematics and/or science achievement are amenable to change by teachers and schools. One of these was self-confidence; the other was students' aspirations to higher education. The finding in relation to self-confidence is not new; many studies have found this variable to be strongly related to mathematics and science achievement, and it is understood that the relationship is complex and circular. Regardless, the strong link between student self-confidence in mathematics and science suggests that it is an important outcome in itself. The more students succeed in mathematics and science, the more likely they are to believe that they can succeed; the more students believe they can succeed, the more confident they will become with learning mathematics and science.

Students' aspirations to higher education are also important. The data analysis presented in this report showed that Indigenous students and non-Indigenous students with similar levels of aspiration had similar levels of achievement. However, the confidence intervals were much larger for the Indigenous students, a finding that suggests many of the Indigenous students with such aspirations will need substantial support if they are to achieve their goals. Nonetheless, such aspirations should be encouraged in schools, and teachers need to ensure that they particularly encourage Indigenous students to consider careers that require them to continue their education.

References

Commonwealth of Australia. (2005). *National report to Parliament on Indigenous education and training 2003.* Available on http://www.dest.gov.au/sectors/indigenous_education/publications_resources/profiles/national_report_indigenous_education_and_training_2003_part1#publication

De Bortoli, L., & Cresswell, J. (2004). *Australia's Indigenous students in PISA 2000: Results from an international study* (ACER Research Monograph 59). Melbourne, VIC: Australian Council for Educational Research.

Frigo, T., Corrigan, M., Adams, I., Hughes, P., Stephens, M., & Woods, D. (2003). *Supporting English literacy and numeracy learning for Indigenous students in the early years* (ACER Research Monograph 57). Melbourne, VIC: Australian Council for Educational Research.

Lokan, J., Ford, P., & Greenwood, L. (1997). *Mathematics and science on the line: Australian middle primary students' performance in the Third International Mathematics and Science Study* (TIMSS Australia Monograph No. 2). Melbourne, VIC: Australian Council for Educational Research.

Ministerial Council on Education, Employment, Training, and Youth Affairs (MCEETYA). (1999). *The Adelaide declaration on national goals for schooling in the twenty-first century.* Available on http://www.mceetya.edu.au/mceetya/nationalgoals/natgoals.htm

Organisation for Economic Co-operation and Development (OECD). (2004). *Learning for tomorrow's world: First results from PISA 2003.* Paris: Author.

Thomson, S., Cresswell, J., & De Bortoli, L. (2004). *Facing the future: A focus on mathematical literacy among Australian 15-year-old students in PISA 2003.* Melbourne, VIC: Australian Council for Educational Research.

Thomson, S., & Fleming, N. (2004a). *Examining the evidence: Science achievement in Australian schools in TIMSS 2002* (TIMSS Australia Monograph No 7). Melbourne, VIC: ACER Press.

Thomson, S., & Fleming, N. (2004b). *Summing it up: Mathematics achievement in Australian schools in TIMSS 2002* (TIMSS Australia Monograph No 6). Melbourne, VIC: ACER Press.

Zammit, S., Routitsky, A., & Greenwood, L. (2002). *Mathematics and science achievement of junior secondary school students in Australia* (TIMSS Australia Monograph No. 4). Melbourne, VIC: Australian Council for Educational Research.

The effect of students' perceptions of the learning environment on mathematics achievement: Explaining the variance in Flemish TIMSS 2003 data

W. Schelfhout, G. Van Landeghem, A. Van den Broeck, and J. Van Damme
Catholic University of Leuven
Leuven, Belgium

Abstract

This paper discusses the merits of a constructivist learning environment in fostering the learning of mathematics in secondary school. A 33-item instrument based on a theoretical model of powerful teaching, and designed to measure individual students' perceptions of the learning environment, was constructed and included in the Flemish part of the TIMSS 2003 assessment. The 33-item measurement was reduced to six scales by means of exploratory factor analysis. An educational effectiveness approach was then employed to determine what differences, if any, might be apparent when Flemish mathematics classes were viewed from the perspective of constructivist learning and whether any such differences related to differences in students' mathematics achievement.

Introduction

During recent decades, research has led to the identification of important characteristics of effective learning processes. These can be summarized in the following definition: learning is a constructive, cumulative, self-regulated, goal-oriented, situated, collaborative, and individually different process of knowledge building and meaning construction (De Corte, 2000). Furthermore, a series of guiding principles for the design of powerful learning environments—in line with the preceding features of effective acquisition processes—have emerged from the available research and literature (De Corte, 2000). Richardson (2003) points to the necessity of elaborating on a framework related to effective constructivist teaching as follows: "A second issue that confronts us in constructivist teaching is that because constructivism is a theory of learning and not a theory of teaching, the elements of effective constructivist teaching are not known" (p. 1629). He continues: "... [an] area of needed development in constructivist pedagogy is theory building. Theories of constructivist teaching will provide us with ways of identifying more and less effective teaching practices for use in teacher education and professional development" (p. 1636).

Researchers need to take account of some important points when constructing a model for the design of learning environments (De Corte, 2000). First, learning is complex. Consequently, when designing appropriate learning environments, we have to start from a "holistic (as opposed to a partial) approach to the learning–teaching environment, i.e. all relevant components of the learning environment should be addressed" (De Corte, 2000, p. 249). This means we need to take account of the various views of the nature of knowledge and learning. These views have different educational implications, and these implications sometimes seem to contradict one another. However, as Anderson, Greeno, Reder, and Simon (2000) point out, because the different perspectives form different aspects of the same learning and educational process, they need to be acknowledged.

This contribution starts with a description of a four-dimensional model of learning environments. Due to space restrictions, we can present only a summary of the model. For an extensive account of this model and an in-depth discussion of its underlying paradigms, see Schelfhout, Dochy, Janssens, Struyven, Gielen, & Sierens (2006a, 2006b). We then discuss how the model was employed in the construction of an instrument for measuring students' perceptions of their learning environment. This instrument was used to allow Flanders to collect supplementary data during its participation in the 2003 cycle of the International Association for the Evaluation of Educational

Achievement's (IEA) Trends in Mathematics and Science Study (TIMSS). A factor analysis of the measurements revealed six dimensions that could be interpreted in terms of the theoretical model.

In a subsequent step, scores for six scales corresponding to these dimensions were calculated and aggregated, yielding six characteristics of the class as a constructivist learning environment (as perceived by the students). We examined the explanatory power of these class characteristics with respect to individual students' mathematics achievement, as measured in TIMSS 2003. Unlike the international TIMSS 2003 design, the Flemish study included two classes in most schools, which made it possible to use three-level models of students within classes within schools to model the relationships. Also, a Flemish extension of the TIMSS questionnaire for teachers included a measure of each teacher's perception of the class as a constructivist learning environment. The teachers' viewpoints provided an interesting addition to the students' perceptions.

Theoretical model related to powerful teaching

When designing powerful learning environments, we need to search for an answer to the following basic educational dilemma. On the one hand, students need to initiate, direct, and regulate the learning processes as much as possible themselves. This becomes possible in activating and student-controlled learning environments (see, for example, Scardamalia & Bereiter, 1991). On the other hand, students need coaching—sometimes even external steering—during these learning processes. A certain degree of teacher control therefore has to be built into learning environments (Schwartz & Bransford, 1998).

De Corte (2000, p. 254) provides a general research-based operationalization of this dilemma through a series of guiding educational principles. We took these guidelines as a starting point for our own research. Based on an extensive literature study, we established a list of different teaching activities that the authors of the various studies identified as important for fostering learning for understanding. We next considered, using a limited number of components, how to describe these many different activities. We identified 13 dimensions within the entire group of teaching activities. Table 1 sets out these dimensions and their associated important teaching activities.

With the aim of designing an iterative, gradually more complex model, we next summarized these 13 sub-dimensions into a basic—but comprehensive—model, with fewer main dimensions. We started by comparing our 13 dimensions with Bransford, Brown, and Cocking's (2000) model of effective learning environments. This model describes "four perspectives on learning environments that seem particularly important given the principles of learning" (p. 133). In the following section, we discuss the four main dimensions of our model, each time giving a brief account of their link to the corresponding perspective in the Bransford et al. model. We indicate certain differences, and describe the sub-dimensions with which we further operationalized the different aspects of each main dimension.

The four main dimensions of the model

1. Motivate students to exert learning effort

This forms the first main group of educational activities that are a major point of attention during the creation of powerful learning environments. In general, the different teaching activities within a total educational approach should motivate students to exert effort to engage in learning activities and to sustain these efforts. Without motivation, there is no deep-level learning; effort is therefore essential (Stipek, 2002). Learning motivation can be fostered in several ways and at different levels. First, there is learning motivation that can be facilitated through initiatives executed within a short time period, for example, by arousing interest (Sub-dimension 1). However, this kind of motivation can be short-lived. Initial interest often disappears when real effort is required. Teachers and teaching coaches therefore need to arouse and maintain a deeper motivation for learning among students (Sub-dimension 2) (Boekaerts, 2001).

Creating this level of motivation requires us to anticipate and capitalize on the different characteristics and needs of learners—their interests, their prior knowledge, their learning strategies. It is also important that we adapt the curriculum to the interests of the learners as far as is possible—to differentiate the educational approach according to the individual needs of learners. Bransford et al. (2000) refer to the importance of these activities as part of their description of the "learner-centered" aspect of learning environments. They refer to the

Table 1: Model of Powerful Teaching Operationalized in 13 Sub-dimensions

Sub-dimension 1: Motivate by arousing interest Give realistic examples Give examples from the news Use realistic situations as a starting point for lessons Base assignments on realistic cases Create a feeling of usefulness General ability to arouse interest	*Sub-dimension 2: Maintain motivation for learning* Create challenging learning tasks Create opportunities to be creative Create pleasant lessons Create a pleasant variation in educational approaches Avoid coming across as a know-all
Sub-dimension 3: Motivate through a structured approach Keep attention by sequencing tasks and explanations Make use of learning conversations Explain in a clear and structured way Keep the class under control Students understand learning content through the approach	*Sub-dimension 4: Activate towards self-regulated learning* Make students think for themselves instead of listening Explain everything first; students have to listen (-) Make students solve difficult assignments Assignments being too difficult (-)
Sub-dimension 5: Activate toward connecting contents Ask for connection between learning contents Give assignments that require the use of prior knowledge Make students search for information to solve problems	*Sub-dimension 6: Give activating feedback during tasks* Give hints without giving the answers Monitor and point to mistakes Motivate with hints to enable a search for a solution
Sub-dimension 7: Give structuring feedback before tasks Repeat prior knowledge necessary for new content Connect prior and new knowledge Indicate what knowledge and skills have to be mastered	*Sub-dimension 8: Give activating feedback after tasks* Give considerable explanation of the problem-solving process Give answers on paper and make the students correct their tasks Encourage students to reflect on the causes of their mistakes Repeat content if necessary
Sub-dimension 9: Structure contents in an organized whole Make students synthesize learning content Show students how to synthesize	*Sub-dimension 10: Create cooperative learning* Make students work cooperatively on assignments Organize cooperative learning in an appropriate way Have students solve problems through group dynamics Teach how to cooperate
Sub-dimension 11: Make use of group discussions Organize group discussions Make students discuss the learning contents Make students substantiate their arguments	*Sub-dimension 12: Tailor evaluation to in-depth knowledge acquisition* Avoid emphasis on knowledge reproduction Avoid literal repetition of questions Stress the importance of understanding the subject matter Require the application of knowledge to new problems
Sub-dimension 13: Give activating feedback about tests Provide ample discussion of tests Make test answers available on paper Make students correct their tests using solutions on paper	

need for environments "that pay careful attention to the knowledge, skills, attitudes and beliefs that learners bring to the educational setting" (p. 133). They also stress that "accomplished teachers 'give learners reason', by respecting and understanding learners' prior experiences and understandings, assuming that these can serve as a foundation on which to build bridges to new understandings" (p.134).

Creating opportunities for learners to determine their own goals and learning paths is also important. As a consequence—and because humans normally strive to integrate themselves into their environments and therefore need and want to learn the necessary skills to do so—learning tasks must be sufficiently relevant. They should give students opportunities to learn something they can use in the real world (Brown, 1996). Of course, what one person wants and needs to learn in order to function in the real world differs from what other people want and need. A constant, however, is that every learner wants to learn in order to become a part of a community; social skills are thus essential. Learners accordingly need opportunity to experiment with different roles within groups of learners. They also need to experiment with using/building certain knowledge in groups, because real insight relating to different domains of knowledge is often bound to how social communities use/co-construct that insight.

What all this means for students' learning effort is that motivation to exert such effort can be created in meaningful communities of learners (Brown & Campione, 1996; Slavin, 1991). This consideration is indicated in the Bransford et al. model through the "community-centered" perspective, which refers to the need for environments to be such that learning is "enhanced by social norms that value the search for understanding and allow students (and teachers) the freedom to make mistakes in order to learn" (Bransford et al., 2000, p. 145). Creating co-operative learning environments (see Sub-dimensions 10 and 11) can have these positive effects, but only if they are realistic. Thus, we need to create group tasks that are meaningful for the learners concerned and/or to set these tasks within environments in which the learners can experiment with different roles and social interactions in a constructive way. However, cooperative tasks can also have an activating quality, even if they happen in less realistic settings and involve relatively traditional learning activities. We discuss this consideration below in relation to the second main dimension, "Activate self-regulated learning."

The above two sub-dimensions are important in creating long-lasting motivation. However, other educational activities are essential to prevent this motivation being undermined. Learners need to feel that their efforts will lead them to understand the purpose of what they are doing. Ongoing feelings of uncertainty decrease their efforts (Good & Brophy, 2000). Therefore, an effectively organized and structured approach to lessons is needed (see also Sub-dimension 3, discussed in relation to Main Dimension 4 below). However, the giving of appropriate feedback can reduce this uncertainty—even within an authentic problem-solving environment. We discuss this matter in relation to Sub-dimensions 6, 7, and 8, as part of the third main dimension, "Give feedback and coach."

2. Activate students' self-regulated learning

Creating motivation to exert effort in learning does not automatically produce the type of deep understanding that learners can apply in different situations. Learners also have to be engaged in effective learning tasks deliberately designed to meet this goal. A first important aspect of these tasks is that they should be what Bransford et al. (2000) call "learner-centered," which aligns with learning environments "that pay careful attention to the knowledge, skills, attitudes and beliefs that learners bring to the educational setting" (p. 133). As the authors explain, "By selecting critical tasks that embody known misconceptions, teachers can help students test their thinking and see how and why various ideas might need to change. The model is one of engaging students in cognitive conflict and then having discussions about conflicting viewpoints" (p. 134). In our model, this dimension incorporates all teacher activities that aim to engage students in learning tasks requiring them to ask questions about or link their prior knowledge to new information so that they can construct meaning. Our model also assumes that students, when engaging in assignments (an activity that requires greater self-regulation on their part), will make more use of their prior knowledge and will construct meaning in a more active way (Schunk, 2001).

Nonetheless, self-regulating processes have to be learned. Teachers accordingly have to move gradually when implementing more student regulation of the complete learning process (Corno & Randi, 1997). As a consequence, learning assignments should be sufficiently complex in order to activate students toward self-regulated learning rather than having them just apply memorized knowledge mindlessly or adding little pieces of information without considering the whole (see Sub-dimension 4). Here, students should aim to connect new knowledge with prior knowledge—and to do so within an expanding, well-organized body of knowledge (Shuell, 1996) (see Sub-dimension 5 and Sub-dimension 9 in Main Dimension 4, "Structure and steer"). However, these activating tasks should not be so complex that students pick up misconceptions and/or become de-motivated (Schwartz & Bransford, 1998) (see Sub-dimension 4 in Main Dimension 1, "Create motivation"). Furthermore, these tasks should be supported by feedback given during the tasks, and this feedback must be such that it activates students to take the next step in their learning process in a self-regulated way (see Sub-dimension 6). If necessary, teachers should also provide feedback after students have completed assignments, and this feedback too should be activating (see Sub-dimension 8 in Main Dimension 3). While part of this feedback will focus on the learning content, another important part will center on the metacognitive learning processes that the learners have to go through (Marton & Booth, 1997).

A co-operative learning environment thus bears special features that can activate learners toward self-regulated learning (see Sub-dimensions 10 and 11). It can encourage them to articulate their thoughts toward the co-learners—to connect these thoughts to what the group has already discussed (Slavin, 1991). In addition, social pressure can encourage learners to make more effort to master the learning contents at hand and thereby stimulate self-regulated learning. However, many aspects of a co-operative learning environment can also can hinder these activating effects. Therefore, co-operative learning should be organized and monitored in an appropriate way (Druckman & Bjork, 1994), and the communication within the groups of students should stay focused on the learning content (Cohen & Lotan, 1995). Assessment can also play an important role in activating students (see Sub-dimension 12)—students adapt their learning efforts toward the tests they will sit. Therefore, it is important that assessments ask students to use their knowledge to solve (new) problems. This practice signals to students that they have to aim for this goal when studying/learning in class. Attempts to achieve this goal increase students' extrinsically motivated learning effort (Dochy & McDowell, 1997).

3. Give students feedback and coach them

For reasons of certification, summative assessment is necessary to indicate whether learners have reached a certain level of mastery. However, the most important educational aspect of assessment lies in its formative use. Bransford et al. (2000) refer to this when describing their "assessment-centered" perspective of a learning environment. They refer to formative environments as those that "provide opportunities for feedback and revision ... [and in which] what is assessed must be congruent with one's learning goals" (p. 140). Our model stresses assessment as just one important constituent of giving effective feedback.

By "assessment," we mean those activities that allow the teacher (or learning coach) to determine where a learner stands at a certain point in time. As Rice (1991) puts it, what are the learner's prior experiences, knowledge, and misconceptions concerning the learning content? What metacognitive skills relevant to the learning content in question does he or she have? By referring to the outcomes of this assessment, teachers can develop and/or select adjusted tasks (see Sub-dimensions 4 and 5) and give adjusted structuring feedback before, during, and after each learning assignment (see Sub-dimensions 7, 6, and 8 respectively). This feedback should focus on learning content as well as on providing learners with the support necessary to build self-regulated learning processes (Schunk, 2001). Constant assessment of the attained level also should be built into the teaching–learning process, with feedback given in a manner that activates learners to take the next self-regulated step in their learning processes (Vygotsky, 1978). As long as they are adapted to the needs of the learner, limited hints or more extensive explanation (for instance, to indicate complex misconceptions) can be given. In each case, teachers and learning coaches have to create situations that allow learners to actively use the feedback they receive to facilitate further learning (Winne, 2001). What is also clear is that summative assessments require appropriate feedback (see Sub-dimension 13).

4. Structure and steer

Strictly speaking, there is no need for a fourth dimension in our model. However, we created this separate dimension to stress the importance of teachers deliberately organizing the activities set under the three former dimensions over a longer period of teaching. The reason for doing this is to foster students' deep-level learning and their ability to transfer that learning to various situations, activities, and other bodies of knowledge. Our fourth dimension therefore resembles, to a certain degree, the "knowledge-centered" perspective in the model developed by Bransford et al. (2000). For Bransford and colleagues, this perspective relates to environments "which take seriously the need to help students become knowledgeable by learning in ways that lead to understanding and subsequent transfer. ... [The perspective] requires well-organized bodies of knowledge that support planning and strategic thinking" (p. 136).

In our model, the fourth dimension incorporates all teacher activities that aim to structure and steer the learning processes. For teachers and learning coaches, the first step relevant to this dimension involves structuring the teaching activities discussed under Dimensions 1 to 3 above:

- Learn to know your learners (their prior experiences, knowledge, interests, etc)
- Create (and/or maintain) motivation to exert effort to learn
- Create activating, self-regulated learning tasks, adjusted to the learners at hand
- Give activating feedback before the task
- Constantly assess the learners, as far as possible during the tasks, and after the tasks
- Give activating feedback during and after tasks
- Assess the whole learning process and the learning outcomes and decide on the assignment of new learning tasks (while maintaining motivation), etc.

We have, of course, already discussed aspects of these structuring tasks, most notably activating learners in a way that allows them to connect learning contents (Sub-dimension 5), and giving learners structuring feedback before, during, and after learning assignments (Sub-dimensions 6, 7, and 8). In Sub-dimension 9, we stressed other aspects: the need to organize a structuring approach within the context of an elaborated view of the curriculum, and the need to encourage students to structure their learning contents as an organized whole (Reigeluth, 1999). We have to remember, however, that (some) learners will be unable to perform certain learning activities independently and will have to be steered toward doing so (Singley & Anderson, 1989). There also will be a need to "automate" certain insights for learners to prevent them experiencing unnecessary cognitive overload (Anderson, 1993). Teachers also need to work directly on any misconceptions learners may develop. Another important point is that teachers cannot bypass the need to take into account the available learning time and resources. Often, this requirement means that teachers have to search for an effective trade-off between goals and educational approaches.

How teachers structure their educational approach is an essential motivator (or otherwise) for students. Students expect that the teaching approach will help them begin to understand the subject matter and to feel that they can do the learning assignments, and that they gain these feelings preferably at an early stage of the lesson or the assignment. For most students, these positive feelings of mastering the situation align with clear explanations from the teacher of the subject or topic at hand and with the teacher maintaining good class management (Good & Brophy, 2000). Teachers can create these feelings more easily in traditional educational settings because students are more used to these settings and because they are easier for teachers to manage. However, these conditions also have to be met in relation to the more complex learning tasks that activate self-regulated learning. If too many lessons are unstructured and unclear, students lose motivation (see Sub-dimension 3) (Schwartz & Bransford, 1998).

Measuring perceptions of the learning environment by using data from TIMSS 2003 for Flanders

In common with students in the other countries participating in TIMSS 2003 (see, for example, Mullis, Martin, Gonzales, & Chrostowski, 2004), the Flemish students sat achievement tests in science and in mathematics. In addition, students, teachers, and principals filled out the international TIMSS 2003 questionnaires.

In Flanders, a numerical and spatial intelligence test for students and a questionnaire for parents supplemented the internationally defined measurements of TIMSS 2003. In addition, some elements were added to the international student, teacher, and principal questionnaires. One of the additions to the student questionnaire is central to the present paper, namely, an instrument designed to measure students' perceptions of the learning environment. The construction of this instrument, which consists of 33 four-point items (refer Table 2), was based on the theoretical model described in the earlier part of this paper. This detailed 33-item instrument also relates to a six-item precursor administered in the Flemish part of TIMSS 1999 (Van den Broeck, Opdenakker, & Van Damme, 2005) in that it includes these six previously used items (see Table 2, items 3, 5, 15, 27, 32, and 33.) Moreover, six closely similar items were included in the Flemish extension of TIMSS 2003 questionnaire for teachers (see Table 3).

Table 2: Six Learning Environment Scales from the Flemish Part of the TIMSS 2003 Student Questionnaire

“Activation” Scale (11 items, α = 0.76): In the math class …

… the teacher asks about relationships between different parts of the subject material during tasks. (8)

… the teacher explains as little as possible, letting us think and try for ourselves instead. (11)

… during team work or when I am working on my own, the teacher inquires about the difficulties I encounter while solving a problem. (15)

… the teacher points out connections between new and previously treated subject matter. (16)

… the teacher gives tasks that encourage us to keep looking for a solution. (17)

… the teacher’s tests require us to apply the subject matter to new problem contexts. (19)

… the teacher gives small clues that help us to find solutions by ourselves. (22)

… when an assignment or test goes somewhat wrong, I am encouraged to think about what has caused the problem and what I can do about it. (24)

… when we start with a new subject, the teacher takes time to repeat previous subject matter that will be relevant to the new topic. (30)

… it is important to understand the subject matter in order to obtain good marks on tests. (31)

… during team work or when I am working on my own, the teacher inquires about the time I need to solve a problem. (33)

“Clarity” Scale (7 items, α = 0.82): In the math class …

… the teacher bears in mind students’ remarks when searching for suitable assignments or practice materials. (3)

… the teacher is able to explain new topics in a clear and well-organized manner. (6)

… we get the opportunity to explain our solution to the teacher. (7)

… the teacher keeps the class under control. (9)

… the teacher tries to make us understand new subject matter by alternating questions to the class with explanations. (23)

… the teacher takes into account students’ answers. (25)

… it’s thanks to the teacher’s approach that I understand the subject matter well. (29)

“Authenticity” Scale (3 items, α = 0.74): In the math class …

… the teacher gives examples of situations in daily life where the subject matter can be applied. (1)

… each new chapter starts with examples from daily life that clarify the new subject. (5)

… situations are described that can happen in the real world and that need a mathematical solution. (14)

“Motivation” Scale (4 items, α = 0.76): In the math class …

… the teacher makes sure that I get interested in the subject matter. (2)

… the teacher uses an agreeable diversity of approaches in his/her teaching. (4)

… we work in a pleasant manner. (12)

… I feel that the subject matter will be useful to me later. (21)

“Feedback” scale (3 items, α = 0.70): In the math class …

… the teacher explains the solution after an exercise. (18)

… the teacher repeats the subject matter when it is not properly understood by some students. (26)

… the teacher clarifies errors in tests. (28)

“Cooperation” Scale (2 items, α = 0.74): In the math class …

… we have the opportunity to ask other students to explain their way of solving a problem. (27)

… we have the opportunity to discuss our approach to math problems with other students. (32)

Items from the 33-item Questionnaire Not Used in the Scales: In the math class …

… after a test we receive the solutions on paper. (10)

… the teacher’s way of handling the subject matter costs me too much effort. (13)

… we have to correct our own tests by means of the right answers on paper. (20)

Note: For each item, the number between brackets indicates its position in the list of 33 items presented to the student. These are four-point items, with the following potential answers: “Nearly always,” “‘Rather often,” “Now and then,” or “‘Never.” The internal consistency of the scales has been expressed by means of Cronbach’s α. The value of α reported here was calculated for the complete Flemish TIMSS 2003 sample, i.e., A-stream and B-stream together.

Table 3: Six Items about the Learning Environment in the TIMSS 2003 Teacher Questionnaire

During math lessons in this class …
… I introduce each new chapter or new topic by means of varied examples from daily life. (a)
… I give students the opportunity to discuss their approach to math problems with other students. (b)
… I give students the opportunity to ask other students to explain their way of solving a problem. (c)
… I bear in mind students' remarks when searching for suitable assignments or practice materials. (d)
… I organize team work or individual work to get a picture of the actual difficulties the students encounter while solving a problem. (e)
… I organize team work or individual work to get a picture of the actual time needed by students to solve a problem. (f)

Note: These are five-point items. Potential answers: "Almost never," "Seldom," "Sometimes," 'Often," or "Almost always." Items a, b, c, d, e, and f are closely similar to Items 5, 32, 27, 3, 15, and 33 in the student questionnaire.

Sample, variables, and analyses

The sample

The Flemish sample of the TIMSS 2003 study consisted of 5,213 Grade 8 students in 276 classes in 148 schools. In Flanders, Grade 8 is divided into a general or academic track (A-stream) and a vocational track (B-stream). The A-stream part of the TIMSS sample had 4,521 students in 228 classes in 119 schools, and the B-stream had 692 students in 48 classes in 29 schools. The remainder of this present text refers to the A-stream, unless we explicitly indicate otherwise.

The fact that the Flemish sample contained more than one class per school constituted an extension to the international study design. In the A-stream (B-stream), we had 109 (19) schools, with two classes in the sample, and 10 (10) schools represented by a single class. As in the Flemish TIMSS 1999 study (Van den Broeck et al., 2005), this design enabled us to distinguish between the class and the school level in the analyses. In most (i.e., 92) of the 128 schools with two classes in the sample, the two classes each had a different mathematics teacher.

Characteristics of the learning environment in the class

We employed exploratory factor analysis to structure and reduce the 33-item measure of the learning environment as perceived by the students. (In the next sections, we present and discuss the results from the perspective of the theoretical framework.) Next, we calculated student-level scale scores aggregated (averaged) for each class. Thus, each scale yielded a class-level learning environment variable.

We also calculated the scores on the six-item learning environment scale based on the teachers' questionnaire (Table 3), and this yielded an additional class-level learning environment variable. We then compared this variable with a variable derived from the six closely similar items in the student questionnaire (which were also available in the Flemish TIMSS 1999 study). We constructed this latter variable in the same way as the class-level variables derived from the 33-item questionnaire, that is, by calculating student-level scale scores and averaging it per class.

Mathematics achievement

Here, we used the TIMSS 2003 Rasch score for mathematics achievement (Martin, 2005) as a response variable to which we sought to relate the class-level learning environment variables. For the A-stream of Grade 8 in Flanders, this mathematics score was available for 4,328 students in 224 classes in 119 schools, with a mean value of 152.8 (0.4) and a standard deviation of 8.3. The total variance partitioned into 69% variance between students within a class, 17% between classes within a school, and 15% between schools.

Further analyses

Having explored the structure of the 33-item measurement of the learning environment characteristics by means of a factor analysis, our next aim was to relate this structure to the theoretical model that served as our point of departure. Before doing so, however, we studied these dimensions further by examining the corresponding learning environment variables in two ways. First, we estimated their within-school and between-classes differences, and then we determined their explanatory power with respect to the mathematics achievement of individual students.

Our final learning environment variables refer to the class level. For each variable, we estimated a two-level model of classes within schools in order to partition its variance into a within-school/between-classes component and a between-schools component. We assessed the explanatory power of each learning environment variable in relation to mathematics achievement by means of a three-level model (of students within classes within schools).

Results

The exploratory factor analysis yielded an acceptable six-factor solution, described in Table 2 (above). For each factor, the table lists all items that loaded significantly on that factor and that were included in the corresponding scale.

Table 4 shows the correlations between the class-level variables derived by aggregating the scales defined in Table 2. We also added the correlations with the variable derived from the six-item scale defined in TIMSS 1999, which combined items from the scales "activation" (Items 15 and 33), "clarity" (Item 3), "authenticity" (Item 5), and "cooperation" (Items 27 and 32), and which had an internal consistency of α = 0.76. Finally, Table 4 shows the correlations with the six-item learning environment scale from the teacher questionnaire (α = 0.74). Note that the correlation between the teachers' scale (CT) and the corresponding variable (CP) derived from the student questionnaire as well as the correlations between CT and the other variables determined by the students are clearly smaller than the correlations among the variables from the student questionnaire.

Table 5 summarizes the basic statistics of these eight class-level learning environment variables. It also shows the partitioning of the variance of each variable into a within-school/between-classes component and a between-schools component. Although there are sizeable between-schools differences for some variables, the between-classes component is dominant for the seven variables derived from the student questionnaire; for the variable CT, the two components are nearly equal. We therefore concluded that the learning environment variables, in keeping with their intended meaning, were true class-level variables (rather than school-level variables). Finally, Table 5 lists the coefficients of the learning environment variables in their role as predictors of mathematics achievement. None of the coefficients differed significantly from zero at the 5% level, although some had a *p*-value below 10%.

Discussion

Table 5 shows that none of the coefficients in the single predictor models was significantly different from zero at the 5% level. Our conclusion here was that the dependent variable in this first exploration was not optimally suited for our purpose. First, the mathematics achievement score referred to an international test that was less than perfectly adapted to the Flemish curriculum. Second, we could envisage a more sensitive analysis that would involve restricting the dependent variable to those items—such as problem-solving items—for which we could expect the constructivist learning environment to have a positive effect. In the remainder of the discussion, we consider the coefficients with a *p*-value below 10% as (marginally) significant from zero.

Table 4: Correlations between Class-level Learning Environment Variables

	CLAR	AUTH	MOTIV	FEEDB	COOP	CP[a]	CT[b]
ACTIV	0.80	0.56	0.72	0.77	0.57	0.78	0.16
CLAR		0.47	0.75	0.88	0.55	0.71	0.16
AUTH			0.59	0.41	0.40	0.69	0.14
MOTIV				0.67	0.60	0.76	0.14
FEEDB					0.52	0.66	0.12
COOP						0.85	0.19
CP							0.19

Notes: [a] CP is based on the TIMSS 1999 six-item measure of the student's perception of the learning environment (see Section 3).

[b] CT is the six-item scale from the teachers' questionnaire.

Each correlation refers to up to 228 A-stream classes.

The variables have been defined in a "positive" sense: a class with score 4 experiences more "activation" than a class with score 1, etc.

Table 5: Characteristics of the Class-level Learning Environment Variables

	Basic statistics			Var. comp.[c]			Single predictor model[d]		
Variable	*N*[e]	Mean	*SD*	Class	School	*N* [f]	Coeff.[g]		*p*[h]
Activation	227	2.56	0.19	67%	33%	4,328	2.9	(1.7)	9%
Clarity	227	3.01	0.35	69%	31%	4,328	1.8	(0.9)	6%
Authenticity	227	2.08	0.33	69%	31%	4,328	-1.5	(1.0)	14%
Motivation	227	2.34	0.34	95%	5%	4,328	-0.1	(0.9)	89%
Feedback	227	3.24	0.35	65%	35%	4,328	1.1	(0.9)	25%
Cooperation	227	2.15	0.36	91%	9%	4,328	-0.8	(0.9)	39%
CP[a]	227	2.15	0.28	86%	14%	4,328	-1.9	(1.1)	10%
CT[b]	216	3.15	0.65	47%	53%	4,088	0.9	(0.5)	9%

Notes: [a] CP is based on the TIMSS 1999 six-item measure of the student's perception of the learning environment.
[b] CT is the six-item scale from the teachers questionnaire. CTpotentially ranges from 1 to 5, whereas the other variables have a potential range from 1 to 4.
[c] Variance components according to a two-level model (classes within schools).
[d] Three-level model (students within classes within schools) explaining mathematics achievement by means of a single predictor.
[e] Number of classes.
[f] Number of students.
[g] The variables have been defined in a "positive" sense: a class with Score 4 experiences more "activation" than does a class with Score 1, etc. Between brackets: standard error.
[h] The *p*-value refers to a deviance test comparing the model with and without the predictor.

The composition of the six scales in Table 2 largely agreed with the combination of several of the sub-dimensions derived from the theoretical framework. The theoretical model assumes that activating students toward self-regulated learning is the main determinant in the design of powerful learning environments. Teachers who assign tasks designed with this objective in mind are in tune with the constructivist nature of learning processes. Students are asked to use their acquired knowledge to put meaning to new information and/or are prompted to solve related or even new problems. The "activation" scale comprises a number of aspects (sub-dimensions) of the main dimension "Activate students' self-regulated learning" in the theoretical framework. This process first involves assigning tasks with a sufficient degree of complexity so that students have to think for themselves (Items 17, 11). This consideration fits the fourth sub-dimension. Second, the activation scale stresses the importance of figuring out the connections between components of the subject matter, either by the teacher (Items 16 and 30) or at the teacher's request during assignments (Item 8). This element corresponds to the fifth sub-dimension. Next, the scale also expresses the importance of providing activating supervision of these assignments (Items 22, 15, 33). Here, we find Sub-dimension 6. Finally, the need for activating assessment is represented in the scale, first through the items referring to the complexity of the tests (Items 19 and 31, see Sub-dimension 12), and second through the concept of tests or assignments as learning moments (Item 24). According to the model, activation is essential for in-depth learning. The three-level model revealed a significant effect (p = 0.09) in this regard.

An important idea in the theoretical framework is that the process of activating students toward self-regulated learning must be such that it does not impair students' confidence in their ability to master the subject matter. This consideration not only implies that a clear-cut and structured approach is essential (Items 6, 29, 23, and 9), but it also tallies with Sub-dimension 3. Such an approach is also likely to foster motivation. In a similar vein, the theoretical model draws attention to students' need for understanding and respect during their efforts to master the subject matter. In addition, students need clarity with regard to their role in the educational process. In our model, this need is embodied in the aspect "Avoid coming across as a know-all" (Sub-dimension 2) and in "Keep attention by sequencing tasks and explanations" (Sub-dimension 3). In the clarity scale, we find the corresponding items (Items 25, 3, and 7). Thus, according to our theoretical framework, the clarity of the learning environment is an equally important a predictor of achievement as is activation, and it therefore should be balanced with activation. The multilevel analysis yielded a significant effect ($p = 0.06$).

Items 1, 5, and 14 constitute the "authenticity" scale (Table 2) and refer to the advisability of promoting interest by the use of real-life examples. This advice aligns with Sub-dimension 1 of the theoretical model. The scale "Motivation" (Items 12, 4, 2, and 21) fits Sub-dimension 2, "Maintain motivation for learning." This scale mainly concerns the fostering of long-term motivation by means of agreeable and diversified instruction. While we consider this an important factor in the implementation of a pleasant learning process, we do not regard it as essential for attaining in-depth mathematical insight. Of course, this does not preclude deadly dull and repetitive lessons from having the detrimental effect of preventing learning because of a total lack of attention among the students and perhaps even rebellious behavior. In our analyses, we did not find a link between the authenticity and motivation scales and mathematics achievement ($p = 0.14$ and $p = 0.89$, respectively). This may mean that most teachers do not let the learning situation come to that described, but rather that they provide enough variation in the lessons to encourage the students to put in sufficient effort.

The scale "feedback" (Items 26, 28, and 18) corresponds to the eighth sub-dimension "Give activation feedback after tasks." We consider this dimension particularly important, especially for students who have yet to benefit from good supervision during activating assignments. The importance of the teacher giving carefully targeted and limited hints, especially during activating tasks, is a point that is stressed by our theoretical framework. The effect of the "feedback" variable, however, was not significantly different from zero ($p = 0.25$), but note that the feedback aspect is partly covered by the "activation" scale (Items 22, 15, and 33) and also by the quality of the interaction with the teacher as expressed in the "clarity" scale (Items 25, 3, 7), which did have some predictive power.

The sixth scale in Table 2—"cooperation" (Items 32 and 27)—matches the 10th sub-dimension of the model (Table 1). The theoretical framework suggests that cooperation has a considerable activating value, but we did not find a significant effect in our analysis ($p = 0.39$). One possible explanation is that the activating value of cooperation is conditional on the quality of the interaction (Slavin, 1991), an aspect not explicitly included in the 33-item questionnaire.

Finally, with regard to the (limited) opportunity within the Flemish TIMSS 2003 assessment to compare the point of view of the students and their teachers concerning the learning environment, we note that while both views were positively related, this relationship was rather weak. Thus, it seems desirable for future studies to question teachers about this learning environment in their classes as thoroughly as we questioned the students in TIMSS 2003.

The main conclusion of this contribution is that some important features of the theoretical framework seem to be reflected in the TIMSS 2003 data. There is substantial agreement between the sub-dimensions postulated by the model and the factor structure found in the data. Also, there is some indication that the aspects of the learning environment that are theoretically the most essential—namely, activation toward self-regulated learning and clarity of the learning environment—relate to achievement in mathematics. However, we could not find any evidence in the TIMSS data for several other potential links between the learning environment and achievement put forward by the theoretical framework. Additional research is necessary to determine if restricting the dependent variable to those items on which a positive effect of a constructivist learning environment may be expected influences the results.

References

Anderson, J. R. (1993). *Rules of the mind.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., Greeno, J. G., Reder, L. M., & Simon, H. A. (2000). Perspectives on learning, thinking and activity. *Educational Researcher, 29*(4), 11–13.

Boekaerts, M. (2001). Pro-active coping: Meeting challenges and achieving goals. In E. Frydenberg (Ed.), *Beyond coping: meeting goals, visions and challenges* (pp. 129–147). Oxford: Oxford University Press.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience and school.* Washington DC: National Academy Press.

Brown, A. L. (1996). Distributed expertise in the classroom. In G. Salomon (Ed.), *Distributed cognitions: Psychological and educational considerations* (pp. 188–228). Hillsdale, NJ: Erlbaum.

Brown, A. L., & Campione, J. C. (1996). Psychological theory and the design of innovative learning environments: On procedures, principles, and systems. In L. Schauble & R. Glaser (Eds.), *Innovations in learning: New environments for education* (pp. 289–325). Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, E., & Lotan, R. (1995). Producing equal-status interaction in the heterogeneous classroom. *American Educational Research Journal, 32*, 99–120.

Corno, L., & Randi, J. (1997). Motivation, volition and collaborative innovation in classroom literacy. In J. Guthrie & A. Wigfield (Eds.), *Reading, engagement: Motivating readers through integrated instruction* (pp. 14–31). Newark, DE: International Reading Association.

De Corte, E. (2000). Marrying theory building and the improvement of school practice: A permanent challenge for instructional psychology. *Learning and Instruction, 10*, 249–266.

Dochy, F., & McDowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation, 23*, 279–298.

Druckman, D., & Bjork, R. A. (Eds.). (1994). *Learning, remembering, believing: Enhancing team and individual performance.* Washington DC: National Academy Press.

Good, T., & Brophy, J. (2000). *Looking in classrooms* (8th ed.). New York: Longman.

Martin, M. O. (Ed.). (2005). *TIMSS 2003 user guide for the international database.* Chestnut Hill, MA: Boston College.

Marton, F., & Booth, S. (1997). *Learning and awareness.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Mullis, I. V. S., Martin, M. O., Gonzales, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report.* Chestnut Hill, MA: Boston College.

Reigeluth, C. M. (1999). The elaboration theory: Guidance for scope and sequence decisions. In C. M. Reigeluth (Ed.), *Instructional design theories and models* (pp. 335–381). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rice, D. C. (1991). *The design and validation of an instrument to identify pre-service elementary teachers' intuitive and school knowledge of the concepts of surface area/volume and states of matter.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Lake Geneva.

Richardson, V. (2003). Constructivist pedagogy. *Teachers College Record, 105*(9), 1623–1640.

Scardamalia, M., & Bereiter, C. (1991). Higher levels of agency for children in knowledge-building: A challenge for the design of new knowledge media. *Journal of the Learning Sciences, 1*, 37–68.

Schelfhout, W., Dochy, F., Janssens, S., Struyven, K., Gielen, S., & Sierens, E. (2006a). Educating for learning-focused teaching in teacher training: The need to link learning content with practice experiences within an inductive approach. *Teaching and Teacher Education, 22*(7), 874–897.

Schelfhout, W., Dochy, F., Janssens, S., Struyven, K., Gielen, S., & Sierens, E. (2006b). Towards an equilibrium model for creating powerful learning environments: Validation of a questionnaire on creating powerful learning environments. *European Journal for Teacher Education, 29*(4), 471–504.

Schunk, D. (2001). Social cognitive theory and self-regulated learning. In B. Zimmerman & D. Schunk (Eds.), *Self-regulated learning and academic achievement.* Mahwah, NJ: Lawrence Erlbaum Associates.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction 16*(4), 475–522.

Shuell, T. (1996). Teaching and learning in a classroom context. In T. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 726–764). New York: MacMillan.

Singley, M. K., & Anderson, J. R. (1989). *Transfer of cognitive skill.* Cambridge, MA: Harvard University Press.

Slavin, R. (1991). Group rewards make group work. *Educational Leadership, 5*, 89–91.

Stipek, D. (2002). *Motivation to learn: Integrating theory and practice.* Boston, MA: Allyn and Bacon.

Van den Broeck, A., Opdenakker, M.-C., & Van Damme, J. (2005). The effects of student characteristics on mathematics achievement in Flemish TIMSS 1999 data. *Educational Research and Evaluation, 11*(2), 107–121.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. Zimmerman & D. Schunk (Eds.), *Self-regulated learning and academic achievement.* Mahwah, NJ: Lawrence Erlbaum Associates.

The relationship between classroom practices related to homework and student performance in mathematics in TIMSS 2003

Ming-Chih Lan and Min Li
University of Washington
Seattle, Washington, USA

Abstract

This study examined the relationship between classroom practices on homework and student performance in mathematics by conceptualizing homework practice as a four-dimensional construct. We used the United States section of data sets for students and the teachers who taught them from the Trends in International Mathematics and Science Study (TIMSS) in 2003 to examine this relationship; our method of analysis was hierarchical linear modeling (HLM). The results showed, at the student level, that seven of the 10 variables we selected from the database were important explanatory variables accounting for variation in student performance. These variables included "language used at home," "mother's level of education," "student's self-efficacy relating to mathematics," and "time spent on homework at home." At the teacher level, three of the four dimensions that made up our proposed construct of homework practice, namely Dimension I, "the frequency of homework assigned," Dimension II, "the amount of time the teacher wanted students to spend on the homework assigned," and Dimension IV, "the strategies teachers used in relation to homework," positively correlated with student learning. The dimension that did not show a significant correlation was Dimension III, "the focus of the homework provided." In general, the results partially supported the notion that homework practice is a multi-dimensional construct, and implied that it is possible to predict higher student performance by increasing both the frequency and the amount of homework assigned to students. We also successfully created a simpler HLM model, with fewer variables at the student level and a transformed "homework intensity" variable that combined Dimensions I and II. With reference to classroom practices related to homework, this study found that increasing the frequency and amount of homework assigned led to higher student scores on the TIMSS mathematics assessment. Thus, for teachers, a productive way to support student learning is to assign homework more frequently and with more work instead of paying attention to one at the exclusion of the other. In addition, the study indicated that the more often teachers spent time in class having students correct their own homework, then the better their students achieved in mathematics.

Introduction

Cooper (1989) defines homework as tasks that teachers assign to students to carry out during non-school time. Homework not only helps students master what they have learned in class, but also extends the time they have available to learn the subject matter. The information self reported by the teachers randomly sampled from all over the United States for participation in the Trends in International Mathematics and Science Study (TIMSS) 2003 showed that nearly all Grade 8 mathematics teachers (over 99%) assigned homework to their students (Martin, 2005). Because homework is a commonly used classroom practice, and therefore is a vital part of instruction, it is important that teachers and students know how to use homework effectively to promote student learning.

Typically, homework serves a twofold purpose. On the one hand, homework is a formative assessment tool that provides teachers with feedback that allows them to adjust their instruction and that provides students with a means of improving or consolidating their learning. Second, homework is a type of summative assessment tool in that its use contributes to students' academic grades. No matter how teachers employ homework, their main area of consideration is how to use it effectively to maximize their students' learning.

The literature documents ample evidence of how homework affects students' performance (see, for example, Cooper, 1989; Cooper, Lindsay, Nye, & Greathouse, 1998; Cooper & Valentine, 2001; Keith & Cool, 1992; Keith, Keith, Troutman, Bickley, Trivette, & Singh, 1993; Mikk, 2006; Rodriguez, 2004). However, much of this body of research focuses on relationships between one specific dimension of homework and student performance, and rarely takes into account the diverse characteristics of homework. As a result, the research findings not only oversimplify the relationship between homework and student learning, but any recommendations arising out of these findings tend to be difficult for teachers to implement in their daily classroom practices.

The purpose of this study, therefore, was to explore the effects of four facets of homework practice on student learning. Our particular aims were to:

1. Identify critical dimensions of homework that would allow effective prediction of student performance; and
2. Provide recommendations to teachers about effective ways to use homework in classroom practices.

We conceptualized homework as a multi-dimensional construct so that we could examine the combined effects of these dimensions and their associated variables on student learning. We considered that a conceptual model of this kind would more accurately capture the relationship between homework practice and student performance, and so provide teachers with more useful guidance than would much of what is presently available on how to make effective assessment and instructional decisions related to homework.

Theoretical framework

In the instructional environments that teachers encounter on a daily basis, homework practice involves a series of important decisions that teachers have to make explicitly or implicitly. First, teachers need to consider how much time their students should spend on homework and how often they should assign homework. Second, teachers need to determine the focus of homework that they expect the students to work on. After students turn in their homework, teachers have to think about the strategies they will use to assess and give feedback on the homework in order to optimize student learning (Cooper, Robinson, & Patall, 2006). These four components of assigning and using homework represent the construct of interest in this paper, that is, well-rounded classroom practice relating to homework. Thus, our four dimensions were:

I. The frequency with which homework is assigned
II. The amount of homework that is assigned
III. The focus of the homework provided
IV. The strategies teachers use when assigning and assessing homework.

Figure 1 presents a diagrammatic representation of this construct.

Method

Research design

We employed a cross-sectional, correlational research design to examine the relationship between classroom homework practices and student performance in mathematics. We used two-level (i.e., student and teacher) hierarchical linear modeling (HLM) with maximum likelihood estimate procedures to analyze the TIMSS database. We used a two-level approach because within TIMSS students were "nested" in classrooms which were taught by their teachers. Figure 2 presents the proposed HLM model that we intended to investigate.

Figure 1: Homework Practice as A Multi-dimensional Construct

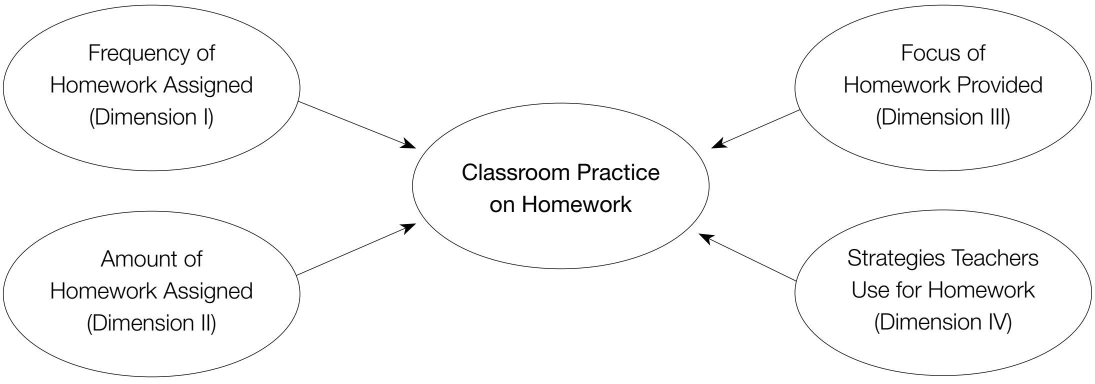
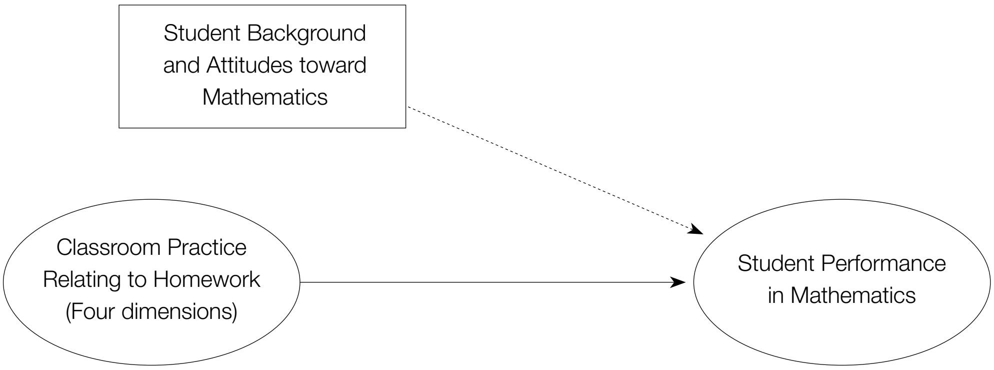

Frequency of Homework Assigned (Dimension I)
Focus of Homework Provided (Dimension III)
Classroom Practice on Homework
Amount of Homework Assigned (Dimension II)
Strategies Teachers Use for Homework (Dimension IV)

Figure 2: HLM Model Explored in this Study


Student Background and Attitudes toward Mathematics
Classroom Practice Relating to Homework (Four dimensions)
Student Performance in Mathematics

Sample

In this study, we chose to focus on mathematics as the subject area because of the national emphasis on mathematics education and testing after the No Child Left Behind Act, which was passed in 2000 with the aim of improving student achievement in the United States. The data used came from the United States portion of the information collected from approximately 46 countries during the administration of TIMSS in 2003. The participants of this study included the middle school mathematics teachers who participated in the TIMSS project and their Grade 8 students. The sample consisted of 8,921 students (48% boys and 52% girls) and 377 teachers (35% male and 65% female), stratified from 232 participating schools.

Instruments

The TIMSS 2003 mathematics achievement test for evaluating student performance consisted of 194 items in multiple-choice and constructed-response formats, which were then grouped into 12 booklets. Each student received one of the booklets (Martin, Mullis, & Chrostowski, 2004). In this study, a standardized national Rasch score derived from the raw score was used to capture individual student performance.[1]

The TIMSS student questionnaire took about 40 minutes to complete, and most questions required students to respond by using a Likert scale. The questionnaire asked students about their background, their attitudes toward subject matter, and aspects pertaining to their parents and their home and school activities. From these questions, we selected 10 items. Two of these dealt with demographic variables, namely gender and mother's level of education, and four related to each student's perception of his or her efficacy in mathematics (refer Bandura, 1986; Schunk, 1994). We decided to consider these variables because empirical research has shown that they effectively predict student performance on mathematics assessments (Brookhart, 1997; Brookhart & DeVoge, 1999; Rodriguez, 2004). In addition, we selected the item "time students spent on homework" because of its positive relationship with junior high school students' learning (Keith et al., 1993; Keith & Cool, 1992). Two items from the student questionnaire (namely, "frequency with which teachers assigned homework" and "amount of time teachers assigned to homework") were identical to items surveyed in the teacher questionnaire. We included these to serve as covariates to predict differences in student performance relative to teachers' practices. Table 1 provides a summary of these 10 student variables.

The teacher questionnaire took approximately 90 minutes to complete. Like the student questionnaire, it generally required respondents to use a Likert scale to answer the questions. The questionnaire asked teachers about their own background, professional preparation, and instructional practices. Because the primary goal of this present study was to explore the relationships between the posited four-dimensional construct of homework practice and student performance, we selected 10 items from the teacher questionnaire to represent the construct. These were "the frequency of homework assigned" (Dimension I), "the amount of homework assigned" (Dimension II), three items under "focus of the homework assigned"

[1] The national Rasch scores were standardized using the IRT model to have a mean score of 150 and a standard deviation of 10 within each country. Because each country has the same mean score and dispersion, these scores are not useful for international comparisons (Martin, 2005).

Table 1: Summary of Item and Variable Characteristics: Student Questionnaire

Variable name	Item #	Item description	Question statement	Used (recoded) codes
BSBGSEX	2	Gender	Are you a girl or a boy?	1= Girl 2= Boy
BSBGOLAN	3	English used at home	How often do you speak English at home?	1= Never 2= Sometimes 3= Almost always 4= Always
BSBGMFED	6A	Mother's level of education	What is the highest level of education completed by your mother (or stepmother or female guardian)?	1= Did not complete elementary school or did not go to school 2= Elementary school 3= Some high school 4= High school 5= Vocational/technical certificate after high school 6= Associate's degree in a vocational/ technical program 7= Two-year or four-year college or university degree 8= Master's degree, teaching certificate program, or professional degree/Doctoral degree
BSBMTWEL	8A	Do well in mathematics	What do you think about learning mathematics? ... I usually do well in math.	1= Disagree a lot 2= Disagree a little 3= Agree a little 4= Agree a lot
BSBMTCLM	8C	Mathematics is difficult	What do you think about learning mathematics? ... Mathematics is more difficult for me than for many of my classmates.	(same as above)
BSBMTSTR	8F	Mathematics is not my strength	What do you think about learning mathematics? … Mathematics is not one of my strengths.	(same as above)
BSBMTQKY	8G	Learn mathematics quickly	What do you think about learning mathematics? … I learn things quickly in mathematics.	(same as above)
BSBGDOHW	17I	Amount of time spent doing homework at home	On a normal day, how much time do you spend before or after school doing each of these things? I do homework …	1= No time 2= Less than a hour 3= 1–2 hours 4= More than 2 but fewer than 4 hours 5= 4 or more hours
BSBMHWMA	19A	Frequency with which teachers assign homework	How often does your teacher give you homework in mathematics?	1= Never 2= Less than once a week 3= 1 or 2 times a week 4= 3 or 4 times a week 5= Every day

Table 1 (contd.): Summary of Item and Variable Characteristics: Student Questionnaire

Variable name	Item #	Item description	Question statement	Used (recoded) codes
BSBMHWMG	19B	Amount of time teachers assign for homework	When your teacher gives you mathematics homework, about how many minutes are you usually given?	1= Fewer than 15 minutes 2= 15–30 minutes 3= 31–60 minutes 4= 61–90 minutes 5 = More than 90 minutes

(Dimension III), and five items under "strategies teachers use when assigning and assessing homework" (Dimension IV). Table 2 provides a summary of these 10 teacher variables, while Table 3 summarizes the four dimensions of the proposed "homework practice in class" construct and the dimensions' corresponding items.

Data analysis

The HLM models were fit to handle the nested data (Raudenbush & Bryk, 2002), based on the samples as described above. Three models were available for testing the HLM model hypothesized in this study.

- *Model I:* Here, we performed a one-way ANOVA with random effects. This served as a base from which to predict variation between and by teachers on student performance and to estimate the possible effects of the later-established HLM models. The base models for the student level and the teacher level were:

  $\mathit{Student\ Performance}_{ij} = \beta_{0j} + r_{ij}$

  $\beta_{0j} = \gamma_{00} + u_{0j}$

- *Model II:* We started with a HLM model that included the 10 student variables and the 10 teacher variables (as described above) to explore the relationship between homework practice and student performance. At student level, the 10 student items from the student questionnaire served as controlling variables to capture the explained variances between teachers and to eliminate the teacher main effects resulting from the covariate effects. At teacher level, the 10 teacher items from the teacher questionnaire ($|r|$ = .005 to .390, without collinearity problems in our preliminary analysis) acted as predictors to capture the variances which could explain the differences in student performance relative to teachers' practices. The models established at the student level and the teacher level were:

  $\mathit{Student\ Performance}_{ij} = \beta_{0j} + \beta_{1j}\mathit{(Gender)}_{ij} + \beta_{2j}\mathit{(English.Used.at.Home)}_{ij} + \beta_{3j}\mathit{(Mothers'.Level.in.Education)}_{ij} + \beta_{4j}\mathit{(Do.Well.in.Math)}_{ij} + \beta_{5j}\mathit{(Math.Is.Difficult)}_{ij} + \beta_{6j}\mathit{(Math.Is.Not.My.Strength)}_{ij} + \beta_{7j}\mathit{(Learn.Math.Quickly)}_{ij} + \beta_{8j}\mathit{(Time/to.Do.Homework.at.Home)}_{ij} + \beta_{9j}\mathit{(Frequency.Teachers.Assign.Homework)}_{ij} + \beta_{10j}\mathit{(Time.Teachers.Assign.Homework)} + r_{ij}$

  $\beta_{0j} = \gamma_{00} + \gamma_{01}\ \mathit{(Frequency.of.Homework.Assigned)}_{j} + \gamma_{02}\mathit{(Time.of.Homework.Assigned)}_{j} + \gamma_{03}\mathit{(Do.Problem/Question.Sets)}_{j} + \gamma_{04}\mathit{(Gather.Data.and.Report)}_{j} + \gamma_{05}\mathit{(Find.Relevent.Application)}_{j} = \gamma_{06}\mathit{(Monitor.Homework)}_{j} + \gamma_{07}\mathit{(Correct.and.Give.Feedback)}_{j} + \gamma_{08}\mathit{(Have.Student.Correct.in.Class)}_{j} + \gamma_{09}\mathit{(discuss.in.Class)}_{j} + \gamma_{010}\mathit{(Contribte.to.Grade)}_{j} + u_{0j}$

  $\beta_{1j} = \gamma_{10}$

  $\beta_{2j} = \gamma_{20}$

  $\beta_{3j} = \gamma_{30}$

  $\beta_{4j} = \gamma_{40}$

  $\beta_{5j} = \gamma_{50}$

  $\beta_{6j} = \gamma_{60}$

  $\beta_{7j} = \gamma_{70}$

  $\beta_{8j} = \gamma_{80}$

  $\beta_{9j} = \gamma_{90}$

  $\beta_{10j} = \gamma_{100}$

- *Model III:* With reference to the results of our analysis of Model II, we next developed a simplified HLM model using fewer predictors. Thus, we analyzed only those variables that were statistically significant in Model II. We kept seven variables—four at student level and three at teacher level. The four student variables related to students' attitudes and motivations in regard to mathematics learning, and we averaged these as an indicator of "perceived self-efficacy." The three significant teacher-level variables included "frequency of homework assigned" (Dimension I), "amount of homework assigned" (Dimension II), and "having students

Table 2: Summary of Item and Variable Characteristics: Teacher Questionnaire

Variable name	Item #	Item description	Question statement	Used (recoded) codes
BTBMHWMC	33	Frequency of homework assigned	How often do you usually assign mathematics homework to the TIMSS class?	1= Some lessons 2= About half the lessons 3= Every or almost every lesson
BTBMHWKM	34	Amount of homework assigned	When you assign mathematics homework to the TIMSS class, about how many minutes do you usually assign? (Consider the time it would take an average student in your class.)	1= Fewer than 15 minutes 2= 15–30 minutes 3= 31–60 minutes 4= 61–90 minutes 5= More than 90 minutes
BTBMKHCP	35A	Do problems/ question sets	How often do you assign doing problems/question sets to the TIMSS class?	1= Never or almost never 2= Sometimes 3= Always or almost always
BTBMKHCG	35B	Gather data and report	How often do you assign the task of gathering data and reporting to the TIMSS class?	(same as above)
BTBMKHCA	35C	Find relevant applications	How often do you assign the task of finding one or more applications of the content covered in the TIMSS class?	(same as above)
BTBMHDAM	36A	Monitor homework	How often do you monitor, via the mathematics homework assignments, whether or not the homework was completed?	(same as above)
BTBMHDAF	36B	Correct and give feedback	How often do you correct assignments and then give feedback to students via the mathematics homework assignments?	(same as above)
BTBMHDAC	36C	Have students correct in class	How often do you have students correct their own homework in class via the mathematics homework assignments?	(same as above)
BTBMHDAD	36D	Discuss in class	How often do you use homework as a basis for class discussion via the mathematics homework assignments?	(same as above)
BTBMHDAG	36E	Contribute to grade	How often do you use homework to contribute toward students' grades or marks via the mathematics homework assignments?	(same as above)

Table 3: Summary of the Proposed Four-Dimensional Homework Construct and the Dimensions' Corresponding Items in TIMSS 2003

Dimension	Description	Corresponding item
I	Frequency of homework assigned	• The frequency of homework assigned
II	Amount of homework assigned	• The time of homework assigned
III	Focus of homework provided	• Do problems/question sets • Gather data and report • Find relevant application for content covered
IV	Strategies teachers use in relation to homework	• Monitor whether homework is completed or not • Correct assignment and give feedback • Have students correct homework in class • Discuss homework in class • Use homework to contribute to grades

correct their homework in class" (Dimension IV). We combined and recoded the first two teacher variables to create a new variable that we named "homework intensity." The tested HLM models for the student level and the teacher level were:

$$\mathit{Student.Performance}_{ij} = \beta_{0j} + \beta_{2j}(\mathit{English.Used.at.Home})_{ij} + \beta_{3j}(\mathit{Mothers'.Level.of.Education})_{ij} + \beta_{8j}(\mathit{Do.Homework.at.Home})_{ij} + \beta_{9j}(\mathit{Perceived.Self\text{-}efficacy}) + t_{ij}$$

$$\beta_{0j} \_ \gamma_{00} + \gamma_{08}(\mathit{Have.Student.Correct.in.Class})_j + \gamma_{011}(\mathit{Homework.Intensity})_j + u_{0j}$$

$$\beta_{3j} = \gamma_{30}$$
$$\beta_{8j} = \gamma_{80}$$
$$\beta_{9j} = \gamma_{90}$$

Results and discussion

TIMSS 2003 involved a sampling procedure that required a carefully drawn random stratification of schools, classes, and students. Any analysis of TIMSS data therefore needs to take into account stratified sampling of subgroups (Martin, 2005), and this consideration should be kept in mind when considering the results presented here.

The means and standard deviations (based on students' and teachers' responses to the items in their respective questionnaires) for the variables selected for analysis in this study are presented in Table 4. Among the student variables, the students' mothers' average level of education was high school (M = 5.45). "Do well in mathematics" had the highest average score (M = 3.16) among the self-efficacy items. The amount of time spent on homework at home reported by the students was about an hour per day (M = 2.73), which was the same as the results found in a study conducted by Walberg (1991). Students also reported that the frequency with which teachers assigned homework was, on average, three or four times a week (M = 4.39), while the amount of time teachers required them to spend on mathematics homework was about 30 minutes per day (M = 2.28). These findings are consistent with the 15- to 30-minute durations reported either by students or by parents in a study conducted by Cooper et al. (1998).

For the teacher variables, teachers assigned homework (Dimension I) for over half of the lessons they taught (M = 2.85). The amount of time they reported assigning to homework (Dimension II) corresponded with the amount of time reported by the students (M = 2.23 vs. 2.28, respectively). In Dimension III (focus of homework provided) "provide problems/questions sets" was what the middle school mathematics teachers typically reported "always or almost always" doing (82%). However, they rarely required students to "gather data and report what they found" or to "find relevant applications" (2% and 12% of the teacher sample respectively reported they "always or almost always" set these tasks). Of the strategies teachers reported using in relation to homework (Dimension IV), teachers tended to give precedence to two of the five relevant items. These two were "monitoring if students had completed homework or not" (90% of the sample reported "always or almost always" doing this) and using homework to "contribute to grading"

Table 4: Summary of Means and Standard Deviations of Variables Involved in Student (Level 1) and Teacher (Level 2) Data Sets

Item Description	Mean	*SE*
Student variables (Level 2)		
Gender	1.48	.01
English used at home	3.76	.01
Mother's level of education	5.45	.04
Do well in mathematics	3.16	.02
Mathematics is difficult	2.15	.02
Mathematics is not my strength	2.41	.02
Learn mathematics quickly	2.83	.02
Amount of time spent doing homework	2.73	.02
Frequency with which teachers assign homework	4.39	.00
Amount of time teachers assign to homework	2.28	.00
Teacher variables (Level 2)		
Frequency of homework assigned	2.85	.03
Amount of homework assigned	2.23	.03
Do problems/question sets	2.82	.02
Gather data and report	1.59	.03
Find relevant applications	1.77	.04
Monitor homework	2.91	.02
Correct and give feedback	2.34	.05
Have student correct in class	2.49	.03
Discuss in class	2.48	.03
Contribute to grade	2.74	.03

(77% of the sample reported "always or almost always" doing this).

Table 5 shows the results of the analysis for the unconditional HLM model (Model I) of student performance in mathematics—the model specified without the explanatory (predictor) variables. The estimate of the grand mean for the teacher level was 150.19. The variance of teacher deviations from the grand mean was 50.369, which was significantly different from zero (χ^2_{382} = 6,518, $p < .001$). Student level accounted for about 48% (47.265 / (50.369+47.265)) of the variance, whereas teacher level accounted for 52% of the variances in student performance. According to Cohen's criteria (1988), and given that teacher level accounts for 52% of the variances, we would have been better to use multilevel regression rather than regular regression. What is particularly relevant here is that we cannot ignore the fact that teacher practices on homework generated so many of the variances.

In Model II, we analyzed a testable HLM model of student mathematics performance with the 10 controlling variables at student level and the 10 teacher variables related to homework practice at teacher level. The estimates of each of the coefficients (γ_s) and the variances of the random effects (τ_s) are presented in Table 6.

At student level, seven of the 10 variables, including "English used at home by students," "mother's level of education," four items representing student self-efficacy on mathematics, and "time to do homework at home," were statistically significant ($|t|$ = 2.343 to 9.567, $p < .02$). Specifically, the findings with "English used at home" ($t = 2.343$, $p < .02$) and "mother's level of education" ($t = 2.655$, $p < .01$) confirmed research conducted by Rodriguez (2004), who used the same HLM technique. However, the observation that gender ($t = 1.089$, $p > .05$) was a not significant predictor was contrary to the results from Rodriguez's research in which gender was a significant predictor of student

Table 5: Unconditional HLM Model of Student Performance without Predictors Involved (Model I)

Fixed effects		Coefficient	*SE*	*T* ratio	*p*
Intercept level 2, grand mean	$\gamma 00$	150.191	.377	398.383	.000
Random effects		*Variance Component*	*df*	$\chi 2$	*p*
Teacher mean residuals, *u0j*	$\tau 00$	50.369	382	6,518	.000
Student residuals, *r*^{i}*j*	σ^2	47.265			

Table 6: HLM Model II of Student Mathematics Performance with 10 Controlling Variables at Student Level and 10 Variables Related to Homework Practice at Teacher Level

Fixed effects		Coefficient	*SE*	*T* ratio	*p*
Model for teacher means, $\beta 0j$					
Intercept level 2, grand mean	$\gamma 00$	150.177	.352	427.169	.000
Frequency of homework assigned	$\gamma 01$	2.521	.903	2.789	.006
Amount of homework assigned	$\gamma 02$	2.874	.619	4.639	.000
Do problems/question sets	$\gamma 03$	1.177	.915	1.286	.199
Gather data and report	$\gamma 04$	-1.004	.710	-1.414	.158
Find relevant applications	$\gamma 05$	.639	.595	1.073	.284
Monitor homework	$\gamma 06$	-.544	1.105	-0.493	.622
Correct and give feedback	$\gamma 07$	-.981	.548	-1.790	.074
Have student correct in class	$\gamma 08$	2.009	.615	3.267	.002
Discuss in class	$\gamma 09$	.243	.699	.348	.728
Contribute to grade	$\gamma 010$	-.941	.766	-1.228	.221
Model for slopes					
Gender, $\beta 1j$	$\gamma 10$	.201	.185	1.089	.277
English used at home, $\beta 2j$	$\gamma 20$	.389	.166	2.343	.019
Mother's level of education, $\beta 3j$	$\gamma 30$	.137	.052	2.655	.008
Do well in mathematics, $\beta 4j$	$\gamma 40$	.969	.160	6.048	.000
Mathematics is difficult, $\beta 5j$	$\gamma 50$	-1.063	.111	-9.567	.000
Mathematics is not my strength, $\beta 6j$	$\gamma 70$	-.360	.111	-3.244	.002
Learn mathematics quickly, $\beta 7j$	$\gamma 80$	.883	.140	6.291	.000
Amount of time to do homework at home, $\beta 8j$	$\gamma 90$	-.530	.109	-4.872	.000
Frequency with which teacher assigns homework, $\beta 9$	$j\ \gamma 90$	.064	.151	.423	.672
Amount of time teacher assigns to homework, $\beta 10j$	$\gamma 100$	-.173	.112	-1.549	.121
Random effects		*Variance component*	*df*	$\chi 2$	*p*
Teacher mean residuals, *u0j*	$\tau 00$	43.680	372	6,374	.000
Student residuals, *rij*	σ^2	40.858			

performance on mathematics for middle school students. All four self-efficacy items were statistically significant ($|t|$ = 3.244 to 9.567, p < .01), findings consistent with the established theory that self-efficacy is relevant to student performance (Brookhart, 1997; Brookhart & DeVoge, 1999; Rodriguez, 2004).

"Time spent doing homework at home," as reported by the students, was a significant predictor, but surprisingly had a negative relationship with performance (t = -4.872, p < .001). It suggested that the more time students spent on homework at home, the lower their scores tended to be on the TIMSS achievement test. This outcome is partly consistent with a study conducted by Cooper and Valentine (2001). They observed a negative correlation for elementary school students but a somewhat positive correlation for students in middle and high schools. The negative relationship leaves us to speculate that too much homework that requires students to work alone without sufficient intellectual support from teachers is not beneficial for student learning (Cooper, 1989).

In addition, both the frequency with which teachers assigned homework and the amount of time they expected students to spend on it (as reported by the students) were not significantly correlated with students' mathematics achievement, a result that we discuss below in relation to our analysis of the corresponding items at the teacher level.

Overall, at student level, these seven variables within HLM Model II explained about 13% of the variances ((47.265-40.858) / 47.265) (cf. the variance in the unconditional HLM model, that is, Model I).

At teacher level, three of the 10 chosen variables were statistically significant ($|t|$ = 2.789 to 4.639, p < .01). These variables were "frequency of homework assigned," "amount of homework assigned," and "have student correct homework in class." The three variables represented, respectively, Dimensions I, II, and IV of the homework practice construct that we developed for this study. It is noteworthy that the finding relating to the amount of homework assigned (t = 4.639, p < .001) is consistent with one of the results from Brookhart's (1995) path analysis of the effects of classroom assessment on student achievement and motivation. Brookhart found that the amount of homework assigned had a positive effect on the mathematics achievement of Grade 8 students. However, the finding of the present study conflicts with empirical evidence that shows weak relationships between the amount of homework assigned and student performance (Cooper et al., 1998).

Another striking finding was that none of the items representing "the focus of homework provided" was statistically significant ($|t|$ = 1.073 to 1.414, p > .05). One possible explanation might be that the prompts related to and quality of the homework assigned, and how the teachers were using homework, may have been more influential factors than the focus of homework per se. We need to conduct further research to examine the possible relationships among these components.

We concluded that these findings partially supported our framework conceptualizing homework practice as a multi-dimensional construct. We also noticed that students' reports and teachers' reports of "frequency of homework assigned" and "amount of homework assigned" differed significantly. Possible reasons might be that the frequency with which individual teachers assigned homework and the amount of time they assigned to homework were estimated averages for their entire class, whereas the students' estimates may have been influenced not only by the class average but also by taking their own individual variations around the average into account. Because of the variances generated among students of individual teachers, the parameter estimated between student and teacher levels might be different.

Overall, at the teacher level, the variances of teacher residuals remained significant ($\chi^2{}_{372}$ = 6,374, p < .001), although 13% of the variances ((50.369-43.680) / 50.369) was explained by these four significant variables within HLM Model II. This outcome implies that other unknown factors were dominating the differences in student performance. The literature suggests that these could be (amongst other factors) the mathematics topics teachers teach (Rodriguez, 2004), the amount of time parents spend helping their children with homework (Balli, 1998), the type and extent of written comments students receive on their homework (Austin, 1976), and the effects of cooperative team homework (Ma, 1996). None of these factors was within the main area of interest of the present study, however.[2]

[2] Those variables were also not captured by the TIMSS questionnaires.

As Table 7 shows, at teacher level, we coded the two variables relating to frequency and amount of homework assigned as a new variable—“homework intensity,” which had 12 (1 to 12) possible values. This new coding system not only combined and simplified the original coding system with more levels but also kept effective the categorization arising out of the teachers’ answers to the original two questions. We also took the averages of the four student-belief items into one variable to represent student self-efficacy in learning mathematics. We then examined, using HLM two-level equations, our third model (Model III), which contained the new variables and eliminated the non-significant variables. Estimates of the coefficients (γ_s) and the variances of the random effects (τ_s) are shown in Table 8.

At student level, all four variables remained significant ($|t|$ = 2.686 to 27.537, $p < .01$), with perceived self-efficacy offering the strongest predictor. At teacher level, the new combined “homework intensity” variable remained significant (γ_{11} = .987, t = 4.569, $p < .001$). The variance of deviation at both student and teacher levels remained almost the same compared with the previous model (II) (40.858 vs. 41.079 at the student level, and 43.680 vs. 45.946 at the teacher level, respectively). This result indicates that HLM Model III was a successful simplification of the 10-variable Model II. With Model III, we were able to assign fewer variables at the student and teacher levels.

As we stated in the theoretical framework section above, homework can be seen as a series of instructional

Table 7: Summary of Coding System Used to Combine the Variables “Frequency of Homework Assigned” and “Amount of Homework Assigned” as an Indicator of “‘Homework Intensity”

Frequency of homework assigned *(options below)*	Raw codes (A)	Amount of homework assigned *(options below)*	Raw (B)	Homework intensity (Recoded codes: A x B)
Every or almost every lesson	3	61–90 minutes	4	Range of values = 1 to 12
About half the lessons	2	31–60 minutes	3	
Some lessons	1	15–30 minutes	2	
		Fewer than 15 minutes	1	

Table 8: HLM Model III of Student Mathematics Performance with Combined Four Significant Controlling Variables at Student Level and the Transformed New Variable Related to Homework at Teacher Level

Fixed effects		Coefficient	*SE*	*T* Ratio	*p*
Model for teacher means, *β0j*					
Intercept level 2, grand mean	γ00	150.263	.357	420.65	.000
Have student correct in class	γ08	2.249	.587	3.832	.000
Homework intensity	γ011	.987	.216	4.569	.000
Model for slopes					
English used at home, *β2j*	γ20	.447	.163	2.739	.007
Mother’s level of education, *β3j*	γ30	.136	.051	2.686	.008
Amount of time doing homework at home, *β8j*	γ80	-.577	.103	-5.600	.000
Perceived self-efficacy, *β9j*	γ90	3.226	.117	27.537	.000
Random effects		*Variance component*	*df*	*χ2*	*p*
Teacher mean residuals, *u0j*	τ00	45.946	385	6,894	.000
Student residuals, *rij*	σ2	41.079			

and assessment actions that teachers have to consider and decide on each day. We suggested that if teachers want to use homework in an effective way, they need to consider their homework-related practices from a perspective involving four dimensions. However, our findings suggest that while these dimensions are helpful, a simpler model (in the case of this study, the third HLM model—Model III—with its "homework intensity" variable, that is, the combined dimensions of frequency and amount of homework) may be particularly useful in helping teachers re-think their homework practices in the suggested direction. It appears that a productive way for teachers to support student learning is to assign homework more frequently and with more work instead of paying attention to one factor at the exclusion of the other. Additionally, our findings suggest that having students correct their own homework in class enhances their learning and gives teachers new ways to think about their homework-related practices.

Conclusions

This project examined the relationship between a proposed four-dimensional construct of homework practice and student learning in mathematics. Our premise was that the construct aligned with the instructional/assessment environments teachers encounter on a daily basis. We used the United States portion of the data sets from TIMSS 2003 to examine the relationship between this construct and student performance. The information we drew on came from 8,912 Grade 8 students and 456 mathematics teachers, and our analysis involved hierarchical linear modeling techniques. We concluded that it is possible to use various variables within the four dimensions of the homework practice construct to predict student performance, but that the relationship between the variable "focus of assigned homework" and student performance is inconclusive.

More particularly, the results of our HLM analyses supported our notion that it is possible to treat homework as a group of factors that have a combined relationship with (effect on) student performance. In addition, we found that combining Dimension I of our construct, "frequency of homework assigned," and Dimension II, "amount of homework assigned," into a new variable, "homework intensity," was effective in predicting student performance. In other words, increasing the frequency and the amount of homework that teachers assign to students serves as a predictor of higher student achievement.

In regard to teachers' classroom practices relating to homework, this study suggested that the more frequently teachers assigned homework and the greater the amount of time they expected students to spend on homework, the more likely their students were to have the higher achievement scores on the TIMSS mathematics assessment. Furthermore, the more often that teachers spent time in class having students correct their own homework, the more likely it was that their students performed well on the TIMSS mathematics achievement test. However, we believe that further research is needed to explore why this strategy had such a positive impact on student learning whereas "the focus of homework assigned" yielded no correlation with student performance.

References

Austin, J. D. (1976). Do comments on mathematics homework affect student achievement? *School Science and Mathematics, 76*(2), 159–164.

Balli, S. J. (1998). When mom and dad help: Student reflections on parent involvement with homework. *Journal of Research and Development in Education, 31*(3), 142–146.

Bandura, A. (1986). *Social foundation of thoughts and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Brookhart, S. M. (1995). *Effects of the classroom assessment environment on achievement in mathematics and science.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied Measurement in Education, 10*(2), 161–180.

Brookhart, S. M., & DeVoge, J. G. (1999). Testing a theory about the role of classroom assessment in student motivation and achievement. *Applied Measurement in Education, 12*(4), 409–425.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associations.

Cooper, H. (1989). *Homework.* White Plains, NY: Longman.

Cooper, H., Lindsay, J. J., Nye, B., & Greathouse, S. (1998). Relationships among attitudes about homework, amount of homework assigned and completed, and student achievement. *Journal of Educational Psychology, 90*(1), 70–83.

Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987–2003. *Review of Educational Research, 76*(1), 1–62.

Cooper, H., & Valentine, J. C. (2001). Using research to answer practical questions about homework. *Educational Psychologist, 36*(3), 143–153.

Keith, T. Z., & Cool, V. A. (1992). Testing models of school learning: Effects of quality of instruction, motivation, academic coursework, and homework on academic achievement. *School Psychology Quarterly, 7*, 207–226.

Keith, T. Z., Keith, P. B., Troutman, G. C., Bickley, P. G., Trivette, P. S., & Singh, K. (1993). Does parental involvement affect eighth-grade student achievement? *School Psychology Review, 22*, 474–496.

Ma, X. (1996). The effects of cooperative homework on mathematics achievement on Chinese high school students. *Educational Studies in Mathematics, 31*(4), 379–387.

Martin, M. O. (Ed.). (2005). *TIMSS 2003 user guide for the international database.* Chestnut Hill, MA: Boston College.

Martin, M. O., Mullis, I. V. S., & Chrostowski, S. J. (Eds.). (2004). *TIMSS 2003 technical report.* Chestnut Hill, MA: Boston College.

Mikk, J. (2006). *Students' homework and TIMSS 2003 mathematics results.* Paper presented at the International Conference, "Teaching Mathematics: Retrospective and Perspectives," Tartu, Estonia.

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

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education, 17*(1), 1–24.

Schunk, D. H. (1994). Self-regulation of self-efficacy and attributions in academic settings. In D. H. Schunk & B. J. Zimmerman (Eds.), *Self-regulation of learning and performance: Issues and educational applications* (pp. 75–99). Hillsdale, NJ: Lawrence Erlbaum Associates.

Walberg, H. J. (1991). Does homework help? *School Community Journal, 1*(1), 13–15.

Gender differences in mathematics achievement among Iranian eighth-graders in two consecutive international studies (TIMSS 1999 and TIMSS 2003)

Ali Reza Kiamanesh
Tarbeiat Moallem University (Teacher Training University)
Tehran, Iran

Abstract

This study investigated the changes observed in the mathematics achievement across TIMSS 1999 and TIMSS 2003 of Iranian Grade 8 boys and girls. The study also looked at changes across the two studies in male and female teachers' characteristics and considered these as factors contributing to the changes in student achievement. Changes in the average percentage of correct scores for three mathematics content areas—number, algebra and data—indicated a descending trend for both genders. The superiority of boys over girls in four different content areas in the 1999 study disappeared in TIMSS 2003. Iranian girls in the middle school tended to score higher than boys on "data, algebra, number, and geometry" content areas. The superiority of Iranian boys over girls in multiple-choice items in TIMSS 1999 disappeared in TIMSS 2003. The gender gap in "open-ended items," which favored boys in TIMSS 1999, also changed in favor of girls in TIMSS 2003. In fact, for TIMSS trend items, girls did not perform better than boys on the open-ended items and boys did not perform better than girls on the multiple-choice items. Iranian female teachers are "younger and have less teaching experience" than males. Moreover, female teachers express higher tendencies toward "using different ways to solve most mathematics problems," including having students "work together in small groups," "relate mathematics learning to their daily lives," and "explain their answers." Furthermore, female teachers place more emphasis than male teachers on "mathematics homework," "using constructed-response items," and "giving mathematics tests and exams." However, they are less likely than the men to favor "learning mathematics through memorization" and "having students do independent work." A higher proportion of female teachers than male teachers thought that "teachers' job satisfaction within their school is high" and "society and their students appreciate their career." The findings show that job satisfaction and the positive perspective of female teachers regarding teaching of mathematics may be the factors behind the better mathematics performance of girls than boys at Grade 8.

Introduction

Given the applicability of mathematics in almost all aspects of every society and the extreme importance of mathematics learning, educationists and researchers have devoted much effort to investigating factors that explain variations in mathematics achievement between boys and girls in schools. Gender differences in scholastic achievement in general and findings on gender differences in mathematics achievement in particular are not newly emerged facts. A long history of research in this area demonstrates that male advantage in mathematics achievement is a universal phenomenon (Beaton, Mullis, Martin, Gonzalez, Kelly, & Smith, 1996; Mullis et al., 2000).

Researchers have shown that boys tend to score higher than girls on problems that include spatial representation, measurement, and proportions, as well as complex problems, whereas girls tend to score higher than boys on computations, simple problems, and graph reading (Beaton, Mullis, Martin, Gonzalez, Kelly, & Smith, 1999). According to some research findings, the "gender gap" in mathematics achievement increases during middle school and becomes more pronounced at the upper secondary level (Fennema, 1985; Fennema, Carpenter, Jacob, Frank, & Levi, 1998).

One of the reasons given for the "gender gap" in mathematics achievement relates to the type of strategies that early elementary school-age children use to solve mathematics problems (Davis & Carr, 2001). Researchers (Carr, 1996; Carr & Jessup, Jessup

& Fuller, Fennema et al., all cited in Davis & Carr, 2001) have found that boys are more likely to retrieve information from memory and use covert cognitive strategies, such as decomposition, whereas girls are more likely to use overt strategies, such as counting on fingers or manipulative strategies, to solve mathematics problems.

Another area related to the "gender gap" in mathematics achievement is item format. Researchers have shown that boys perform better than girls on multiple-choice items and that girls perform relatively better than boys on open-ended items (Bell & Hay, 1987; Bolger & Kellaghan, 1990). One possible explanation for the above-mentioned difference between the two genders might be the fact that open-ended items require language skills, and multiple-choice items do not entail verbal ability (Murphy, 1982). Another possible explanation could be the fact that male students tend to guess answers more than girls do. Work by Wester and Henriksson (2000) shows no significant changes in gender differences when the item format is altered. Females perform slightly better than males when using multiple-choice items; this difference in favor of females remains the same when open-ended items are used.

In addition, educators and researchers have debated for many years about the school variables that influence students' achievement. An early study by Coleman et al. (1966) suggested that, independent of a child's background and general social context, schools bring little influence to bear upon a child's achievement. Recent studies on teachers' effects at classroom level have shown that differential teacher effectiveness is a strong determinant of differences in students' learning. According to Darling-Hammond (2000), teacher quality variables appear to be more strongly related to students' achievement than do class sizes, overall spending levels, and teacher salaries. From among variables assessing teacher "quality," the percentage of teachers with full certification and majoring in the field is a more powerful predictor of students' achievement than is teacher's level of educational attainment (e.g., Master's degree).

Darling-Hammond and Hudson (1998) also suggest that teacher, school, and student quality influences the teaching quality, which in turn affects students' outcomes. Teachers' effects are dominant factors affecting students' achievement gain (Wright, Horn, & Sanders, 1997). Other studies suggest that factors like class size (Broozer, 2001) and teacher quality variables (Darling-Hammond, 2000) may play an important role in what students learn.

Some studies in developing countries have shown mixed results related to the role of the teacher and school effects in students' learning (Fuller & Clarke, 1994; Kremer, 1995). Opdenakker and Van Damme's study (2006) on the relationship between teacher characteristics and teaching styles showed a small positive correlation between job satisfaction and classroom management for the second grade mathematics teachers. The researchers indicated that job satisfaction and teaching style dimensions were not gender-related. More specifically, male and female mathematics teachers differed with respect to classroom management skills. However, a study by Peire and Baker (1997) indicated that levels of job satisfaction are higher among young and less experienced teachers than among older and experienced teachers. Workplace conditions, including parental support and involvement in school activities, are among factors that have positive correlations with teachers' job satisfaction. Satisfaction with teaching as a career is thus a precondition for teacher effectiveness, which in turn leads to effective teaching and students' achievement. Teachers who do not receive support in their work may be less motivated to teach and perform well in the classroom (Ostroff, 1992). Teachers who are satisfied with their teaching career are less likely to leave the career than are those who are less satisfied.

Research findings related to the effect of class size on students' performance vary. For instance, Woessmann and West (2006) investigated the class-size effects in different countries using data from IEA's TIMSS studies. They found that in Greece and Iceland, class size has an effect on students' performance. However, they ruled out any noteworthy class-size effect in Canada, Portugal, Singapore, and Slovenia. Woessmann and West concluded that class-size effects estimated in one school system could not be interpreted as a general finding for all school systems.

One of the most remarkable results of TIMSS 2003 for Iranian eighth-graders was the significant decrease in the boys' mathematics achievement score from the time of TIMSS 1999 and the significant improvement in the girls' achievement over the same period. The mean mathematics achievement scores for the Iranian eighth-graders across the three TIMSS studies (1995, 1999, and 2003) showed a significant decline from 1999 to 2003 and no significant change from 1995 to 1999. On the one hand, the Iranian girls' mean score in 2003 was higher than their mean scores for 1999 and 1995 (by 9 and 12 scale score points, respectively). On the other hand, the Iranian boys' mean score in TIMSS 2003 was significantly lower than their mean scores for 1999 and 1995 (by 24 and 21 scale score points, respectively). As shown in Table 1, the superior achievement of Iranian boys in TIMSS 1995 and 1999 had moved to a reverse trend by TIMSS 2003.

Purposes of the study

Mathematics achievement involves a complex interaction of factors that have specific direct effects and/or indirect effects through other factors on school outcome. The study presented in this paper investigated the differences between boy and girl students' mathematics achievement in regard to variables such as mathematics content areas, performance expectation levels of the test items, test formats, and teachers' characteristics. In general, the present study was designed to investigate the changes observed in Iranian boy and girl eighth-graders' mathematics achievement across TIMSS 1999 and 2003 and to identify the role of teachers' gender in the observed changes. More specifically, the study sought to investigate:

1. Gender differences observed in the "entire mathematics achievement test" as well as the "trend items" regarding "content areas" and "performance expectation levels."
2. Gender differences observed in the trend items regarding "item format."
3. Gender differences observed in "mathematics teachers' characteristics" in order to explain the observed differences between the mathematics achievement of boy students and the mathematics achievement of girl students.

Method

Data sources

The data for this study were obtained from 5,301 and 5,122 Iranian eighth-graders as well as from 170 and 180 teachers (one class per school) who participated in TIMSS 1999 and TIMSS 2003, respectively. The average age of the sampled students at the time of testing was 14.4 and 14.6, respectively. More specifically, the data related to the students who took part in the mathematics achievement tests of the two TIMSS studies as well as to the teachers who completed the entire number of required items on the teacher background questionnaires. In both studies, teachers did not constitute representative samples of Iranian teachers, but were the teachers for nationally representative samples of students. Therefore, in this study, I analyzed the teacher data at the teacher level through student–teacher linkage files, using the IEA's IDB analyzer in the TIMSS 1999 user guide (Gonzalez & Miles, 2001) and the TIMSS 2003 user guide (Martin, 2005). Table 2 shows the distribution of the sampled students as well as of the teachers who participated in TIMSS 1999 and TIMSS 2003, by gender.

Variables

All TIMSS 1999 and 2003 mathematics items as well as 79 trend items were used to compare the performance of both genders in relation to the following sets of variables:

1. *Mathematics achievement:* Average percentage correct and the international mathematics student achievement plausible values;
2. *Item content areas:* including "number," "algebra," "measurement," "geometry," and "data;"
3. *Item performance expectation levels:* "applying concepts," "knowing facts and procedures," "solving routing procedures," and "reasoning;"
4. *Item format:* "multiple-choice" or "open-ended;"

Table 1: Average Scores for Iranian Eighth-graders on Mathematics Test Items in TIMSS 2003, 1999, 1995

Study	Both genders	Girls	Boys
TIMSS 2003	411	417	408
TIMSS 1999	422	408	432
TIMSS 1995	418	405	429

Table 2: Distribution of Iranian Students and Teachers Participating in TIMSS 1999 and 2003, by Gender

Study		Total	Female		Male	
		N	*N*	Percent	*N*	Percent
Students	TIMSS 1999	5,301	2,096	39.5	3,205	60.5
	TIMSS 2003	5,122	2,140	41.8	2,982	58.2
Teachers	TIMSS 1999	170	52	30.6	118	69.4
	TIMSS 2003	180	69	38.1	111	61.3

5. *Teacher characteristics:* almost all items from the teacher background questionnaires for both studies were used to compare the characteristics of both genders in regard to variables such as age, teaching experience, teachers' expectations of students, teachers' job satisfaction, and the amount of time spent on homework, tests, or quizzes, as well as teaching and grouping strategies.

Data analyses

The present study utilized the independent samples *t*-test to investigate student gender differences across the two studies in relation to the content areas, performance expectation levels, and format of the mathematics items. Teacher gender differences in the two studies regarding variables such as the degree of satisfaction teachers reported in relation to their jobs, their respective ages and years of teaching experience, their use of small-group teaching, and the extent to which they agreed with teaching mathematics through problem solving and memorization were investigated through *t*-tests and chi square tests.

Results

The findings relating to Iranian students who reported in TIMSS 1999 and in TIMSS 2003 "always or almost always" using Farsi at home (the language of the test) are particularly worth noting in terms of the gender difference across the two studies. In both studies, more girls than boys reported "always or almost always" using Farsi at home. In 1999, the proportion of girls reporting this incidence was 2.2% higher than the proportion of boys, whereas by 2003, the difference in proportion had risen to 13.3%. In addition, of the students who reported they "sometimes or never" used Farsi at home, girls were 2.7% and 13.9% less likely than boys to give this response in the 1999 and 2003 studies, respectively.

The comparison of performance in the five different mathematics content areas in TIMSS 1999 and 2003 revealed a decline in the average scale score for both genders in the content areas of "number, geometry, and data." The extent of the decline in these areas was greater for the girls than for the boys. In the other two content areas, that is, measurement and algebra, the girls' average scale score increased and that of the boys declined. In TIMSS 1999, the average scale score for boys in all five content areas was higher than that for the girls. However, in TIMSS 2003, the average scale score for girls in four out of the five content areas was higher than that for the boys, while the boys' average scale score in the "measurement" content area was higher than that for the girls. In general, though, the differences between the girls and the boys were not significant in the two studies for the five content areas. Table 3 shows the average percentages of students who correctly answered items relating to the different content areas. It also shows the *t*-test results by study and gender.

The trend data showed that the average percentage of items in all four performance expectation levels correctly answered by boys decreased from TIMSS 1999 (3.71) to 2003 (5.47). However, the average percentage of items reflecting the "reasoning" category correctly answered by girls increased by 2.81% across the two studies. These indices remained almost the same for "knowing facts and procedures" as well as for "solving routine problems," but "applying concepts" showed a decrease of 2.64%. In general, the differences between girls and boys were not significant in the two studies for the four performance expectation levels. Table 4 shows the average percentages of items relating to the different performance expectation levels that were correctly answered and the *t*-test results, by study and gender.

Table 3: Average Percentage of Items within Different Mathematics Test Content Areas Correctly Answered by Students, by Study and Gender

Content areas	*N* of items	Study	Gender	Mean	Standard error	*t*-value	*p*-value
Numbers	24	1999	Female	39.22	3.77	-.917	.364
			Male	42.67			
		2003	Female	38.13	3.66	-.038	.97
			Male	38.27			
Algebra	16	1999	Female	32.57	6.99	-.109	.914
			Male	33.34			
		2003	Female	33.67	6.51	.91	.37
			Male	27.74			
Measurement	17	1999	Female	18.55	4.12	-1.43	.162
			Male	24.44			
		2003	Female	19.12	4.67	-.556	.582
			Male	21.72			
Geometry	12	1999	Female	35.38	5.49	-.98	.337
			Male	40.78			
		2003	Female	35.31	5.16	-.21	.907
			Male	35.93			
Data	10	1999	Female	46.42	7.93	-.579	.57
			Male	51.01			
		2003	Female	45.97	6.98	-.066	.95
			Male	46.43			

With regard to the 79 trend items, the average percentage of items correctly answered by Iranian students in TIMSS 1999 was almost 3% higher than the average percentage of items correctly answered in TIMSS 2003. In TIMSS 1999, the average percentage correct for the boys was 4% higher than the average percentage correct for the girls (37.63 and 33.75, respectively). In TIMSS 2003, the superiority of the boys disappeared, and the girls' average percentage correct was 0.45 of a percent higher than the average percentage correct for the boys (33.7 and 33.25, respectively).

In TIMSS 1999, the average percentage of the open-ended format trend items (there were 20 of these) correctly answered by the boys was slightly higher than the average percentage of such items correctly answered by the girls (23.1 and 21.33, respectively). In TIMSS 2003, the average percentage of these items correctly answered by the girls was higher than that for the boys (18.8 and 17.4, respectively). Both observed differences were not significant.

In TIMSS 1999, more boys than girls correctly answered the multiple-choice format trend items. (There were 59 of these items.) The average percentages of items correctly answered by the boys and by the girls were 42.52 and 37.96, respectively. In TIMSS 2003, the average percentage correct for both genders was almost the same (38.73 for the girls and 38.68 for the boys). Both observed differences were not significant. Table 5 shows the students' average percentage correct for the different item types. It also shows the *t*-test results, by study and gender.

Because the content of the teacher questionnaire for TIMSS 1999 and for TIMSS 2003 generally

Table 4: Average Percentage of Items Relating to Different Performance Expectation Levels Correctly Answered by Students, by Study and Gender

Performance Expectations	*N* of items	Study	Gender	Mean	Standard error	*t*-value	*p*-value
Applying Concepts	15	1999	Female	46.93	6.42	-.482	.634
			Male	50.04			
		2003	Female	44.29	5.39	-.359	.722
			Male	46.27			
Knowing Facts and Procedures	22	1999	Female	32.49	4.32	-.902	.372
			Male	36.39			
		2003	Female	32.84	4.19	.462	.646
			Male	30.91			
Solving Routine Problems	27	1999	Female	29.99	4.42	-1.042	.352
			Male	34.6			
		2003	Female	29.46	4.22	-.023	.816
			Male	30.45			
Reasoning	15	1999	Female	29.19	6.38	-.516	.61
			Male	32.48			
		2003	Female	31.99	6.93	.465	.654
			Male	28.77			

differed, I was unable to conduct a comprehensive trend analysis using the two data sets. We also need to note that in Iran's education system, male teachers teach at boys' middle schools and female teachers teach at girls' middle schools. As a result, comparing the achievement of boy and girl students is the same as comparing the performance of male and female teachers.

The findings of the present study showed that female teachers teaching in the girls' schools were younger than the male teachers teaching in the boys' schools. The association between the teachers' gender and their age was significant in both TIMSS 1999 and 2003 (χ^2 = 171.36 , df = 4 ; χ^2 = 447.95, df = 4, respectively). In both studies, most of the female teachers were under the age of 40, and most of the male teachers were above the age of 40. In the 1999 study, the female teachers teaching in the girls' schools had less teaching experience than the male teachers in the boys' schools (12.33 and 14.29 years, respectively). In the 2003 study, the average teaching experiences for female and male teachers were 11.65 and 15.37 years, respectively. The observed differences between the two groups in both studies were significant at the .01 level.

In TIMSS 1999, the average percentage of mathematics lesson time spent by students in a typical month on tests and/or quizzes in the female classes was lower than the average percentage of this time spent in the male classes (19.04 and 20.24, respectively). The observed difference between the two groups was significant at the .01 level. In TIMSS 2003, the percentage of lesson time that students spent in a typical week on tests and quizzes compared to the 1999 study decreased sharply. The percentage of time in female classes was 12.13 and in male classes 10.97. The observed difference between the two groups was significant at the .01 level.

In the 1999 study, the average class size for female classes was significantly lower than the average class size for male classes (31.91 and 33.32 students, respectively). The average class size for both groups in the 2003 study was almost the same (29.84 and 29.77 students, respectively). In the 1999 study, the average percentage of time that girl students in a typical week of mathematics lessons spent on "working problems with their teacher as guide" was significantly higher than the average percentage of corresponding time spent by boy students (20.1 and 18.05, respectively). In the 2003 study, the average percentage of time that the girl students in a typical week of mathematics lessons spent "working problems with their teacher as guide" was significantly lower than the average percentage of time that the boys spent on this activity (10.94 and 15.5, respectively). The observed differences between the two groups in both studies were significant at the .01 level.

In the 1999 study, the percentage of time that girl students spent "working problems on their own without their teacher's guidance" was significantly lower than the amount of time spent by the boy students (11.94 and 14.14, respectively). In the 2003 study, the percentage of time that girl students spent on working problems on their own without their teacher's guidance was significantly higher than that of the boy students (14.16 and 13.52, respectively). The observed differences between the two groups in both studies were significant at the .01 level.

In the 1999 study, the percentage of time that girl students spent "listening to their teacher re-teach and clarify content/procedures" was almost the same as that of the boy students (19.52 and 19.96, respectively). In the 2003 study, the percentage of time that girl students spent "listening to their teacher re-teach and clarify content/procedures" was significantly higher than that of the boy students (14.91 and 14.14, respectively).

Table 6 shows the differences between the male and the female teachers in relation to the different variables.

Although, in the 1999 study, teaching was the first chosen career for both genders before they began their higher education, more male teachers than female teachers said they would move away from teaching as a career if they could ($\chi^2 = 713.97$, df = 1). The association between teachers' gender and their belief about whether "society appreciates their work" and "students appreciate their work" was significant. In all of the three mentioned variables, the standardized residuals (Haberman, cited in Hinkle, Wiersma, & Jurs, 1998) indicated that the category "female teachers" was the major contributor to the significance of χ^2. In other words, the female teachers were significantly more likely than the male teachers to state that they would not change their teaching career and that society as well as their students appreciated their work.

In the 1999 study, the association between teachers' gender and most of the other variables assessed in the teacher background questionnaire was significant. These variables (and here I name just some of them) included the following (see also Table 7):

- *Hours outside the formal school day spent per week on activities like preparing or grading student tests or examinations:* more than four hours in female classes and three to four hours in male classes;
- *Hours outside the formal school day spent per week on planning lessons:* under one hour in female classes and three to four hours in male classes;
- *The importance of "remembering formulas and procedures" for students:* rated as very important in female classes and as not important in male classes;
- *Being able to think creatively:* somewhat important in female classes and not important in male classes;
- *Understanding how mathematics is used in the real world:* somewhat important in female classes and not important in male classes;
- *Mathematics is primarily an abstract subject:* somewhat important in female classes and not important in male classes;
- *Some students have a natural talent for mathematics and others do not:* teachers of female students strongly disagreed and teachers of male students strongly agreed;
- *A liking for and an understanding of students is essential for teaching mathematics:* teachers of female students agreed and teachers of male students strongly disagreed;
- *The subject matter that teachers emphasized most in mathematics classes:* teachers of female students used a combination of algebra, geometry, number, etcetera, and teachers of male students used mainly numbers;
- *The main sources of written information for deciding which topics to teach:* teachers of female students used the teacher edition of the textbook, and the teachers of male students used other resource books and examination specifications;
- *How to present a topic:* teachers of female students used the teacher edition of the textbook and

Table 5: Average Percentage of Differently Formatted Items Correctly Answered by Students, by Study and Gender

Number of Items	*N* of items	Study	Gender	Mean	Standard error	*t*-value	*p*-value
Multiple-choice Items	59	1999	Female	37.96	4.84	-.382	.704
			Male	42.52			
		2003	Female	38.73	4.1	.32	.751
			Male	38.68			
Open-ended Items	20	1999	Female	21.33	2.82	-1.614	.109
			Male	23.18			
		2003	Female	18.87	2.51	.041	.967
			Male	17.4			

Table 6: Differences between Female and Male Teachers on Selected Variables, TIMSS 1999 and TIMSS 2003

Variable	Study	Gender	Mean	Standard error	*t*-value	*p*-value
Teachers' teaching (Number of students)	1999	Female	12.33	.78	-2.51	.01
		Male	14.29			
	2003	Female	11.65	.23	-3.72	.01
		Male	15.37			
Average class size Experience (years)	1999	Female	31.91	.192	-3.12	.01
		Male	33.32			
	2003	Female	29.84	.21	.33	.01
		Male	29.77			
Percentage of weekly total mathematics time students spent working problems with teacher as guide	1999	Female	20.1	.50	-3.81	.01
		Male	18.05			
	2003	Female	10.94	.26	-10.79	.01
		Male	15.5			
Percentage of weekly total mathematics time students spent working problems without teacher's guidance	1999	Female	11.94	.49	-4.49	.01
		Male	14.14			
	2003	Female	14.16	.24	2.66	.01
		Male	13.52			
Percentage of monthly mathematics lesson time spent on tests and/or quizzes	1999	Female	19.04	.5	-2.4	.01
		Male	20.24			
	2003	Female	12.13	.18	4.64	.01
		Male	10.97			

examination specifications, and teachers of male students used the student edition of the textbook and the national curriculum guidelines;

- *Asking students to explain the reasoning behind an idea:* teachers of female students did this in some lessons and teachers of male students did this in every lesson;
- *Asking students to represent and analyze relationships using tables, charts, and/or graphs:* teachers of female students did this in most lessons and teachers of male students did this in every lesson;
- *How often in mathematics classes students worked individually with assistance from the teacher:* teachers of female students did this in some lessons and teachers of male students worked this way in every lesson;
- *Low morale among fellow teachers:* teachers of female students reported "not at all" or "a little" and teachers of male students reported "quite a lot;"
- *Low morale among students:* teachers of female students said "not at all" or "a little" and teachers of male students said "a great deal;"
- *Assigning mathematics homework:* teachers of female students assigned mathematics homework three or

Table 7: Association between Gender of Teachers and Selected Variables from the Teacher Background Questionnaire, TIMSS 1999

Association between teacher's gender and …	χ^2	df	Contingency coefficient	*N*	*p*-value
1. Would change to another career if had the opportunity	713.96	–	.3479	5,211	.001**
2. Extent to which thinks society appreciates his or her work	296.04	1	.230	5,284	.001**
3. Extent to which thinks students appreciate his or her work	803.33	1	.364	5,246	.001**
4. How often usually assigns mathematics homework	638.17	3	.329	5,220	.001**
5. Agreement that students understand how mathematics is used in the real world	73.3	2	.118	5,226	.001**
6. Agreement that some students have natural talent for mathematics and others do not	199.34	3	.191	5,284	.001**
7. Belief that a liking for and an understanding of students is essential for teaching mathematics	44.89	3	.092	5,284	.001**
8. Extent to which he/she emphasizes mathematics homework	48.12	2	.096	5,156	.001**
9. Extent to which he/she emphasizes problem solving	119.18	2	.149	5,219	.001**
10. His/her degree of confidence to teach mathematics	26.41	2	.071	5,284	.001**
11. Which subject matter he/she emphasizes most in mathematics classes	138.16	4	.163	5,064	.001**
12. Main sources of written information he/she uses to decide which topics to teach	138.16	4	.163	5,064	.001**
13. Agreement that students need to be able to think creatively	142.55	2	.163	5,193	.001**
14. Extent to which gives feedback on homework to whole class	103.69	3	.153	4,514	.001**
15. Extent to which thinks parents' interest in their children's learning and progress limits how teacher teaches	42.03	3	.089	5,224	.001**

Note: ** Significant at $p < .001$.

four times a week or every day and teachers of male students did so approximately once a week;

- *Number of minutes assigned for mathematics homework:* teachers of female students assigned more than 90 minutes and teachers of male students 61 to 90 minutes;
- *Teachers' reported level of confidence in teaching mathematics:* teachers of female students reported a medium level and teachers of male students reported a high level;
- *Teachers' emphasis on mathematics reasoning and problem solving:* the level was medium for teachers of female students and low for teachers of male students;
- *Teachers' emphasis on mathematics homework:* teachers of female students reported a low level and teachers of male students reported a medium level;
- *Giving feedback on homework to the whole class:* teachers of female students said they did this "always" and teachers of male students said they "never" or "rarely did this;"
- *The extent to which parents' interest in their children's learning and progress limits how the teacher teaches:* for teachers in female classes, this was "a little" and for teachers in male classes, it was "quite a lot;"
- *The extent to which parents' lack of interest in their children's learning and progress limits how the teacher teaches:* teachers in female classes said "a little" and teachers in male classes said "not at all."

In TIMSS 2003, similar to the 1999 study, the association between teachers' gender and most of the other variables assessed in the teacher background questionnaire was significant. Among these variables were the following (see also Table 8):

- *Planning lessons outside school:* teachers in female classes reported usually two to three hours and teachers in male classes usually reported more than five hours;
- *Learning mathematics involves memorizing:* teachers of female students disagreed a lot and teachers of male students agreed;
- *There are different ways to solve most mathematics problems:* teachers of female students agreed a lot and teachers of male students disagreed a lot;
- *Teacher job satisfaction within school:* this was very high or high in the girls' schools and low or very low in the boys' schools;
- *Teachers' expectations for student achievement:* expectations were very high in the girls' schools and low or very low in the boys' schools;
- *Teachers' characterization of parental support for student achievement within school:* this was medium in the girls' schools and very low in the boys' schools;
- *Teachers' characterization of parental involvement in school activities:* this was reported as medium in the girls' schools and low in the boys' schools;
- *Teachers' characterization of students' desire to do well in school:* reported as very high in the girls' schools and very low in the boys' schools;
- *Frequency of working together in small groups:* teachers in the girls' schools said this occurred every or almost every lesson, while teachers in the boys' schools said it occurred in some lessons;
- *Relating what students are learning in mathematics to their daily lives:* this occurred during every or almost every lesson in girls' schools and for about half of the lessons in boys' schools;
- *Requiring students to explain their answers:* the teachers in the girls' schools said this requirement held for every or almost every lesson, while the teachers in the boys' schools said this requirement was in place for about half or some of the lessons;
- *Frequency of assigning mathematics homework to the TIMSS class:* homework was assigned during every or almost every lesson in girls' schools and during about half or some of the lessons in boys' schools;
- *Giving mathematics tests or examinations to the TIMSS class:* this occurred about once a week or every two weeks in girls' schools and a few times a year or never in boys' schools;
- *Item format usually used in mathematics tests or examinations:* the format used in girls' schools was mostly or only constructed-responses, and in boys' schools the items used about 50% of the time were constructed-responses;
- *Agreement that few new discoveries in mathematics are being made:* teachers in girls' schools agreed a lot and teachers in boys' schools disagreed a lot;
- *There are different ways to solve mathematical problems:* teachers in girls' schools agreed a lot and teachers in boys' schools disagreed, or disagreed a lot;
- *Teachers' perception of school as a safe environment:* teachers gave a medium rating in girls' schools and a high rating in boys' schools;

- *Teachers' emphasis on mathematics homework:* this was high in girls' schools and low in boys' schools;
- *Teachers' perception of quality of school climate:* this was high or medium in girls' schools and low in boys' schools.

Summary and conclusions

The change between TIMSS 1999 and TIMSS 2003 in the number of students whose native language was the language of the test was more evident for the Iranian girls who participated in the studies than for the Iranian boys who participated in these studies. This finding offers some explanation for the general increase in girls' achievement across the two tests and the general decrease in boys' achievement. Those students speaking the test language (Farsi) at home probably understood the test items better than those who did not.

Changes across the two studies in the average scale scores for the three content areas of the mathematics tests—number, algebra, and data—indicated a descending trend for both genders. The superiority of boys over girls in the four different content areas in the 1999 study disappeared in TIMSS 2003, and girls performed almost the same as or better than boys in 2003. The only superiority of boys that remained unchanged in both studies was in the "measurement" content area. However, the width of the gap between the two genders decreased. In contrast to the research findings of Beaton et al. (1999), Fennema (1985), and Fennema et al. (1998), Iranian girls in the middle school tended to score higher than Iranian boys in the middle school in the "data, algebra, number, and geometry" content areas from 1999 to 2003.

Approximately 26% of the total test scores in TIMSS 1999 and TIMSS 2003 were allocated to constructed-response items, which required students to generate and write their answers or provide explanations for

Table 8: Association between Gender of Teachers and Selected Variables from the Teacher Background Questionnaire, TIMSS 2003

Association between teacher's gender and ...	χ^2	df	Contingency coefficient	*N*	*p*-value
1. Hours per week spent on planning lessons	142.41	3	.169	4,839	.001**
2. Asking students to work together in small groups	250.57	3	.213	4,874	.001**
3. How he or she characterizes teachers' job satisfaction within school	334.04	4	.253	4,890	.001**
4. How often gives mathematics tests or examinations	115.61	4	.163	4,210	.001**
5. Uses formats typically used in mathematics tests or examinations	69.92	3	.148	3,188	.001**
6. Rating of school climate	84.65	2	.13	4,890	.001**
7. Extent to which emphasizes mathematics homework	122.51	2	.157	4,878	.001**
8. Perception of the safety of the school	32.32	2	.081	4,917	.001**
9. Thinks learning mathematics mainly involves memorizing	102.53	3	.143	4,917	.001**
10. Characterizes students' desire to do well within school	158.92	4	.177	4,917	.001**
11. Characterizes parental support for student achievement within school	177.67	4	.184	4,917	.001**
12. Characterizes parental involvement in school activities	95.81	4	.139	4,890	.001**
13. Frequency with which he or she assigns mathematics homework	103.01	2	.146	4,778	.001**
14. How often he or she gives a mathematics test or examination to the class	115.61	4	.163	4,219	.001**

Note: ** Significant at $p < .001$.

their responses. Findings from the present study show that the superiority of Iranian boys over girls in multiple-choice items in the earlier study disappeared in the later one. In fact, both genders had almost equal performances on this type of item in TIMSS 2003. The gender gap in the "open-ended items," which favored boys in TIMSS 1999, also changed to the superiority of girl students in TIMSS 2003. Thus, despite the findings of Bell and Hay (1987), Bolger and Jessup (cited in Davis & Carr, 2001), and Carr (1996), we can conclude that, for the TIMSS trend items, girls did not perform better than boys on the open-ended items and that boys did not perform better than girls on the multiple-choice items.

Data from the trend items for the performance expectation levels indicated that the average percentage of test items correctly answered by boys decreased in all four performance categories, and the average percentage of test items correctly answered by girls increased in two performance categories, namely "solving routine problems" and "reasoning." The TIMSS 1999 data showed that the average percentage of items correctly answered by boys in all of the four performance expectations was higher than the percentage correctly answered by the girls. However, in TIMSS 2003, the girls had superiority over boys in "knowing facts and procedures" and "reasoning."

One of the main objectives of the present study was to address this question: "Do male and female teachers vary in their dispositions and teaching characteristics; if so, what characteristics explain the observed mathematics achievement differences between boy and girl students at Grade 8?" Findings from the quantitative and categorical data indicated that those teachers who were teaching in girls' schools and those who were teaching in the boys' schools differed in regard to some of their characteristics. The teachers in the girls' schools were younger than the teachers in the boys' schools and had less experience in teaching. Trend data indicated that the number of years of teaching for teachers in the girls' schools decreased and for teachers in the boys' schools increased from 1999 to 2003. The percentage of the class time spent on teacher-guided student practice in the girls' schools decreased across the two studies more than it did in the boys' schools. However, the percentage of time students spent on independent study and working together in small groups increased from 1999 to 2003. Using group learning in education has many advantages, including enhancing achievement, social skills, and efforts (Johnson & Johnson, 2004). In addition, the direction of differences between the two groups in the amount of time spent on tests and quizzes reversed from 1999 to 2003. In the 2003 study, the teachers in the girls' schools spent more time than the teachers in the boys' schools on tests and quizzes.

In the 2003 study, more teachers in the girls' schools than in the boys' schools thought that teachers' job satisfaction within their school was high. Some of the findings from TIMSS 1999 also indicated a relatively high level of job satisfaction among female teachers. A sizeable number of female teachers in the 1999 study indicated that society and their students appreciated their work. Furthermore, in the 2003 study, more teachers in the girls' schools than in the boys' schools said teaching had been their first job priority at the beginning of their higher education. Also, more female teachers than male teachers said they would want to continue in the teaching profession if they had opportunity to change jobs.

The findings of this study do not support the outcomes of other studies carried out in Japan and Germany, which showed more men than women enjoying teaching as a profession (Lissmann & Gigerich, cited in Mwanwenda, 1997). However, the present study confirms findings indicating that women experience more job satisfaction than do men (Park, 1992). In addition, the findings of the present study confirm findings by Peire and Baker (1997) showing that younger and less experienced teachers report higher levels of job satisfaction than older and more experienced teachers report.

In the 2003 study, the girl students were being taught by teachers whose perspectives on teaching methodologies differed from those of the teachers of boy students. Teachers in the girls' schools were considerably more likely than the teachers in the boys' school to give high ratings to the following: teacher expectations for student achievement; students' desire to do well in school; frequency of assigning mathematics homework to class; emphasis on mathematics homework; parental support for students; and parental involvement in school activities. The last two features might offer other underlying reasons for the higher job satisfaction in the girls' schools than in the boys' schools. As Opdenakker and Van Damme (2006) argue, teachers with a high level of job satisfaction give more instructional support to their class than do those with a low level of job satisfaction.

Moreover, female teachers expressed higher tendencies toward using different ways to solve most mathematics problems. They also were more likely to have students work together in small groups, relate mathematics learning to their daily lives, and explain their answers. In addition, they put more emphasis on mathematics homework, using constructed-response items, and giving mathematics tests and examinations to the TIMSS class. In contrast, they showed a lower tendency than the male teachers toward having students learn mathematics through memorization and toward students engaging in independent work.

In general, the conclusions we can draw are that job satisfaction, parental support for student achievement, and involvement in school activities as well as the positive perspective of the female teachers regarding mathematics and the teaching of mathematics are the factors behind the effective teaching in girls' schools and the generally better performance of girl students than boy students on the TIMSS mathematics items. Two other non-equivalent plausible explanations can also be offered. First, the current results may be considered as the beginning of a local change in girls' mathematics performance, which aligns with the global trend in gender differences in mathematics achievement. In TIMSS 1999, boys performed significantly better than girls, but in TIMSS 2003, the average performance of boys and girls was almost equal (Mullis, Martin, Gonzalez, and Chrostwoski, 2004). The second explanation for the better performance of the girls across time is the prohibition on co-education in Iran.

References

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study* (TIMSS). Chestnut Hill, MA: Boston College.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1999). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study-Repeat* (TIMSS-R). Chestnut Hill, MA: Boston College.

Bell, R. C., & Hay, J. A. (1987). Differences and biases in English language formats. *British Journal of Educational Psychology, 57*, 200–212.

Bolger, N., & Kellaghan, T. (1990). Methods of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement, 31*, 275–293.

Broozer, M. (2001). Intra-school variation in class size: Patterns and implications. *Journal of Urban Economics, 50*, 163–189.

Carr, M. (1996). Metacognitive, motivational, and social influences on mathematics strategy use. In M. Carr (Ed.), *Motivation in mathematics* (pp. 189–111). Caskill, NJ: Hampton Press.

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. (1966). *Equality of educational opportunity.* Washington, DC: Department of Health, Education, and Welfare.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy analysis variables. *Educational Policy Analysis Archives, 8*(1). Available online at http://epaa.asu.edu/epaa/v8n1/

Darling-Hammond, L., & Hudson, L. (1998). Teachers and teaching. In R. J. Shavelson, L. M. McDonnell, & J. Oakes (Eds.), *Indicators for monitoring mathematics and science education.* Los Angeles, CA: Rand.

Davis, H., & Carr, M. (2001). Gender differences in mathematics: Strategy, use, and the influence of temperament. *Learning and Individual Differences, 13*, 83–95.

Fennema, E. (1985). Explaining sex-related differences in mathematics: Theoretical models. *Educational Studies in Mathematics, 16*, 303–312.

Fennema, E., Carpenter, E. T., Jacob, V. R., Frank, M. L., & Levi, L. W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher, 27*(5), 6–11.

Fuller, B., & Clarke, P. (1994). Raising school effects while ignoring culture? *Review of Educational Research, 64*, 119–157.

Gonzalez, J. E., & Miles, A. J. (Eds.). (2001). *TIMSS 1999: User guide for the international database.* Chestnut Hill, MA: Boston College.

Hinkle, D. E., Wiersma, W., & Jurs, G. S. (1998). *Applying statistics for the behavioral sciences.* Boston, MA: Houghton Mifflin Company.

Johnson, W. D., & Johnson, T. R. (2004). *Assessing students in groups: Promoting group responsibility and individual accountability.* Thousand Oaks, CA: Corwin Press/Sage Publications.

Kremer, M. (1995). Research on schooling: What we know and what we don't (a comment on Hanushek). *World Bank Research Observer, 10*, 247–254.

Martin, M. O. (Ed.). (2005). *TIMSS 2003 user guide for the international database.* Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Gregory, K. D., Garden, R. A., & Murphy, R. J. L. (2000). Sex differences in objective test performance. *British Journal of Educational Psychology, 52*, 213–219.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades.* Chestnut Hill, MA: Boston College.

Murphy, R. J. L. (1982). Sex differences in objective test performance. *British Journal of Psychology, 52*, 213–219.

Mwanwenda, T. S. (1997). Teacher gender differences in job satisfaction in Transkei. *Research in Education, 8*, 1–3.

Opdenakker, M. C., & Van Damme, J. (2006). Teacher characteristics and teaching styles as effective factors of classroom practice. *Teaching and Teacher Education, 22*(1), 1–21.

Ostroff, C. (1992). The relationship between satisfaction, attitudes and performance: An organizational level analysis. *Journal of Applied Psychology, 77*, 963–974.

Park, A. (1992). Women, men and the academic hierarchy: Exploring the relationship between rank and sex. *Oxford Review of Education, 18*, 227–239.

Peire, M., & Baker, D. P. (1997). *Job satisfaction among America's teachers: Effects of workplace conditions, background characteristics, and teacher compensation* (statistical analysis report). Washington, DC: United States Department of Education.

Wester, A., & Henriksson, W. (2000). The interaction between item format and gender differences in mathematics performance based on TIMSS data. *Studies in Educational Evaluation, 26*, 79–90.

Woessmann, L., & West, M. (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review, 50*(3), 696–739.

Wright, S. P., Horn, S. P., & Sanders, L. (1997). Teacher and classroom context effects on students' achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57–67.

Differences in teaching and learning mathematics in relation to students' mathematics achievement in TIMSS 2003

Radovan Antonijević
Faculty of Philosophy
University of Belgrade
Belgrade, Serbia

Abstract

This paper considers the main results of the TIMSS 2003 assessment that pertain to the mathematics achievement of students in Grade 8. It also considers the extent to which these results were affected by certain characteristics of teaching and learning mathematics cross-nationally. To assist the appropriate analysis of student achievement in mathematics, the paper also presents contextual pieces of information, obtained from the mathematics teachers and relating to mathematics curricula and to the teaching and learning of mathematics in primary school obtained from the mathematics teachers. More specifically, the research focuses on the following topics: the presence of specified types of teacher–student interactions in the process of knowledge and concepts' attainment; the presence of certain teaching and learning methods; different opportunities offered for solving complex problems in teaching; the assignment of mathematics homework and its role in teaching and learning mathematics; and the assignment of mathematics tests and examinations. The analyses include students' mathematics achievement in four of the countries that participated in TIMSS 2003: Bulgaria, the Netherlands, Serbia, and the United States. Different statistical procedures were applied in this study in order to analyze connections and dependences between chosen variables, including TIMSS mean scale scores, mean percentage, Spearman's correlation, statistical significance, and Fisher's coefficient, amongst others.

Background

The presence of certain opportunities for teaching and learning mathematics in primary school represents a group of crucial factors that has an important influence on students' achievement in mathematics. The scope and quality of students' attained knowledge and concepts and the abilities and skills they develop in this field depend on the overall quality of teaching and learning mathematics. It is not possible to form complete and clear demarcation lines between the two general contexts of students' achievement. The first refers to curriculum content and the other refers to existing opportunities for teaching and learning mathematics that appear in the teaching process. However, it is interesting to conceptualize a theoretical and empirical model that allows us to examine and describe how certain characteristics of the teaching and learning process can affect students' achievement in mathematics in the final grade of primary school.

As a starting point for establishing this kind of investigation model, my colleagues and I referred to a broad spectrum of published works that generated their research topics from different analyses of data from IEA's mathematics and science achievement studies (TIMSS). Many of the conducted analyses, represented in reviewed papers, involve exploration of the contextual dependences of students' achievement. Especially interesting are those that consider the quality of students' knowledge and attainment of concepts and their abilities and skills in relation to the teaching process. Each of the chosen variables in the research belongs to the classroom context for teaching and learning mathematics, which Turner and Meyer (2000) define (and discuss) as the "instructional-motivational context" of teaching and learning.

My initial idea was to confirm the presence and intensity of some characteristic influences within this context on students' achievement by drawing on and analyzing data from the sample for Serbian eighth-graders. One example of these characteristics in relation to this sample is the difference in mathematics achievement between those students who had access to computers and those who did not during mathematics teaching and learning. Another area of interest was to apply a cross-country comparison of these kinds of influences on students' achievement. An appraisal suggested it would be particularly interesting to compare these influences on students' achievement as they appear in highly developed countries with these same kinds of influences as they appear in countries undergoing social and economic transition.

Cross-country comparisons of the TIMSS data are often made in order to establish differences in factors affecting students' achievement. An example is Papanastasiou's (2002) analysis of information relating to Grade 4 students from Cyprus, Hong Kong SAR, and the United States who participated in TIMSS 1999. Other TIMSS cross-country studies relevant to establishing a conceptual framework for this present study include those by Birenbaum, Tatsuoka, and Yamada (2004), and Schümer (1999).

Despite the many advantages of conducting comparative analyses of international surveys, some shortcomings and problems in international comparative studies have been identified and discussed from the time that secondary analyses were first conducted of TIMSS data. Accordingly, in this paper, I also discuss specific types of information in the studies' questionnaires, and the consequences this information has in relation to of the reliability of the analyses and the validity of conclusions made. This "lack of information" appears to be underpinned, in part, by the general intention of the designers of the questionnaires to develop items that, in some cases, allowed researchers to obtain an extensive volume of information, but not necessarily in-depth, robust information. Adolfsson and Henricsson (1999), for instance, point to the need for greater standardization of research materials and procedures and for methods that allow participating countries to include in the questionnaire items that are particularly relevant to them. These two researchers also identify problems associated with translating the test material into the languages of each country. Some of the drawbacks of the conceptions and procedures associated with obtaining contextual data are also discussed in this study.

Data and purposes

The data used in this paper came from the main results of the TIMSS 2003 mathematics assessment for eighth-graders in four participating countries—Bulgaria, Serbia, the Netherlands, and the United States, and my focus here is on how much those results were affected by specified characteristics of the teaching and learning of mathematics. In this paper, I consider and discuss in particular factors that are known to contribute to the quality of teaching and from there influence the quality of student achievement. More specifically, I look at the following:

- Factors that determine the quality of teacher–student interactions that contribute to the process by which students gain knowledge and attain concepts associated with a field of learning;
- The extent to which certain teaching and learning methods and activities are employed in schools;
- The extent to which teachers provide students with different opportunities for solving complex problems in mathematics;
- The extent to which calculators and computers are used in mathematics teaching;
- The quality of mathematics homework that teachers assign to students; and
- The types of mathematics tests and examinations that teachers set for students.

The specific combined model of analysis of these separate variables that I employ in this paper includes both statistical and content analyses. For example, when considering the obtained statistical indices relating to how and when teachers employed certain methods and activities, I did not adhere only to a statistical analysis but endeavored to give a more complete and thorough (content) analysis by considering the broader context of influences underpinning teachers' use of these methods and activities. This process thus required examination of the degree of relationship between the teaching method and subject-matter characteristics, such as the structure of lessons, the time available to use a particular method, the amount of time given over to that method, and the frequency with which the method was used. The essential intention was not to examine each item separately, but to consider its natural contextual relationships and interconnectedness. Each content analysis in the paper is based on this model.

The main research question informing the analyses in this paper therefore was: *What are the levels of influence of some factors of teaching and learning mathematics in the classroom on students' achievement in the field of mathematics across a four-country sample of Grade 8 students?* Bos and Kuiper (1999) asked and considered a similar question in one of their secondary analyses of the TIMSS 1999 data: "What can be learned about mathematics achievement of grade 8 students, and the factors at student and classroom levels that may be associated with that achievement across 10 education systems?"

To explore the above question, I connected variables from the four countries' BTM files (data from the questionnaire for Grade 8 mathematics teachers) and BSA files (achievement data for Grade 8 students), and then merged these files using student–teacher linkage files (BST). The next step involved applying a variety of statistical procedures and measures so as to provide robust (statistical) descriptions and delineation of the characteristics of the chosen group of mathematics teachers and their students in each of the four countries, as well as of some cross-country characteristics.

The statistical procedures and measures used included the following: TIMSS mean scale score (*M*), standard deviation (*SD*), standard error (*SE*), percent measure, Spearman's correlation (rho), Fisher's coefficient (*F*), degrees of freedom (df), and statistical significance measure (*p*). All estimations of student achievement in mathematics were made by calculating the "first mathematics plausible value," available for each student in the student achievement files, as well as in the aforementioned merged files. DPC IDB analyzer software was then used to carry out each defined analysis in the research. This software has links with the SPSS software, and both offer an appropriate approach to analyzing the information held in the TIMSS international database.

Results and discussion

Along with the differences in the overall achievement results across the four countries, there were many cross-national differences relating to the separate factors influencing teaching and learning mathematics and consequently students' achievement. These are presented and discussed below.

Overall student achievement

Table 1 presents the overall achievement results in mathematics for the Grade 4 students in each of the four countries, as well as the number of students in each national sample. The total number of students across the four-country sample was 20,390 students. In the remaining tables in this paper, the numbers of students generally differ from those given in the Table 1 data. This is because the numbers of students belonging to the subgroups of teachers varied, depending on the number of teachers who gave answers to the teacher background questionnaire items that I considered.

Teacher–student interactions in mathematics teaching

The presence of different types of interactions between the teacher and his or her students in mathematics teaching, as well as between and among students, is important for overall teaching efficiency. There are many opportunities to improve these kinds of interactions. Among the factors that affect the organization and quality of teacher–student interactions in mathematics teaching are class size, the lesson topic and its contents, and the type of lesson taught.

One of the items in the teacher questionnaire asked teachers to give the number of students in their mathematics class. Table 2 presents the average class size for each of the four countries. Class size doubtless has an influence, albeit indirect, on the effectiveness of the teacher–student interactions that take place during classroom activities. More particularly, we can hypothesize that higher quality interactions are achieved in smaller rather than in bigger classes. As Keys (1999) emphasizes in his consideration of work conducted by Mortimore and Blatchford, there is a widespread belief amongst parents, teachers, and others that students learn more effectively in small classes. Another common supposition is that smaller class sizes are advantageous to maximizing the effective implementation of the intended curriculum.

Table 1: Students' Average Achievement in Mathematics, across Countries

Country	*M*	*N*	*SE*
Netherlands (NLD)	536.27	3,065	3.82
United States (USA)	504.37	8,912	3.31
Serbia (SCG)	476.64	4,296	2.60
Bulgaria (BGR)	476.17	4,117	4.32

Table 2: Average Number of Students in TIMSS Classes in Each Country

Country	*M*	*N*	*SE*
BGR	23.49	3,806	5.07
NLD	25.79	2,843	3.65
USA	24.54	7,504	6.35
SCG	26.53	3,899	4.94
TOTAL	24.95	18,052	5.54

However, the TIMSS 2003 results show evidence contrary to these notions. The average test scale scores of students were lowest in classes of up to 24 students (461 points, *SE* = 1.9), highest in classes of 25 to 32 students (473 points, *SE* = 1.4), and just slightly lower than the previous group in classes of 33 to 40 students (470 points, *SE* = 2.1) (Mullis, Martin, Gonzales, & Chrostowski, 2004, p. 266).

The TIMSS 2003 questionnaire for Grade 8 mathematics teachers contains only one variable in one item that directly relates to teacher–student interaction in mathematics teaching. One of the deficiencies in the structure of this questionnaire is the lack of variables in the domain of exploring diversity of teacher–student interaction, especially the interaction relevant to processes of cooperative learning and of students' attainment of concepts brought about the nature, type, and quality of the teaching.

Methods that teachers employ and activities they use when teaching mathematics

The different methods that teachers use when teaching mathematics reflect different didactic attitudes to the teaching and learning of mathematics and so, not surprisingly, can produce different achievement results for students, both quantitatively and qualitatively. In the mathematics teachers' questionnaire, Item 20 contains eight variables relevant to the methods and activities that teachers use in their teaching of mathematics. These are:

1. Reviewing students' homework;
2. Having students listen to lecture-style presentations;
3. Having students work on mathematics problems with the guidance of the teacher;
4. Having students work on mathematics problems of their own without teacher guidance;
5. Having students listen to the teacher re-teaching and clarifying content/procedures;
6. Requiring students to take tests or quizzes;
7. Having students participate in classroom-management tasks not related to the lesson's content/procedures; and
8. Having students engage in other activities.

Table 3 presents, for each of the four countries, the percentage (*M*) of teachers who reported using each of these activities. The table also lists the standard deviations (*SD*). It is interesting to see from the table which activities occupied the highest percentage of mathematics instructional time across the countries. The highest percentage in Bulgaria was for the activity "students working on problems with teacher's guidance" *(M* = 26.39, *SD* = 10.790). In the Netherlands, the activity attracting the highest percentage was "students working on problems of their own without teacher's guidance" (*M* = 27.27, *SD* = 20.401). In the United States, the corresponding activity was "students working problems with teacher's guidance" (*(M* = 21.07, *SD* = 10.806), and in Serbia it was "having students listen to lecture-style presentations" (*(M* = 25.59, *SD* = 14.607).

In the teacher questionnaire, Item 21 asks the question: "In teaching mathematics to the students in the TIMSS class, how often do you usually ask them to do the following?" The question is then followed by nine different activities (variables). The ones particularly relevant to this present analysis are: (21.c) "work on problems for which there is no immediately obvious method of solution;" (21.g) "relate what the students are learning in mathematics to their daily lives;" (21.h) "have students explain their answers;" and (21.e) "have students determine their own procedures for solving complex problems." Teachers could then respond to each variable using the following scale: "every or almost every lesson;" "about half the lessons;" "some lessons;" and "never."

Table 4 presents the results of Spearman's correlation coefficients (rho) carried out in relation to

Table 3: Mean Percentages of and Standard Deviations for Teachers who Reported Using the Eight Applied Activities, by Country

		Activities							
Country		1	2	3	4	5	6	7	8
BGR	*M*	9.58	17.79	26.39	16.96	16.49	7.91	3.09	1.81
	SD	6.367	13.257	10.790	9.598	11.873	5.970	4.013	2.831
NLD	*M*	14.31	13.38	20.34	27.27	7.47	8.59	5.03	3.66
	SD	9.744	7.567	17.545	20.401	4.993	4.808	5.532	3.805
USA	*M*	13.29	18.24	21.07	17.80	10.66	10.77	4.75	3.47
	SD	6.956	10.919	10.806	10.438	5.433	6.011	4.550	5.919
SCG	*M*	6.96	25.59	22.81	20.46	9.48	7.33	3.12	4.36
	SD	4.232	14.607	12.219	13.464	6479	5.520	3.482	5.203

the combined data from the four countries for these activities. The table shows that the highest level of the Spearman's correlation coefficient (0.427) was that for the activities "explain their answers" and "determine their own procedures for solving complex problems." The lowest level (0.056) was obtained for the activities "work on problems for which there is no immediately obvious method of solution" and "explain their answers."

A factorial analysis conducted in relation to the factors influencing the mathematics performance of Iranian students in TIMSS 1999 confirmed that variables included in the "teaching factor" had the following measures of correlation (Kiamanesh, 2004): "discussing practical problems while teaching a new math topic" (0.646); "using things from everyday life to solve mathematics problems" (0.645); "working together in pairs or small groups" (0.609); "discussing completed homework" (0.547); and "working on mathematics projects" (0.536). These results show not only high levels of positive connections between these teaching variables and the achievement in mathematics of students in the Iranian sample, but also their overall significance as predictors of student achievement.

Problem-solving in mathematics teaching

Many researchers consider that, as much as it is possible to do so, having students use problem-solving tasks in each mathematics domain is a very important part of successful mathematics teaching. Many studies confirm the important place that problem solving holds in regard to the overall quality of teaching in mathematics and for students' achievement in this subject. An assumption is that a group of students who have (a sufficient number of) problem-solving tasks as part of their mathematics learning will be more successful in mathematics than will a group of students who have no opportunity to solve these problems. In the mathematics teachers' questionnaire, Item 21 was designed to collect information on the presence of elements of problem solving in mathematics teaching. The teachers were asked this question: "In teaching mathematics to the students in the TIMSS class, how often do you usually ask them to do the following?" Two of the nine listed answers (variables) are problem-solving activities. The overall achievement of the students in the four-country sample shown in Table 5 confirms the previous hypothesis. The presence of problem-solving tasks in the mathematics curriculum and its implementation enables students to gain a deeper understanding of certain mathematical concepts and from there to gain the higher achievement scores in mathematics achievement tests.

The frequency with which teachers have their students do problem-solving tasks is also relevant. The TIMSS students of mathematics teachers who "often" used problem-solving activities gained higher achievement scores than the students of the teachers who said they "rarely" or "never" used problem-solving activities when teaching mathematics. Those students in the four-country sample whose teachers chose the option of "every or almost every lesson" (5.35% of the total student sample) achieved an average scale score of 517.54. Based on the results given in Table 5, it seems that when teachers enable students to solve different kinds of problem tasks, their students gain

Table 4: Relationships between the Applied Activities Used by Mathematics Teachers, across Countries

Activities		How often work on problems	How often relate to daily life	How often explain answers	How often decide on their own
How often work on problems	Correlation Sig. (2-tailed)	1.000 .	.121 .000	.056 .000	.258 .000
How often relate to daily life	Correlation Sig. (2-tailed)	.121 .000	1.000 .	.279 .000	.346 .000
How often explain answers	Correlation Sig. (2-tailed)	.056 .000	.279 .000	1.000 .	.427 .000
How often decide on their own	Correlation Sig. (2-tailed)	.258 .000	.346 .000	.427 .000	1.000 .

higher scores in mathematics assessments. However, there is a problem. As is evident from the findings of the TIMSS 1999 Video Study, even when the mathematics curriculum content provides problems set within their conceptual context, introducing the problems within that context to the students does not mean the students will understand them (Hiebert et al., 2003).

Accordingly, the variable of giving students opportunity to solve problems, whether the answer is immediately obvious (Table 6) and/or whether they have opportunity to determine their own method of solving them (Table 7), is not a completely adequate predictor of students' mathematics achievement. Table 6 shows no consistent relationship between increasing these opportunities for students and increasing their mathematics achievement. The only country where this relationship was at all evident in the present study was the United States. Here, the more frequently students had opportunity to solve problems, the more likely they were to gain the higher achievement scores. The range of average achievement for the United States students was 497.47 scale score points for the "never" option of the mathematics teachers' answer up to 535.37 scale score points for the "every or almost every lesson" option.

Furthermore, in relation to the problem-solving variable, and as Table 7 shows, the students of the teachers in the four-country sample who said they allowed students to find their own mode of solving problems during "about half the lessons" achieved the best results (505.63 scale score points). This group of students comprised 25.5% of the total sample. Students who had opportunity to do this kind of activity in "every or almost every lesson" achieved somewhat lower results (an average of 495.80 scale score points), while students whose teachers said they "never" provided this opportunity for their students had the lowest average achievement score. It seems, then, that the frequency with which students have opportunity to engage in independent activities needs to be carefully considered. Both too much and too little independent activity appears to adversely influence student achievement.

Table 5: How often Teachers Had Students Work on Problems for Which There Was No Immediately Obvious Method of Solution, across Countries

Options	*M*	*N*	Percentage	*SD*
Every or almost every lesson	517.54	997	5.35	81.862
About half the lessons	513.16	2,996	16.06	82.785
Some lessons	497.37	12,886	69.11	82.609
Never	486.58	1,767	9.48	84.609
TOTAL	499.95	18,646	100.00	83.180

Note: F = 171.708; df = 1; *p* =.000.

Table 6: How often Teachers Had Students Work on Problems for Which There Was No Immediately Obvious Method of Solution, by Country

Country		Every or almost every lesson	About half the lessons	Some lessons	Never
BGR	*M*	465.94	511.64	480.39	445.92
	SD	77.27	90.84	81.18	83.13
NLD	*M*	594.04	511.53	550.08	514.28
	SD	38.68	66.46	64.96	69.00
USA	*M*	535.37	521.43	498.73	497.47
	SD	72.97	81.56	77.04	74.78
SCG	*M*	475.77	493.86	474.12	491.29
	SD	89.41	85.70	87.41	94.16

The results presented in Table 8 also support this premise. Individual student activities are an important component of problem solving and of the process of attaining knowledge and concepts in general. Thus, giving students some degree of opportunity to find their own way to solve problems and discover mathematical concepts seems to be a necessary part of the mathematics teacher's teaching repertoire. These results also suggest that balance is needed in relation to students' individual activities in problem solving, students' group activities, and common teacher–student activities. It seems that such a balance is what teachers need to achieve to enhance their students' performance in these kinds of activities in mathematics.

Use of calculators and computers in mathematics teaching

The TIMSS 2003 testing procedures allowed students to use calculators for both the mathematics and science tests (Mullis et al., 2004). Data from the TIMSS assessments show relationships between mathematics teachers' policies on allowing students to use calculators and students' achievement in mathematics. A relevant hypothesis here is that those students whose teachers did not allow them to use calculators when studying mathematics achieved better results than those students whose teachers allowed them to use calculators when studying mathematics.

The thinking behind this hypothesis is that students who are not allowed to use calculators have the advantage of having to master all needed abilities and skills in the area of calculating, including the development of some concepts related to mathematical procedures and operations. A related consideration is that students who are allowed to use calculators are less likely than students who are not allowed to use them to advance their understanding of the concepts needed for doing calculations. In particular, we could argue that these students have less opportunity to exercise the mechanical operations of calculating, and so do not reach a deeper understanding of the essence and interconnectedness of these concepts. Reynolds and Farrell (cited in Keys, 1999) suggest that the early

Table 7: How Often Teachers Let Students Decide Their Own Procedures for Solving Complex Mathematics Problems, across Countries

Options	*M*	*N*	Percentage	*SD*
Every or almost every lesson	495.80	4,250	22.75	86.708
About half the lessons	505.63	4,703	25.18	80.698
Some lessons	500.19	8,535	45.68	82.822
Never	488.39	1,190	6.37	79.167
TOTAL	499.81	18,678	100.00	83.088

Note: F=.458; df =1 ; *p* =.499.

Table 8: How Often Teachers Let Students Decide Their Own Procedures for Solving Complex Mathematics Problems, by Country

Country		Every or almost every lesson	About half the lessons	Some lessons	Never
BGR	*M*	499.20	493.77	476.63	456.89
	SD	80.55	88.29	84.62	71.68
NLD	*M*	518.25	530.92	547.73	520.59
	SD	84.44	76.64	65.64	55.71
USA	*M*	506.77	511.59	502.61	470.22
	SD	84.11	75.83	74.68	91.79
SCG	*M*	483.01	477.57	471.36	501.64
	SD	89.45	86.68	87.19	86.19

introduction of calculators, and their too frequent use, is one of the reasons for the relatively poor performance of students in England in mathematics.

From Tables 9 and 10, we can see the group of students in the four-country sample who had the highest mathematics achievement in regard to the variable of calculator use was the group whose teachers reported allowing their students to use calculators in all mathematics lessons. This finding does not support the above hypothesis. A study relevant to this finding is one conducted by House (2002). Using data from TIMSS 1999, he found a significant negative relationship between the frequency of calculator use and students' mathematics achievement in Japan and a non-significant relationship between this use and achievement in the United States. In another study,

Table 9: Availability of Calculators during Mathematics Lessons, across Countries

Options	*M*	*N*	Percentage	*SD*
All lessons	522.74	7,770	49.69	76.86
Most	505.46	2,883	18.44	79.36
About half	469.31	1,049	6.71	85.04
Some	479.45	3,589	22.95	82.57
None	457.78	346	2.21	72.01
TOTAL	504.59	15,637	100.00	81.86

Note: F = 1006.474; df = 1; *p* = .000.

Table 10: Availability of Calculators during Mathematics Lessons, by Country

Country		All	Most	About half	Some	None
BGR	*M*	522.74	505.46	469.31	479.45	457.78
	SD	76.86	79.36	85.04	82.57	72.01
NLD	*M*	554.01	509.15	432.62	-	-
	SD	62.42	64.74	63.	-	-
USA	*M*	511.82	508.35	466.34	491.82	482.73
	SD	78.16	76.28	75.76	75.34	50.61
SCG	*M*	461.93	473.17	468.42	476.24	490.76
	SD	84.22	84.13	91.32	90.01	88.47

Keys (1999), drawing on the TIMSS 1995 results, found that across the countries participating in the assessment there was very little association between the extent of calculator use and mean mathematics score (Spearman's rho = -0.17).

In addition to considering the influence on achievement of students being able to use calculators during mathematics lessons, I was interested in determining what connection, if any, existed between students' mathematics achievement and the extent to which their teachers were able to provide them with access to computers to do various mathematics-based activities. One of the items in the mathematics teachers' questionnaire asked this question: "In teaching mathematics to the TIMSS class, how often do you have students who use a computer for the following activities?" Among the four activities that the teachers could check in the answer to this question was "discovering mathematics principles and concepts." This activity was particularly relevant to this present study because it acknowledges the computer as a tool that can be used for "discovering mathematics concepts and principles."

Not surprisingly, students who understand these concepts and principles gain the higher scores on tests of mathematics assessments. However, as shown in Table 11, teachers across the four countries in this present analysis were rarely using computers in their teaching of Grade 8 mathematics students. Only slightly more than 5% of the mathematics teachers reported computer use for this purpose. The table also shows that almost 95% of the students across the four-country sample were using computers only in "some lessons" or "never." The most successful subgroup of students in the four-country sample was the one that had access to computers when engaged in activities designed to help them discover mathematics concepts and principles. In general, these important findings indicate that computers were being mostly inadequately used in mathematics teaching and learning at the time of TIMSS 2003.

In Bulgaria and the Netherlands, not one teacher reported using computers to aid students' discovery of mathematics concepts and principles "every or almost every lesson" or for "about half the lessons" (see Table 12). In the other two countries, only very low

Table 11: Extent to which Computers Used for "Discovering Mathematics Concepts and Principles," across Countries

Options	*M*	*N*	Percentage	*SD*
Every or almost every lesson	480.58	105	2.15	86.08
About half the lessons	491.04	153	3.13	95.55
Some lessons	510.19	2,290	46.78	82.46
Never	501.30	2,347	47.95	77.49
TOTAL	504.69	4,895	100.00	80.85

Note: F = .047; df = 1; *p* = .828.

Table 12: Extent to which Computers Used for "Discovering Mathematics Concepts and Principles," by Country

Country		Every or almost every lesson	About half the lessons	Some lessons	Never
BGR	*M*	-	-	534.05	478.95
	SD	-	-	80.07	72.38
NLD	*M*	-	-	550.25	537.94
	SD	-	-	65.92	60.21
USA	*M*	408.39	495.75	502.54	497.38
	SD	54.36	94.34	80.69	78.45
SCG	*M*	509.45	461.41	436.02	460.41
	SD	79.36	100.09	102.07	84.59

percentages of the mathematics teachers selected these options. These findings again emphasize how little mathematics teachers were using the computer for this purpose at the time of TIMSS 2003. In an analysis of the factors affecting the mathematics achievement of students from Hong Kong SAR, Cyprus, and the United States who participated in TIMSS 1999, Papanastasiou (2002) confirmed that the students who attained the highest average achievement score were those who never used computers, while the lowest mathematics average score (490.28) belonged to the students who used computers for most of their lessons. The results of this analysis show that using computers frequently in mathematics teaching does not necessarily increase students' mathematics achievement.

Assignment of mathematics homework

For teachers, the aim of homework is generally to improve the quality of their regular teaching by giving students assigned tasks or some types of other activities that allow them to exercise and reinforce their previously developed abilities and skills, and to advance in some areas of their mathematics learning. Opportunity for assessing student learning is also a typical feature of mathematics homework, and the frequency with which homework is assigned also has relevance for the effectiveness of the role of homework within the teaching and learning of mathematics.

However, it is also important to take into consideration some other aspects of homework, such as the structure of homework and the types of items or activities included. Unfortunately, the TIMSS contextual data are not robust enough to supply this type of information. The only suitable such information available was that relating to the frequency with which teachers reported assigning homework. Table 13 shows how often teachers across the four countries assigned homework to their students. Students with the higher achievement scores were also the students whose teachers most often required them to complete homework. Furthermore, 85.92% of the students in the four-country sample received homework "every or almost every lesson." The findings here are confirmed by House (2002), who analyzed data relating to Japan from TIMSS 1999. He found that the more often students in Japan were given homework, they more likely they were to achieve the higher results on the mathematics test.

Table 14 shows the results for the "frequency of homework assigned" variable for the student samples from the four countries. The result for Serbia is particularly interesting, because it was unexpected. The best-performing group among the Serbian eighth-grade students was the one for which teachers reported assigning homework in "some lessons." This finding raises many questions about the basic role, purpose, contents, and overall quality of mathematics homework in the Serbian primary school mathematics curriculum. The findings for the samples of the other three countries show the expected results of student achievement across the defined subgroups of students. However, Keys' (1999) secondary analysis of the TIMSS 1995 data for England indicates that we cannot necessarily assume that simply assigning students more homework will increase their mathematics achievement. Rather, as Keys suggests, setting more homework needs to be associated with teacher feedback and follow-up activities if student achievement is to be raised.

A comparative study of the main characteristics of mathematics education in Japan, the United States, and Germany, with data drawn from TIMSS 1995 and the TIMSS 1999 Video Study, also reinforces the importance of assigning homework to improve students' performance in mathematics. This study, conducted by Schümer (1999), stresses the importance of the amount of homework that teachers assign. According to Schümer, the amount of homework that

Table 13: Frequency with which Teachers Assigned Homework, across Countries

Options	*M*	*N*	Percentage	*SD*
Every or almost every lesson	504.29	15,776	85.92	82.67
About half the lessons	480.35	1,796	9.78	83.96
Some lessons	477.23	790	4.30	83.37
TOTAL	500.79	18,362	100.00	83.28

Note: F = 183.668; df = 1; *p* = .000.

Table 14: Frequency with which Teachers Assigned Homework, by Country

Country		Every or almost every lesson	About half the lessons	Some lessons
BGR	*M*	481.65	489.34	467.34
	SD	84.23	91.25	79.05
NLD	*M*	545.05	521.17	-
	SD	65.22	70.85	-
USA	*M*	510.72	477.32	464.68
	SD	77.75	71.30	83.72
SCG	*M*	479.66	471.57	488.95
	SD	88.96	88.59	82.31

elementary school teachers assigned their students had a significantly greater positive impact on student achievement in Japan than in the United States and Germany. Based on the mathematics teachers' estimations, Grade 5 students in the United States had two hours and 20 minutes of mathematics homework weekly, while Grade 5 students in Japan had to complete an average of four hours and 19 minutes of homework each week.

Tests and examinations in mathematics teaching

The presence of some kind of final examination in mathematics is another important factor influencing students' performance in mathematics. The kind and degree of influence depends on several characteristics of the final intended examination, such as the topic areas covered, compatibilities with the TIMSS mathematics subtopics, when the examination is administered, the types and format of items in the examinations, the degree of difficulty of the items, and so on. To pass the examination, students obviously have to experience some process of adequate preparation for it. This preparation is usually provided in the schools, but it can also involve assistance from elsewhere, such as private tuition at home. The various kinds of preparation for the final examination that students experience can significantly affect their performance on the TIMSS mathematics domains. However, in this paper, I look at just one relevant variable: the format of the test items (Tables 15 and 16).

Five types of item formats are typically used in mathematics tests and examinations: (1) constructed-response only; (2) mostly constructed-response; (3) about half constructed-response and half objective-response (e.g., multiple-choice); (4) mostly objective; and (5) objective only. However, most tests of students' mathematics achievement employ a combination of these different item formats. The item format variable in TIMSS does not cover this diversity of options. Future assessments could consider constructing items that represent this diversity, so allowing students to answer test questions representative of the question formats their teachers typically use.

Table 16 shows that the subgroup of students whose teachers used only constructed-response items achieved the highest results in the TIMSS 2003 mathematics assessment across the four countries. It seems that students whose teachers give them opportunity to work mostly with constructed-response tasks have greater opportunity than other students to advance the abilities and skills they need for these kinds of activities and consequently achieve better performance in the field of mathematics. However, given the characteristic difficulties that students experience with some types of mathematics tasks and students' motivation for working on them, it may be that the other types of mathematics test item formats named as "objective" (e.g., multiple-choice) are easier for students to work with and so provide greater motivation for them to prepare for tests and examinations and to attempt to answer the questions once working on them.

Table 15: Types of Question Item Formats Used in the TIMSS 2003 Mathematics Assessment, across Countries

Options	*M*	*N*	Percentage	*SD*
Only constructed-response	512.52	6,522	35.08	86.47
Mostly constructed-response	498.86	6,211	33.41	80.29
About half constructed-response	483.07	3,987	21.44	76.61
Mostly objective	504.33	1,440	7.75	81.27
Only objective	493.16	432	2.32	101.54
TOTAL	500.56	18,592	100.00	83.09

Note: F = 156.663; df =1; p =.000.

Table 16: Types of Question Item Formats Used in the TIMSS 2003 Mathematics Assessment, by Country

Country		Only constructed-response	Mostly constructed-response	About half constructed-response	Mostly objective	Only objective
BGR	*M*	519.31	475.33	475.73	478.97	454.67
	SD	90.86	81.35	73.73	79.95	108.46
NLD	*M*	546.71	524.36	524.76	570.78	437.96
	SD	65.87	67.09	93.08	58.43	29.83
USA	*M*	524.12	511.88	486.06	511.88	545.13
	SD	83.38	75.53	74.34	77.99	77.40
SCG	*M*	480.19	476.14	480.47	485.29	468.20
	SD	88.88	87.21	85.38	97.82	67.14

Summary and conclusions

This study considered some contextual factors related to the teaching and learning of mathematics. The analyses involved exploring a chosen subset of variables from the TIMSS 2003 mathematics assessment and examining the influence of this subset on student achievement both within and across the four countries that were the focus of this study.

The results showed differences in the influence of some of these factors across the four countries and also indicated that some of the variables used in the TIMSS teacher questionnaire are not reliable predictors of students' success in the field of mathematics. In some cases, the variables had the expected or predicted influence, but in other cases, the results were unexpected. This lack of predictability was particularly evident in relation to those variables in which the mathematics teachers were expected to express their attitudes rather than provide objective information. It seems, therefore, that teacher subjectivity had some level of influence on teachers' views on the mathematics teaching and learning opportunities covered by the questionnaire.

Differences in complex cultural contexts also appear to have had an impact on the influence of these factors in the four countries. As Papanastasiou (2002) has noted, and as appears to be the case in this present study, the same variables do not always have the same effects on different students, depending on the cultural context that the students are in. This is the reason why anyone making decisions about educational outcomes needs to take such factors into account, as Papanastasiou stresses.

As previously indicated, some of the items and their variables from the TIMSS 2003 teacher questionnaire chosen for examination in this study did not adequately cover all possible influencing factors. For example, the intention behind the question about use or non-use of calculators in mathematics teaching was to collect what could be termed general facts, and the supposition behind that intention was that those students who were not allowed to use calculators when engaged in mathematics activities would achieve better results on the mathematics test. The reason why was that they would have had greater opportunity than students who used calculators to advance their own calculating abilities and skills. Essentially, the questionnaire contains no items that allow determination of whether students who do not use calculators also develop concepts specifically in the area of calculating.

In summary, the students' achievement results across the four countries frequently showed a broader diversity in regard to the influence of the variables on achievement than might have been considered during construction of the questionnaire. As such, it is difficult to fully rely on these variables as predictors of students' achievement. This consideration represents one of the shortcomings in the structure of the mathematics teachers' questionnaire, as well as some of the secondary analyses based on the data generated by this questionnaire.

In general, we can conclude that students' mathematics achievement depends on many factors from the contexts of school, classroom, and home. However, crucial factors are also found in the context of teaching and learning mathematics through classroom activities. These factors of students' success in the field of mathematics make for a complex system of influences, most particularly as they relate to teaching and learning methods, the content of activities designed to help students discover and attain mathematical concepts, and the different opportunities teachers provide to advance their students' understanding of attained concepts.

The across-country results in this study also give us some structural sense of how the teaching and learning of mathematics influences students' achievements in this field. However, we need to acknowledge the variability of findings relating to some variables both across and within the countries. In line with the findings of Papanastasiou's (2002) cross-country analyses of differences in mathematics teaching, what we can conclude from the data available from the TIMSS 2003 teacher questionnaire and analyzed here is that this information provides some, but not complete, guidance on how teachers in the four countries can improve the teaching and learning of mathematics.

References

Adolfsson, L., & Henricsson, W. (1999). Different methods—different results: How the Swedish results in mathematics vary depending on methodological approach. *Educational Research and Evaluation, 5*(2), 127–138.

Birenbaum, M., Tatsuoka, C., & Yamada, T. (2004). Diagnostic assessment in TIMSS-R: Comparison of eighth graders' mathematics knowledge states in the United States, Japan, and Israel. In C. Papanastasiou (Ed.), *IEA International Research Conference: Proceedings of the IRC-2004 TIMSS* (Vol. 2), (pp. 19–30). Nicosia: Cyprus University Press.

Bos, K., & Kuiper, W. (1999). Modelling TIMSS data in a European comparative perspective: Exploring influencing factors on achievement in mathematics in Grade 8. *Educational Research and Evaluation, 5*(2), 157–179.

Hiebert, J., Gallimore, R., Garnier, H., Givven, K. B., Hollingsworth, H., Jacobs, J., Chui, A. M. et al. (2003). Understanding and improving mathematics teaching: Highlights from the TIMSS 1999 Video Study. *Phi Delta Kappan, 84*(10), 768–775.

House, J. D. (2002). Instructional practices and mathematics achievement of adolescent students in Chinese Taipei: Results from the TIMSS 1999 assessment. *Child Study Journal, 32*(3), 157–178.

Keys, W. (1999). What can mathematics educators in England learn from TIMSS? *Educational Research and Evaluation, 5*(2), 195–213.

Kiamanesh, A. R. (2004). Factors affecting Iranian students' achievement in mathematics. In C. Papanastasiou (Ed.), *IEA International Research Conference: Proceedings of the IRC-2004 TIMSS* (Vo1. 1), (pp. 157–169). Nicosia: Cyprus University Press.

Mullis, I. V. S., Martin, M. O., Gonzales, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades.* Chestnut Hill, MA: Boston College.

Papanastasiou, E. (2002). Factors that differentiate mathematics students in Cyprus, Hong Kong, and the USA. *Educational Research and Evaluation, 8*(1), 129–145.

Schümer, G. (1999): Mathematics education in Japan. *Journal of Curriculum Studies, 31*(4), 399–427.

Turner, J. C., & Meyer, D. K. (2000). Studying and understanding the instructional contexts of classrooms: Using our past to forge our future, *Educational Psychologist, 35*(2), 69–85.

The International Association
for the Evaluation of
Educational Achievement

Across almost 50 years and more than 20 research studies of cross-national achievements, IEA has contributed substantially to the development of a worldwide community of researchers in educational evaluation. The aim of the IEA International Research Conference-2006 (IRC-2006) was to provide an international forum for the exchange of ideas and information on important educational issues investigated through IEA research.

An IEA study typically takes four to five years from inception through to publication of the first round of international reports and release of the international database to a broad community of researchers. Some researchers working directly on projects have the chance to meet one another during the period of participation in a study. However, geography largely constrains further opportunities to exchange ideas, approaches, findings, and concerns. The biennial IEA IRC provides a forum for creative dialogue among researchers, leading to a greater understanding of the numerous roles that education plays in the development of nations and in shaping individuals.

The IRC-2006 featured researchers from six continents presenting results of their analyses of IEA data in the fields of mathematics and science (TIMSS), reading and literacy (PIRLS), civic education and citizenship (CivEd), and information technology in education (SITES). Papers in this volume and its companion volume cover a range of topics around these themes.