The contribution of TIMSS to the link between school and classroom factors and student achievement

Marjolein Drent , Martina R.M. Meelissen & Fabienne M. van der
Kleij
Published online: 09 Oct 2012.

PLEASE SCROLL DOWN FOR ARTICLE

Routledge
Taylor & Francis Group

The contribution of TIMSS to the link between school and classroom factors and student achievement

MARJOLEIN DRENT, MARTINA R.M. MEELISSEN and
FABIENNE M. VAN DER KLEIJ

Worldwide, the interest of policy-makers in participating in studies from the International Association for the Evaluation of Educational Achievement (IEA), such as Trends in International Mathematics and Science Study (TIMSS) has been growing rapidly over the past two decades. These studies offer the opportunity to relate the teaching and learning context to student achievement. This article presents the results of a systematic review of the research literature on TIMSS. Its main purpose is to find out to what extent TIMSS has contributed to insights into 'what works in education and what does not', particularly with regard to school and classroom factors. The review was guided by a generic framework developed within the tradition of educational effectiveness research. The review showed that: (a) since 2000, the number of publications which use TIMSS data for secondary analyses aimed at explaining differences in student achievement has increased strongly; (b) a number of studies, especially older ones, did not take account of the specific sample and test design of TIMSS; and (c) there are large differences between countries in school and classroom factors associated with student achievement. In the light of these results, we discuss the benefits and limitations of country and system comparisons.

Keywords: school effectiveness; academic achievement; mathematics education; science education; literature reviews

Introduction

Founded in the late 1950s, the International Association for the Evaluation of Educational Achievement (IEA) has initiated 32 international comparative studies in education. These studies have addressed a variety of educational subjects, such as achievement in reading literacy, mathematics, science, civic education and the use of computers in education. The interest of policy-makers in participating in these international studies has been growing rapidly over the years. With only 12 countries involved in The First International Mathematics Study (FIMS, 1963–1967), there are currently 69 countries or educational systems (such as

Marjolein Drent works since 2006 as an researcher on TIMSS, and she is an information specialist at the Department Library and Archive of the University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands; e-mail: marjolein.drent@utwente.nl. Her main interests are information literacy and international comparative educational research.

Martina R.M. Meelissen is an assistant professor in the Department of Educational Organisation and Management of the University of Twente. As National Research Coordinator she is involved in TIMSS since 2003. Her main interest is educational effectiveness research with a special interest for gender differences.

Fabienne M. van der Kleij has a master degree in Educational Science and Technology (EST) of the University of Twente. She is currently a PhD student working at CITO (Dutch institute for testing and assessment).

Belgian Flanders) from all over the world taking part in TIMSS 2011 (Trends in International Mathematics and Science Study) and PIRLS 2011 (Progress in International Reading Literacy Study).

In general, participating in an IEA study is quite an enterprise in terms of effort, time and costs. In TIMSS and PIRLS, for example, around 4500 students from a representative sample of at least 150 schools are tested in each country. Besides the test, the students, their teachers, their school principals and the curriculum experts from each country fill in the questionnaires about the context of teaching and learning. Both test and questionnaires need to be translated and adapted for each country without compromising the international validity of these instruments. Extensive administrative procedures have been developed to ensure that the conditions for the assessment are identical in each classroom. Trained scorers then score students' answers on the open-ended test items (about half the items). At least 200 student responses to each item are scored independently by two scorers, in order to determine the inter-rater agreement. Finally, the data from all the countries are cleaned, weights are calculated and item response analyses are conducted to calculate the achievement scores at country and individual levels.[1]

However, in return for these investments, these studies offer more than just a ranking of countries on their average achievement scores. First, the information from the questionnaires provides an in-depth insight into the diversity of the characteristics of education systems all over the world. In addition, countries which have participated in more than one of the TIMSS or PIRLS projects (TIMSS and PIRLS are repeated every fourth or fifth year, respectively) are able to analyse trend data across assessments. But, most importantly, many IEA studies are classroom-based, which means that the context of teaching and learning can be related to educational outcomes, such as student achievement.

This last benefit was one of the initial goals of the founders of IEA in 1958 (Postlethwaite 1995, Postlethwaite and Ross 1992). They viewed the world as 'a natural educational laboratory, where different school systems experiment in different ways to obtain optimal results in the education of their youth' (http://www.iea.nl). By sharing the outcomes and characteristics of these 'experiments', international comparative studies are intended to offer empirical-based information on what matters in education.

However, due to the scale and complexity of these projects, as well as the realization that the determination of causality is difficult because these studies are cross sectional, the goals of the IEA projects became less ambitious during the 1990s (Gustafsson 2008). The current international reports on these studies provide mainly descriptive information, leaving the more in-depth, secondary analyses to the participating countries (e.g. Mullis *et al.* 2008). To make this possible, the data from these studies have been made freely accessible and are well documented for all those who are interested in conducting further analyses of the data (see http://timss.bc.edu).

The question remains, however, to what extent have researchers and policy-makers taken advantage of this opportunity? TIMSS is the largest

IEA international comparative study in terms of longitude, participation and number of research populations. TIMSS is not only classroom-based but also, unlike other international assessment studies such as the Programme for International Student Assessment (PISA), aims to assess what the curricula in the participating countries intend their students to learn. Despite these opportunities, Beaton and Robitaille (2002) concluded, six years after the first TIMSS data became available, that the number of studies conducting secondary analyses of TIMSS remained limited. Most of what was published at that time focused mainly on the achievement results and the ranking of the participating countries (Beaton and Robitaille 2002).

With so much data collected over the past 15 years (TIMSS-1995, -1999, -2003, -2007 and TIMSS-Advanced 2008), it seems likely that the number of studies that use TIMSS data for secondary analyses will have increased compared to what was reported by Beaton and Robitaille (2002). This is encouraged by IEA, who have organized the International Research Conference since 2004. This conference is held every two or three years and has the specific aim of promoting the exchange of results between researchers of secondary analyses of the data from the IEA projects (Papanastasiou 2004).

The purpose of this study is to analyse the scientific contribution and impact of TIMSS by conducting an in-depth analysis of TIMSS-based studies which address the relation between the characteristics of the educational setting at school and classroom levels and the student achievement. To achieve this, a systematic review was conducted of all TIMSS publications since the first TIMSS in 1995. This systematic review includes an extensive search of different databases, a selection of publications based on a number of inclusion and exclusion criteria; a quality assessment; and a content analysis of the studies finally selected.

Research questions

The aim of this study was addressed by systematically analysing studies of TIMSS as they have been published in scientific journals and books since the release of the TIMSS-1995 data. For this systematic review, each study conducting secondary analyses of TIMSS data was examined with the following research questions in mind:

(1) What are the main characteristics and the impact of studies using TIMSS achievement data to address the relation between country, school, class and student characteristics and student achievement?
(2) How can studies using TIMSS achievement data to address the association of malleable school and classroom factors with student achievement be characterized in terms of their scientific quality?
(3) To what extent have secondary analyses of TIMSS achievement data contributed to theories of educational effectiveness regarding school and classroom factors, in terms of the extent to which the findings of these studies support or add to existing knowledge?

The first research question aims to provide a general overview of the characteristics and scientific impact of TIMSS-related studies in which student achievement is the dependent variable. The second and third research questions focus specifically on studies in which school and classroom factors in relation to student achievement are analysed. For the identification of these factors, a conceptual framework by Scheerens (2008) was used, which was developed within the tradition of educational effectiveness research. The main goal of educational effectiveness research is to analyse the association of the conditions of schooling that may potentially enhance effectiveness with outcome measures such as student achievement (Creemers 2006, Scheerens and Creemers 1989). A further explanation of this framework is given in the next section.

Conceptual framework

The generic framework we used to guide our systematic review is the 'Basic system model on the functioning of education' (Scheerens 2008) (see Figure 1). This classic input-process-output model fits within the tradition of educational effectiveness research (Creemers 1994). The framework has a multilevel structure, which means that the model includes factors at student, classroom, school and system levels. Input factors, such as teacher characteristics or parental support, are assumed to be related to outputs (e.g. student achievement) through process factors. Process factors are part of the so-called 'black box', in which teaching and learning take place and in which inputs are transformed into outputs. Process factors can be divided into factors at the school level (e.g. educational leadership and school climate) and factors at the classroom level (e.g. effective learning time, structured instruction and opportunity to learn). In order to assess the added value of schooling, student achievement should be adjusted first by student background characteristics, such as previous achievement and socio-economic status. Furthermore, it is assumed that the relation of input and process factors with student achievement is moderated by contextual factors at system and school lev-
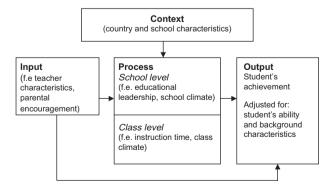

Figure 1.   A basic system model on the functioning of education.
Source: Adapted from Scheerens 2008: 4).

Table 1. Overview of effective process factors in education based on meta-analyses by Scheerens *et al*. (2007).

Process factors school level	Process factors class level
Achievement orientation and expectations	Achievement orientation and expectations
Curriculum quality/opportunity to learn	Curriculum quality/opportunity to learn
Structured instruction	Structured instruction
Differentiation, adaptive instruction	Differentiation, adaptive instruction
Feedback and reinforcement	Feedback and reinforcement
Evaluative potential	Evaluative potential
School climate	Class climate (creation and dimensions of)
Educational leadership	Effective learning time
Consensus and cohesion among staff	
Parental involvement	

els, such as student body composition, school size or educational policy (Creemers 2006, Scheerens 2008).

Based on meta-analyses which resulted in the framework discussed above, Scheerens *et al*. (2007) identified a number of process factors that enhance effectiveness or 'black box' factors related to high achievement. These factors are summarized in Table 1.

In this systematic review, the list of process factors at school and classroom levels was the starting point for the identification and categorization of factors in the selected TIMSS studies, including national options (national questionnaire items that were not part of the international TIMSS instruments).

Method

Systematic review

The present review followed the method of Petticrew and Roberts (2006). Using this method, comprehensive searches were conducted to find all relevant peer-reviewed articles. Next, inclusion and exclusion criteria were formulated to determine the suitability of the studies for answering the research questions of the review. This form of review implies careful reading and an analysis of studies. Explicit criteria were formulated for evaluating the methodological quality of each selected study. The findings of the selected studies are summarized by a narrative synthesis of the different studies. In the following section, the different steps within the review are further discussed.

Search keys and databases

The search was carried out in March 2010. The following online databases were used: Web of Science (www.isiknowledge.com); Scopus (www.scopus.com); Picarta (a Dutch electronic database, www.picarta. nl); and ERIC and Psycinfo (provided through Ebscohost). These

databases were explored using the following search key: 'TIMSS' OR 'trends in international mathematics and science' OR 'third international mathematics and science'. In TIMSS 1995 and 1999, the first 'T' of TIMSS referred to 'Third' because it was the successor to the FIMS and the Second International Mathematics Study (SIMS) as well as the First International Science Study (FISS) and the Second International Science Study (SISS). These studies had not been set up to be trend studies. However, TIMSS 1999 was set up as a trend study, so the 'T' became the abbreviation of 'Trends'. The search keys were used for the title, abstract or list of keywords.

Inclusion and exclusion criteria

In order to select the studies for the review, five inclusion criteria were formulated. A pilot of the review was first conducted on 25 studies. This pilot was used to evaluate the usefulness of the systematic review and to assess the number and type of studies which have been carried out. Furthermore, through the pilot the usefulness of the selection criteria was evaluated (Petticrew and Roberts 2006). The pilot also determined the final order of the criteria and was used to sharpen these criteria. As a result, the following five inclusion criteria were formulated:

(1) The study is published in a peer-reviewed journal or book. Non-research publications, like book reviews, editorials or popular articles, were excluded. Published or unpublished dissertations were also included, provided it was possible to acquire the full text (online or on paper) of the dissertation. If a study had been published both as a dissertation and as a peer-reviewed article, only the article was included.

(2) The study is published in English. The use by researchers of studies published in English, and therefore their impact, was assumed to be higher compared to studies published in other languages.

(3) The study uses 'regular' TIMSS data for analysis. This means that studies using data from the TIMSS video study and the TIMSS performance test (1995) were excluded. This selection step was necessary for criterion 4.

(4) The dependent variable of the study is the mathematics or science achievement of students. Studies conducting only a descriptive analysis were excluded, as were those studies in which the dependent variable was not achievement but another variable, such as attitudes of students. Although the attitudes, motivation or self-confidence of students are sometimes considered as non-cognitive output factors of education, it can also be argued that these affective factors are student characteristics for which achievement needs to be adjusted. Self-confidence, for example, can be regarded as an indicator of prior achievement in cases where a true measure of prior achievement is unavailable, which is the case in TIMSS (Dumay and Dupriez 2007, Kaya and Rice 2010, Van den Broeck et al. 2006,

Xin *et al.* 2004). Therefore, in this study, affective factors were not regarded as an educational output factor.

(5) At least one malleable process factor at school and/or classroom level, related to mathematics or science achievement, is included in the study. This study focuses specifically on process factors which enhance effectiveness that are part of the 'black box' mentioned earlier. Studies that focus only on input, context or student characteristics (or combinations of these factors) were excluded in this step. The decision to include only studies at school and classroom levels also meant that malleable factors at system or country level, for example, the influence of central exit examinations (Jurges *et al.* 2005), were not part of this review.

For the application of the first four selection criteria, bibliographic information like abstract, article title, keywords and source title was used. When it was not clear whether a study fulfilled a certain criterion, it continued to be included in the selection. For the fifth criterion, the full text of the remaining publications was reviewed. For this final selection step, all remaining studies were evaluated independently by two reviewers.

The first research question was addressed by categorizing the studies after the application of selection criterion 4. This means that all studies in which mathematics or science achievement was the dependent variable were categorized by the year of publication, year of data collection, subject, population (age and country) and the number of citations, using the citation index of the Web of Knowledge (December 2010–January 2011).

Quality assessment

A data extraction form was completed for each of the studies selected after the application of selection criterion 5. The main goal of using this form was to assess the quality of the remaining studies (research question 2). The evaluation of the quality of the final selection was carried out by three reviewers. Each study was assessed by two reviewers to check for inter-rater reliability during the reviewing process.

The data extraction form was based on the framework for appraising a survey as formulated by Petticrew and Roberts (2006: box 5.7, 142–143). The data extraction form entailed questions about the research question, use of literature, sampling issues, measurement, method and statistical analyses, presentation of the results and the validity of conclusions. In the data extraction form, a short summary was given for each of the above aspects. The critical appraisal of each article was aimed at the identification of possible susceptibility to bias. Both reviewers evaluated the quality of the articles on a three-point scale. When the reviewers did not agree about the quality of an article, the study was further discussed by the reviewers. In this study, three aspects concerning the data and the statistical issues played a decisive role in the judgement of the quality. Firstly, in accordance with the theory of educational effectiveness research (e.g. Creemers 1994), achievement measures should be controlled for or

adjusted by student characteristics (such as social economic status and previous achievement) in order to assess the added value of schooling. TIMSS does not measure previous achievement, but several indicators are available in the TIMSS data that refer to student background characteristics, such as the educational level of the parents and the gender of the student. Studies not including any of these variables in their analyses were considered to be of low quality. Studies of a low quality were excluded from the final analysis.

Secondly, the study should take into account the specific research design of TIMSS, the so-called nested or clustered design. In TIMSS, schools are sampled first and then within each school one or more classes are sampled. Students are therefore not sampled randomly, but are nested in classes. This means that a study should either use a method like hierarchical linear modelling (multilevel analysis) or, when analyses are conducted at one level, adjust the standard errors for clustering effects.

Finally, we examined how the TIMSS plausible values were used in these studies. In the TIMSS data-set, five plausible values of mathematics and science achievement levels for each student were estimated through a process of imputation, even though each student answered only a part of the TIMSS assessment item pool. Plausible values cannot be viewed as estimates of individual student scores, but rather as imputed scores of students with similar response patterns and background characteristics in the sampled population (Foy *et al.* 2008, Mislevy *et al.* 1992). Ideally, for each subject all five plausible values should be included in the analysis. However, if this was not the case (for example, only the first plausible value was used), the study was not automatically evaluated as being of low quality. Only when studies indicated that they had averaged the plausible values, a study was assessed as low quality. According to von Davier *et al.* (2009), averaging plausible values leads to biased estimates.

Content analysis

The last step in this review was the synthesis of the results of the remaining studies. Often a systematic review of quantitative studies is followed by a quantitative meta-analysis. In order to decide whether such a meta-analysis would be useful, it is important to determine the heterogeneity of the studies under review in terms of population, method and variables (Petticrew and Roberts 2006). According to Petticrew and Roberts, a quantitative meta-analysis is not useful if many different variables or indicators are included in the review, which is the case in this study. In comparison to systematic reviews of studies on one subject (for example, feedback or class size), this study includes all the different process factors addressed in TIMSS. This means that there are only a few studies addressing the same variable or indicator.

We therefore summarized our results through a narrative synthesis. This means that primarily words and text are used to summarize and

explain the findings of the synthesis (Rodgers *et al.* 2009). Based on the data extraction forms, all studies were characterized by their most important characteristics, such as the population studied, country studied and the TIMSS data used. In this study, we decided to make use of a content analysis (Dixon-Woods *et al.* 2004) to summarize the research results of the study. Through a content analysis we were able to systematically categorize the process factors studied in the articles, based on the list as described within our conceptual framework (see Table 1). Two reviewers coded the factors for each study. Differences in the categorization were discussed until agreement was reached. The factors which showed an association with achievement were described, as was the level of significance (research question 3). Because of the large number of process factors available in TIMSS data that may potentially have an influence, this review has limited itself to reporting only the direct relations of malleable process factors with achievement.

Results

Search results

Querying the selected literature databases resulted in 1644 hits. The highest percentage (39%) of TIMSS-related publications was found in ERIC and the lowest percentage (10%) was found in Web of Science. After removing the duplicates, 985 unique publications were identified.

After applying the first selection criterion (peer-reviewed journal, book or dissertation), the number was reduced by almost half. As two of the books specially dedicated to TIMSS were peer reviewed, all 46 chapters (individual articles) of these books were added to the selection (Howie and Plomp 2006, Robitaille and Beaton 2002). This resulted in 594 publications, most of which (523) were published in English. Only 330 publications remained in the review after applying selection step 3 (the study uses 'regular' TIMSS data). In the excluded studies, 'TIMSS' was only referred to and the regular TIMSS data were not used for analyses. The application of the fourth selection criterion (the dependent variable in the study should be the mathematics or science achievement of the students) resulted in a further reduction down to 201 publications. Of all TIMSS-related publications found in the initial search (including the individual book chapters), 8% satisfied all five inclusion criteria. This means that, since TIMSS 1995, 78 studies have been published that include at least one malleable process factor at school or classroom level. The studies selected after criterion 5 consisted of 59 journal articles, 15 book sections and four dissertations. However, it became apparent that 10 studies could not be used for further analysis, either because the full text was not accessible or because the effects of variables for the entire mathematics or science test were not reported. This means that only 68 studies provided satisfactory information for the quality assessment.

Characteristics of the selected studies after criterion 4

In order to answer our first research question, this section provides a description of the characteristics of TIMSS studies that meet the first four criteria in this review. This includes all studies ($n = 201$) in which mathematics or science achievement is the dependent variable, and its association with students' background characteristics, contextual characteristics input factors and/or process factors was analysed. Table 2 provides an overview of the characteristics of TIMSS data used in the studies that remained in the selection after criterion 4.

Table 2 shows that only 4% of TIMSS-related studies in which mathematics or science achievement is the dependent variable were published before 2000. In 2006, a new TIMSS book, *Contexts of Learning Mathematics and Science* by Howie and Plomp, was published and in 2008 the journal, *Studies in Educational Evaluation*, had a special issue on TIMSS. These years have therefore been the most 'productive' years in terms of the number of TIMSS-related studies with achievement as the dependent variable.

Only a few authors systematically use TIMSS data for secondary analysis. The 201 publications were written by 242 unique authors and 49 authors contributed to more than one study. The most productive author in the selection had published 20 articles on TIMSS.

Table 2. Overview of the characteristics of TIMSS-related studies based on title, keyword and abstract, in percentages studies that fulfilled criterion 4,[a] $n = 201$.[b]

Characteristics	% studies
Publication years	
<2000	4
2000–2002	27
2003–2005	22
2006–2008	40
2009–early 2010	8
TIMSS study	
TIMSS 1995	54
TIMSS 1999	35
TIMSS 2003	21
TIMSS 2007	1
Research population[c]	
Population 1 (grade 3 and/or 4)	14
Population 2 (grade 7 and/or 8)	80
Population 3 (grade 12 pre-university/final year)	10
Subject of TIMSS test	
Mathematics	54
Science (or science domains)	13
Both tests	29

[a]Inclusion criterion 4: (a) published in peer-reviewed journal or book, (b) published in English, (c) used regular TIMSS data and d) student achievement is dependent variable.
[b]Percentages do not add up to 100%: some used data of more than one study or population, 11 studies did not provide information about year of study, population or subject.
[c]Only in TIMSS 1995 two adjacent grades were tested. TIMSS 1999 focused only on grade 8 and TIMSS 2003 and 2007 focused on grades 4 and 8.

Eighteen articles (9% of the selected articles) used data from more than one TIMSS study. Although most of the studies were published after 2003, Table 2 shows that the majority of studies used data from TIMSS 1995 or TIMSS 1999. The TIMSS data are freely accessible one year after the main data collection. Because it also takes some time before a manuscript is accepted and published, it is not surprising that, until the beginning of 2010, only one publication had used the data from TIMSS 2007.

This is probably also one of the reasons why there are fewer studies using grade 4 data. Grade 4 students were assessed in 1995 (and in 2003 and 2007), but not in 1999. Furthermore, the number of countries participating with grade 4 is lower compared to countries participating with grade 8. However, even when this is taken into account, the contribution of grade 4 studies still seems to be quite low: only 14%.

Population 3 in 1995 consisted of two subpopulations. The first subpopulation, students in their final year of secondary education, was tested only in TIMSS 1995. The study among the second subpopulation, grade 12 pre-university students, was conducted in 1995 and again very recently in 2008. This limited number of assessments explains why only 10% of the studies used the data from upper secondary education (see Table 2)

There seems to be less interest in science achievement (or the separate science subjects), as the majority of studies used the TIMSS mathematics achievement scores as outcome measure (Table 2). A possible explanation is the differences between countries in how science is taught—as separate subjects or as a comprehensive science topic—and the differences between countries in the role of science subjects in their curriculum (Martin *et al.* 1997, 2008). This complicates the conducting of studies in which the results of more than one country are compared.

The data of 64 countries were used in the TIMSS-related studies in which mathematics or science achievement is the dependent variable. We found that about half the studies used data from more than one country in their analyses (not in table). Most studies focused on data from the western and high-scoring Asian countries such as the USA, Australia and Japan (Table 3). All countries in the top five of Table 3 have participated in all TIMSS studies so far. The data of African and Arab countries were used limitedly within the selected studies. An exception was the use of

Table 3. Top five of countries of which country data are used in percentages studies that fulfilled criterion 4, $n = 201$.[a]

Countries	% studies
United States	41
Japan	30
Australia	26
Netherlands	24
Hong Kong	24

[a]Sixteen studies did not provide information about the country data used.

the TIMSS data from South Africa, which is one of the lowest scoring countries in TIMSS, (16% of all studies, e.g. Howie 2002, 2004, 2005a, b, Howie *et al.* 2008, Howie and Scherman 2008).

Impact

The impact of each peer-reviewed article that fulfilled criterion 4 (201 articles) was calculated by using the citation index of the Web of Knowledge (December 2010–January 2011). This means that the calculation was limited to citations in articles published in journals that have a so-called 'impact factor'. The impact factor of a journal reflects the average number of citations to articles that are published in the journal. In this study, the 201 selected articles were cited 480 times. On average, an article from our selection was cited 2.4 times. The study of Rindermann (2007) was cited the most (40 citations). This article discusses the meaning of the TIMSS, PIRLS and PISA tests as indicators of student achievement. Rindermann (2007) concludes that these tests largely measure intelligence rather than student performance. The article of Baumert *et al.* (2009) is an immediate reaction to this. However, most of the 40 citations do not discuss Rindermanns' conclusions. More than half the articles that referred to Rindermann focused on intelligence or cognitive ability (e.g. Lynn *et al.* 2009, Rushton and Templer 2009).

Until January 2011, about half the articles had not, or not yet, been cited in journals with an impact factor. The impact of the 68 articles that included process factors (criterion 5) is a little higher. On average, they were cited 2.8 times. The 59 articles in peer-reviewed journals that remained in the selection after the application of criterion 5 were published in 31 journals: 16 of these 31 were journals with an impact factor.

Quality assessment

After the application of the fifth criterion, 68 articles, book chapters or dissertations were included in the quality assessment. Despite the fact that all selected publications were peer reviewed, 34 publications (49%) were assessed as being of low quality. Of all 68 studies, 29% were labelled as 'high quality' and 21% as 'satisfactory'. The studies that were rated as being of high or satisfactory quality have been included in the content analysis. Seven of the 13 articles which were evaluated as of high quality were published in a journal with an impact factor. In almost half the 'low quality' publications, the authors did not take the nested design of TIMSS into account and did not mention the consequences of this design for the standard errors. Also, there were quite a number of studies in which student achievement was not adjusted by student background characteristics at all and were therefore judged as being of low quality.

Based on what we found, we decided to be less strict with the assessment of how these studies dealt with the so-called plausible values of TIMSS (indicators for achievement). It emerged that many studies pro-

vided very little, if any, information on how the plausible values were used in the analysis. Therefore, in line with von Davier *et al.* (2009), only studies that used the average of the five plausible values were evaluated as being of low quality.

Other quality issues did not result in direct exclusion. It was often a combination of issues that resulted in a negative judgement. Some examples of these kinds of issues were:

- The level of analysis was different from the level of conclusion/implications. For example, analyses at student level (e.g. students' individually perceived characteristics of teacher instruction) resulted in conclusions assuming causality at class or teacher level.
- The effects of many variables were explored, without any theoretical argumentation. The risk of including many variables (in one study it was over 50 variables) in one analysis is that the results may be spurious, due to capitalization on chance.
- The research was limited to the effects of individual items, without any argumentation provided, while data reduction by constructing composites or indices would have been possible. Data reduction refers to the limitation of variables in the analysis by constructing composites or indices based on a set of interrelated items. The underlying assumption is that these items together represent one factor. One example is self-confidence in learning mathematics, which is measured by a number of statements in the TIMSS student questionnaire.
- Across studies, the same factors (variables) were given different labels or that the same labels were given to different variables. For example, the indicator for 'opportunity to learn' was also labelled as 'number of topics taught', 'content coverage' or 'topic coverage index'. As long as authors indicate how each factor or composite is measured in their study, this is only a problem for the comparison of studies. However, some authors were not clear about this, and that limits the reproducibility of the study.

In most cases, studies with more than just one of these issues were considered as being of low quality.

In 'high' quality studies, none of these issues was present. One example of a recent study that was rated 'high quality' was that of Kaya and Rice (2010). In our final selection, this study is unique, because it is the only study that used the science achievement data from the fourth grade data-set from TIMSS 2003. The study provided a clear rationale for the problem statement, the selection of the five countries analysed in this study, as well as for the variables included in the analysis. The study included all five plausible values in their analyses (using multilevel analyses). Factor analyses were conducted to develop composites and the number of variables included in the analysis was limited, but did include student background variables. Finally, the results were discussed in the light of some of the limitations of TIMSS (e.g. not accounting for stu-

dents' previous performance and limited conceptualization of students' socio-economic status).

Content analysis

The data extraction forms were used as a starting point for the content analysis. These forms included an overview of all process factors that enhance effectiveness (see Table 1). The analysis was conducted to find out to what extent the results of these studies supported the list of effective process factors at school and classroom levels (research question 3). During the categorization of TIMSS indicators into effective process factors, it became clear that the definitions of the process factors were not very precise and sometimes overlapped. However, it was possible to categorize all the variables analysed in the TIMSS articles as indicators of one of the effective process factors.

Table 4 shows the results of the analyses for the 34 studies that were assessed as being of satisfactory or high quality. Often, studies in which no effects are found will either not be published or the non-significant effects of variables will not be reported within a published study (so-called publication bias). It is therefore not possible to provide a complete overview of all the variables analysed, whether they show a significant relation or not. For this reason, only indicators that showed a significant association ($p < 0.05$) in at least one study, country or data-set were included in Table 4. The reported non-significant relations were only included in the country comparison of a selected set of indicators, presented in Table 4.[2]

Table 4 also shows the direction of significant associations. If the results were contradictory (for example, a positive relation was found for an indicator with achievement in one country, subject or study, while in another context a negative relation was found for the same indicator with achievement), the indicators were marked with a question mark. In some studies, the response options of variables were analysed as individual (dummy) variables. In some cases, these variables showed a non-linear association. For example, the study of Ma and Papanastasiou (2006) found that 'working in small groups' was negatively related to achievement when it occurred very often during lessons, positively related when it occurred sometimes and not related if it did not occur during lessons. These cases are indicated with a plus–minus in Table 4.

In the selected TIMSS studies, indicators can be found for most of the process factors that enhance effectiveness mentioned by Scheerens et al. (2007). At school level, none of the studies analysed found associations for the factors such as 'structured instruction', 'feedback and reinforcement' and 'evaluative potential'. At classroom level, there are no associations reported for indicators of 'achievement orientation and expectations'. These factors have also hardly been conceptualized, if at all, in the TIMSS questionnaires, but they could have been part of the national versions of the questionnaires.

A relatively large number of the studies used indicators of 'curriculum quality' and 'opportunity to learn'. In the TIMSS framework, the main

Table 4. Results of the systematic review of TIMSS studies on process factors derived from the school effectiveness model of Scheerens (2007).

Process factors school level	Indicator with a significant effect (<0.05) on math and/or science achievement		
	Grade 7/8 ($n = 27$)	Grade 4 ($n = 5$)	Grade 12 ($n = 3$)
Achievement orientation and expectations	− Number of periods scheduled for teacher to counsel students		
Curriculum quality/ opportunity to learn	+ Number of topics taught + Written statement of curriculum + Amount of students tracked in top stream − Amount of students tracked in bottom stream + Extent of teachers emphasis on homework ? No streaming or tracking	+ Number of topic taught + Extent of ability grouping	
Differentiation, adaptive instruction	+ Extent of relating subjects to everyday life ± Extent of trying to solve an example related to a new topic ± Extent of teacher ask us what we know related to new topic		
School climate	+ Academic climate + Safety at school (teachers' perception) + Average perceived safety in school by students + Extent of cheating of students	+ Disciplinary climate (or lack of violence, average student perception)	− Extent of schools' report on incidents with students

(*Continued*)

Table 4 (*Continued*)

Process factors school level	Indicator with a significant effect (<0.05) on math and/or science achievement		
	Grade 7/8 (*n* = 27)	Grade 4 (*n* = 5)	Grade 12 (*n* = 3)
	+ Teachers dedication towards lessons preparation – Teachers beliefs about math – Extent of absence of students – Extent of violations of dress code – Extent of students causing injury to other students – Frequency school has to deal with class disturbance – Extent of problems with late arrival on school		
Educational Leadership	+ Extent of principal responsible for community relationships		
Consensus and cohesion among staff	+ Staff-cooperation		
	+ Extent of teachers use of professionalization activities – Extent of teachers meet regularly to discuss instructional goals/issues		
Parental involvement	+ Extent of parental involvement (all items) + Extent of direct parental involvement + Extent of school expectations of parents ensuring child's homework completion		
Curriculum quality/ opportunity to learn	+ Amount of homework	? Extent of use of inquiry in science	

(*Continued*)

Table 4 (*Continued*)

Process factors school level	Indicator with a significant effect (<0.05) on math and/or science achievement		
	Grade 7/8 (*n* = 27)	Grade 4 (*n* = 5)	Grade 12 (*n* = 3)
	+ Extent of providing practice and application opportunities + Extent of using various teaching means + Minutes of instruction per week in subject + Extent of use of inquiry in science + Extent of teacher's emphasis on understanding concepts + Extent of using calculator for different purposes − Extent of calculator use − Percentage no homework in class ? Amount of time using textbook		− Extent of passive teaching
Structured instruction	+ Extent of practicing basic math operations without calculator + Extent of whole class teaching + Extent of listening to the teacher giving instruction − Extent of students are asked to provide explanations ± Extent of looking at the textbook while the teacher talks about it ± Extent of having the teacher explain the rules and definitions		
Differentiation, adaptive instruction	+ Extent of providing opportunities for collaboration − Extent of working out problems on their own	− Extent of student oriented teaching + Grouping (segregated grouping)	+ Extent of active teaching

(*Continued*)

Table 4 (*Continued*)

Process factors school level	Indicator with a significant effect (<0.05) on math and/or science achievement		
	Grade 7/8 (*n* = 27)	Grade 4 (*n* = 5)	Grade 12 (*n* = 3)
	± Extent of teacher ask us what we know related to new topic ? Extent of working in small groups ? Extent of relating subjects to everyday life		
Feedback and reinforcement	+ Teacher's time spent on scrutiny of exams/tests + Extent of discussion on homework made		
Evaluative potential	– Extent of having quiz or test		
Class climate (creation and dimensions of)	+ Class climate + Attitude towards math (class average) + Self-confidence (class average) – Extent of disruptive students limits teaching in math class? – Extent of perceived limitations in teaching due to problem students – Extent of uncontrollable attribution (class average) ? Extent of self-pressure (class average)	+ Teachers' support perceived by students + Self-confidence (class average)	+ Self-confidence (class average)
Effective learning time	– Student self-rated attentiveness – Extent of lesson interruption		

Note: Only significant effects are included, + = positive effect in at least one country or one subject; – = negative effect in at least one country or one subject; ± = effect is not linear; and ? = contradicting (positive and negative) results for countries, subjects or TIMSS year.

focus is on these curriculum-related indicators. In line with the results of the meta-analyses of Scheerens *et al.* (2007), most studies showed positive relations of indicators between these factors and achievement.

Table 4 also shows that for class and climate factors, both composites and individual items were used. Most of the relationships found for these factors are comparable across the different studies. An orderly, safe school or classroom environment with few or no disruptive incidents seems to be positively related to achievement in different contexts. Other factors related to instruction (structured instruction, differentiated and adaptive instruction) are included in the selected studies relatively often as well. Most studies used individual items rather than composites to study these factors. It seems that data reduction by constructing composites of these items in TIMSS is more difficult compared to, for example, academic climate. The relationship of a number of these indicators proved to be contradictory in direction.

Table 5 shows the associations of a selected set of process factors (analysed in at least four countries), including variables with no association reported, by country. Comparing countries for these indicators shows that there are substantial differences between countries. For the number of topics taught (opportunity to learn), nine countries showed a positive relation and for five countries no relation with achievement was found. The differences between countries in indicators associated with achievement are even more apparent for the amount of homework given by the teacher. This indicator has been reported in the selected studies in 21 countries. The results are mixed: in 10 countries a positive relation was found and in 11 countries there was no relation with achievement. One explanation could be that in some countries (such as the Netherlands) homework in grade 4 is mostly assigned to students who are falling behind (Marte 2011).

Table 5.   Country comparison of selected set of process factors based on school and class level[a] included in systematic review TIMSS.

Indicator	Number of countries			
	No effect	Positive effect	Negative effect	Non-linear
School level				
Number of topic taught	5	9		
Academic climate	3	2		
Class level				
Amount of homework	11	10		
Attitude towards math (class average)	8	12		
Self-pressure (class average)	11	4	5	
Whole-class teaching	8	2		
Working together in small groups	7		2	1
Perceived limitations in teaching due to problem students	2		2	

[a]Included in analysis of ≥ 4 countries.

Another example of these mixed results is the class mean of self-pressure: it showed no association in 10 countries, a positive relation in four countries and a negative relation in five countries (O'Dwyer 2005). In the aforementioned study of Kaya and Rice (2010), the instructional factor 'science inquiry' (based on five classroom activities reported by grade 4 students) showed no association with science achievement in Australia, Japan and Scotland, a positive relation in Singapore and a negative relation in the USA.

Conclusions and discussion

The aim of this study is to give an overview of the contribution of TIMSS-based studies to theories of educational effectiveness by systematically analysing these studies as they have been published in scientific journals since the release of the TIMSS 1995 data. The method for this systematic review was based on that of Petticrew and Roberts (2006) and included a selection of relevant articles based on the framework of school effectiveness, and specifically process factors at school and classroom levels, as well as an assessment of the quality. The final selection (studies that were indicated as satisfactory or of high quality) was analysed further to find out to what extent the results of these studies supported the list of process factors that enhance effectiveness at school level and class level. This list is based on meta-analyses in the field of educational effectiveness research.

Characteristics and impact (research question 1)

The current situation with regard to the use of TIMSS data for secondary analyses, with student achievement as the dependent variable, is more positive compared to the situation reported by Beaton and Robitaille in 2002. After four different data collections—from 1996 until the beginning of 2010—around 200 studies used TIMSS data for secondary analyses in order to 'explain' differences in student achievement. Special TIMSS issues of journals or books (Howie and Plomp 2005, 2006, Papanastasiou and Plomp 2008) and the IEA conferences (Papanastasiou 2004) have stimulated researchers to take greater advantage of the opportunities that TIMSS has to offer, and from that perspective, they have helped IEA in achieving its initial goals of international comparative studies in education. However, the scope of secondary analyses of TIMSS data has been found to be somewhat limited. As the educational level of primary education also affects the level of secondary education, it would be a positive development if the data from TIMSS 2003, TIMSS 2007 and TIMSS 2011 were to result in more publications about malleable and other factors related to the mathematics and science achievement of grade 4 students in the coming years. Furthermore, more studies on the TIMSS data from Arab or African countries would be a worthwhile addition to educational effectiveness research as well, because the list of process factors that

enhance effectiveness of Scheerens *et al.* (2007) is based mainly on studies in western, developed countries. In addition, the limited interest in the data from Arab countries does not mirror the growing participation of these countries in TIMSS.

This study also shows that the scientific impact of TIMSS-related studies is still limited. This would mean that one of the goals of IEA (sharing the outcomes and learning from each other to increase empirical-based insights into what 'matters' in education within different contexts) is still far from being achieved. However, our study was limited to the citation index of the Web of Knowledge and we only included studies that had conducted secondary analyses with achievement as the dependent variable.

This study focused solely on the scientific impact of TIMSS, as we did not analyse the use of TIMSS in other types of publications, such as educational policy documents. An analysis of these types of documents would be interesting for future research as it could provide information about the impact of TIMSS on educational policies, curricula and the way teaching and learning are organized in schools.

Quality (research question 2)

For the second and third research questions, we limited our selection of TIMSS-related studies further by concentrating on factors in the black box: malleable school and classroom factors. The assessment of these studies revealed several quality issues. One of the recurring issues was that the level of analysis (student) was different from the level of conclusions (class or teacher level) and that causality between the levels was assumed based on student-level analyses. This assumption is problematic in two ways. First, TIMSS is a cross-sectional study and the results can only indicate a relationship and not the causal direction of the relationship. This means that one should be careful with the commonly used terms 'effect' or 'influence'. Secondly, ignoring the levels could result in wrong assumptions about relations between variables at different levels (e.g. Scheerens and Creemers 1989).

We found that in most of the recent 'high quality' studies, multilevel analyses were applied. This is a positive development compared to earlier studies, in which the specific sampling issues of TIMSS were often ignored. However, the main disadvantage of most of the multilevel statistical software is that this software only analyses direct 'effects' of variables and interaction 'effects' on the dependent variable. Until recently, it was technically impossible to build multilevel models with mediators or indirect effects (Muthén and Muthén 1999). For future research, the use of path analysis in a multilevel structure is expected to become more important.

However, this systematic review also showed that in several studies, a large number of variables (without theoretical foundation) were included in the analysis. With so many variables already in the model, adding mediators would make the interpretation of the outcomes too complex

and undermine the validity of the findings. Therefore, instead of 'a grand fishing expedition', in which as many variables as one can think of are analysed, secondary analysis would benefit more from a theoretically driven reduction of variables and items analysed in relation to achievement, but also in relation to each other. A more thematic approach including direct and indirect effects would lead to a greater in-depth understanding of how, and under what conditions, certain process factors are related to achievement.

The last main quality issue found was the treatment of the plausible values of TIMSS. Often, it was unclear if and how the authors used the five plausible values. Ideally, for each subject or subdomain, all five plausible values should be included in the analysis (von Davier et al. 2009).

Content analysis (research question 3)

Several remarks could be made about the outcomes of the selected studies in relation to the school and class process factors found in educational effectiveness research. First, empirical evidence of the importance of some of the process factors mentioned by Scheerens et al. (2007) can also be found in the selected TIMSS studies. However, there were also a number of process factors that were not addressed, or only to a limited extent, in TIMSS publications or they showed mixed results. These mixed results could be due to the limited availability of 'good' indicators and composites in TIMSS for these specific process factors. For example, in most studies the instructional characteristics of TIMSS were analysed at item level because constructing reliable composites did not seem possible. It is also likely that these mixed results are caused by the different contexts in which these relations were analysed. The country comparison showed that the importance of factors related to achievement could vary a great deal between countries. These results support the argument of Kyriakides (2006a,b) about the use of IEA data for international comparisons. He stated that researchers and policy-makers using these data should be aware that if something works in one country, it does not mean that it will also work in another country. Looking at the educational system and curriculum of high-achieving countries (such as the Asian countries) is not necessarily the best way to improve the educational level of lower scoring countries.

Finally, for a number of process factors mentioned by Scheerens et al. (2007), indicators were hardly available, if at all, in the TIMSS instruments. Indicators for three of these factors (reinforcement and feedback, educational leadership and evaluative potential) are included in the contextual framework and questionnaires of TIMSS 2011 (Mullis et al. 2009). Specifically, the evaluative potential is currently attracting growing interest from the research field of educational effectiveness (e.g. Schildkamp et al. 2009, Verhaeghe et al. 2010). The data collected by TIMSS 2011 will offer the opportunity to analyse both the direct and the indirect effects of these process factors on student achievement, controlling for student characteristics and within different educational contexts. More

in-depth studies that also analyse indirect effects, taking advantage of newly available statistical techniques, will make better use of the wide range of opportunities that the TIMSS data have to offer and will help achieve IEA's initial goals of finding factors related to effective education worldwide.

Notes

1. These achievement scores cannot be viewed as estimates of individual student scores, but rather as imputed scores of students with similar response patterns and background characteristics in the sampled population (Foy *et al.* 2008, Mislevy *et al.* 1992).
2. An overview by article of all indicators of process factors can be requested from the authors.

References

* = studies included in the content analysis of this systematic review.

*Ammermuller, A., Heijke, H. and Wossmann, L. (2005) Schooling quality in Eastern Europe: educational production during transition. *Economics of Education Review*, 24 (5), 579–599.
*Baker, D. P., Goesling, B. and Letendre, G. K. (2002) Socioeconomic status, school quality, and national economic development: a cross-national analysis of the 'Heyneman-loxley effect' on mathematics and science achievement. *Comparative Education Review*, 46(3), 291–312.
*Bankov, K., Mikova, D. and Smith, T. M. (2006) Assessing between-school variation in educational resources and mathematics and science achievement in Bulgaria. *Prospects: Quarterly Review of Comparative Education*, 36(4), 447–473.
Baumert, J., Luedtke, O., Trautwein, U. and Brunner, M. (2009) Large-scale student assessment studies measure the results of processes of knowledge acquisition: evidence in support of the distinction between intelligence and student achievement. *Educational Research Review*, 4(3), 165–176.
Beaton, A. and Robitaille, D. (2002) A look back at TIMSS: what have we learned about international studies?. In D. Robitaille and A. Beaton (eds), *Secondary Analysis of TIMSS-Data* (Dordrecht: Kluwer), 409–417.
*Bos, K. T. and Meelissen, M. (2006) Exploring the factors that influence grade 8 mathematics achievement in the Netherlands, Belgium, Flanders and Germany: results of secondary analysis on TIMSS 1995 data. In S. J. Howie and T. J. Plomp (eds), *Contexts of Learning Mathematics and Science: Lessons Learned from TIMSS* (London: Routlegde), 195–210.
*Chepete, P. (2009) *Modelling of the Factors Affecting Mathematical Achievement of Form 1 Students in Botswana Based on the 2003 Trends in International Mathematics and Science Study,* Dissertation (US: ProQuest Information and Learning).
Creemers, B. P. M. (1994) *The Effective Classroom* (London: Cassell).
Creemers, B. P. M. (2006) The importance and perspectives of international studies in educational effectiveness. *Educational Research and Evaluation*, 12(6), 499–511.
Dixon-Woods, M., Agarwal, S., Jones, D., Young, B. and Sutton, A. (2004) *Integrative Approaches to Qualitative and Quantitative Evidence*. London: NHS, Health Development Agency. Available online at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.96.8783&rep=rep1&type=pdf
*Dumay, X. and Dupriez, V. (2007) Accounting for class effect using the TIMSS 2003 eighth-grade database: net effect of group composition, net effect of class process, and joint effect. *School Effectiveness and School Improvement*, 18(4), 383–408.

Foy, P., Galia, J. and Li, I. (2008) Scaling the data from the TIMSS 2007 mathematics and science assessments. In J. F. Olson, M. O. Martin and I. V. S. Mullis (eds), *TIMSS 2007: Technical Report* (Boston, MA: IEA TIMSS & PIRLS).

Gustafsson, J. E. (2008) Effects of international comparative studies on educational quality of educational research. *European Educational Research Journal*, 7(1), 1–17.

Howie, S. J. (2002) *English Language Proficiency and Contextual Factors Influencing Mathematics Achievement of Secondary School Pupils in South Africa*, Dissertation (Enschede, the Netherlands: University of Twente).

Howie, S. (2004) A national assessment in mathematics within an international comparative assessment. *Perspectives in Education*, 22(2), 149–162.

Howie, S. (2005a) Contextual factors at the school and classroom level related to pupils' performance in mathematics in South Africa. *Educational Research and Evaluation*, 11(2), 123–140.

Howie, S. (2005b) System-level evaluation: language and other background factors affecting mathematics achievement. *Prospects: Quarterly Review of Comparative Education*, 35(2), 175–186.

*Howie, S. J. (2006) Multi-level factors affecting the performance of South African pupils in mathematics. In S. J. Howie and T. J. Plomp (eds), *Contexts of Learning Mathematics and Science: Lessons Learned from TIMSS* (London: Routlegde), 58–176.

Howie, S. and Plomp, T. (2005) TIMSS-mathematics findings from national and international perspectives: in search of explanations. *Educational Research and Evaluation*, 11(2), 101–106.

Howie, S. J. and Plomp, T. (eds) (2006) *Contexts of Learning Mathematics and Science: Lessons Learned from TIMSS* (London: Routledge).

Howie, S. and Scherman, V. (2008) The achievement gap between science classrooms and historic inequalities. *Studies in Educational Evaluation*, 34(2), 118–130.

Howie, S., Scherman, V. and Venter, E. (2008) The gap between advantaged and disadvantaged students in science achievement in South African secondary schools. *Educational Research and Evaluation*, 14(1), 29–46.

*Istrate, O., Noveanu, G. and Smith, T. M. (2006) Exploring sources of variation in Romanian science achievement. *Prospects: Quarterly Review of Comparative Education*, 36(4), 475–496.

Jurges, H., Buchel, F. and Schneider, K. (2005) The effect of central exit examinations on student achievement: quasi-experimental evidence from TIMSS Germany. *Journal of the European Economic Association*, 3(5), 1134–1155.

*Jurges, H. and Schneider, K. (2004) International differences in student achievement: an economic perspective. *German Economic Review*, 5(3), 357–380.

*Kaya, S. and Rice, D. C. (2010) Multilevel effects of student and classroom factors on elementary science achievement in five countries. *International Journal of Science Education*, 32(10), 1337–1363.

*Kunter, M. and Baumert, J. (2006) Linking TIMSS to research on learning and instruction: a re-analysis of the German TIMSS and TIMSS video data. In S. J. Howie and T. J. Plomp (eds), *Contexts of Learning Mathematics and Science: Lessons Learned from TIMSS* (London: Routlegde), 335–351.

*Kupari, P. (2006) Student and school factors affecting Finnish mathematics achievement: results from TIMSS 1999 data. In S. J. Howie and T. J. Plomp (eds), *Contexts of Learning Mathematics and Science: Lessons Learned from TIMSS* (London: Routlegde), 127–140.

Kyriakides, L. (2006a) Introduction international studies on educational effectiveness. *Educational Research and Evaluation*, 12(6), 489–497.

*Kyriakides, L. (2006b) Using international comparative studies to develop the theoretical framework of educational effectiveness research: a secondary analysis of TIMSS 1999 data. *Educational Research and Evaluation*, 12(6), 513–534.

*Lamb, S. and Fullarton, S. (2002) Classroom and school factors affecting mathematics achievement: a comparative study of Australia and the United States using TIMSS. *Australian Journal of Education*, 46(2), 154–171.

Lynn, R., Harvey, J. and Nyborg, H. (2009) Average intelligence predicts atheism rates across 137 nations. *Intelligence*, 37(1), 11–15.

*Ma, X. (2001) Stability of socio-economic gaps in mathematics and science achievement among Canadian schools. *Canadian Journal of Education*, 26 (1), 97–118. Available online at: http://www.csse.ca/CJE/Articles/FullText/CJE26-1/CJE26-1-Ma.pdf.

*Ma, X. and Papanastasiou, C. (2006) How to begin a new topic in mathematics: Does it matter to students' performance in mathematics?. *Evaluation Review*, 30(4), 451–480.

Marte, R. (2011) Who benefits from homework assignments?. *Economics of Education Review*, 30(1), 55–64. DOI: 10.1016/j.econedurev.2010.07.001.

Martin, M. O., Mullis, I. V. S., & Foy, P. (with Olson, J. F., Erberber, E., Preuschoff, C., & Galia, J.) (2008) *TIMSS 2007 International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades* (Chestnut Hill, MA: TIMSS & PIRLS International Study Center).

Martin, M. O., Mullis, I. V. S., Beaton, A. E., Gonzalez, E. J., Smith, T. A. and Kelly, D.L. (1997) *Science Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study (TIMSS)* (Boston, MA: TIMSS International Study Center).

*Meelissen, M. and Luyten, H. (2008) The Dutch gender gap in mathematics: small for achievement, substantial for beliefs and attitudes. *Studies in Educational Evaluation*, 34(2), 82–93.

*Mere, K., Reiska, P. and Smith, T. M. (2006) Impact of SES on Estonian students' science achievement across different cognitive domains. *Prospects: Quarterly Review of Comparative Education*, 36(4), 497–516.

Mislevy, R. J., Beaton, A. E., Kaplan, B. and Sheehan, K. M. (1992) Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29(2), 133–161.

Mullis, I., Martin, M. and Foy, P. (2008) *TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades* (Boston, MA: TIMSS & PIRLS International Study Centre).

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y. and Preuschoff, C. (2009) *TIMSS 2011 Assessment Frameworks* (Boston, MA: TIMSS & PIRLS International Study Centre).

Muthén, L. K. and Muthén, B. O. (1999) *Mplus: The Comprehensive Modelling Program for Applied Researchers: User's guide* (Los Angeles, CA: Author).

*O'Dwyer, L. (2005) Examining the variability of mathematics performance and its correlates using data from TIMSS '95 and TIMSS '99. *Educational Research and Evaluation: An International Journal on Theory and Practice*, 11(2), 155–178.

Papanastasiou, C. (2004) *Proceedings of the IRC—2004: IEA International Research Conference* (University of Cyprus , Department of Education).

Papanastasiou, C. and Plomp, T. (2008) Introduction to this special issue. *Educational Research and Evaluation*, 14(1), 1–7. DOI: 10.1080/13803610801896349.

*Park, C. and D. Park (2006) Factors affecting Korean student achievement in TIMSS 1999 In S. J. Howie and T. J. Plomp (eds), *Contexts of Learning Mathematics and Science. Lessons Learned from TIMSS* (London: Routlegde), 177–191.

Petticrew, M. and Roberts, M. (2006) *Systematic Reviews in the Social Sciences: A Practical Guide* (Oxford: Blackwell).

*Pong, S.-l and Pallas, A. (2001) Class size and eighth-grade math achievement in the United States and abroad. *Educational Evaluation and Policy Analysis*, 23(3), 251–273.

Postlethwaite, T. N. (1995) International empirical research in comparative education: an example of the studies of the International association for the evaluation of Educational Achievement (IEA). *Journal für Internationale Bildungsforschung*, 1 (1), 1–19. Available online at: http://213.198.68.42/zs/tc1–95/postleth.pd.

Postlethwaite, T. N. and Ross, K. N (1992). *Effective Schools in Reading: Implications for Educational Planners. An Exploratory Study.* (The Hague: IEA).

*Pugh, G. and Telhaj, S. (2008) Faith schools, social capital and academic attainment: evidence from TIMSS-r mathematics scores in Flemish secondary schools. *British Educational Research Journal*, 34(2), 235–267.

*Ramirez, M.-J. (2006a) Understanding the low mathematics achievement of Chilean students: a cross-national analysis using TIMSS data. *International Journal of Educational Research*, 45(3), 102–116.

*Ramirez, M.-J. (2006b) Factors related to mathematics achievement in Chile. In S. J. Howie and T. J. Plomp (eds), *Contexts of Learning Mathematics and Science: Lessons Learned from TIMSS* (London: Routlegde), 97–112.

Rindermann, H. (2007) The g-factor of international cognitive ability comparisons: the homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality*, 21(5), 667–706.

Robitaille, D. F. and Beaton, A. E. (eds) (2002) S*econdary Analysis of the TIMSS Data*. (Dordrecht: Kluwer Academic Publishers).

Rodgers, M., Sowden, A., Petticrew, M., Arai, L., Roberts, H., Britten, N. and Popay, J. (2009) Testing methodological guidance on the conduct of narrative synthesis in systematic reviews. *Evaluation*, 15(1), 49–73.

*Rodriguez, M. C. (2004) The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education*, 17(1), 1–24.

Rushton, J. P. and Templer, D. I. (2009) National differences in intelligence, crime, income, and skin colour. *Intelligence*, 37(4), 341–346.

Scheerens, J. (2008). *Een overzichtsstudie naar school- en instructie-effectiviteit* [Review of research on school and instructional effectiveness]. (Enschede: University of Twente, Department of Educational Organisation and Management).

Scheerens, J. and Creemers, B. P. M. (1989) Conceptualizing school effectiveness. *International Journal of Educational Research*, 13(7), 691–706.

Scheerens, J., Luyten, H., Steen, R. and Luyten-de Thouars, Y. (2007) *Review and Meta-analyses of School and Teaching Effectiveness*. (Enschede: University of Twente, Department of Educational Organisation and Management).

Schildkamp, K., Visscher, A. and Luyten, H. (2009) The effects of the use of a school self-evaluation instrument. *School Effectiveness and School Improvement*, 20 (1), 69–88.

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

*Schreiber, J. B. and Chambers, E. A. (2003) American high school seniors' mathematics literacy achievement. *North American Journal of Psychology*, 5 (1), 15–30. Retrieved from EBSCOhost Psychology and Behavioural Sciences Collection.

*Tarr, J. E., Mittag, K. C., Uekawa, K. and Lennex, L. (2000) A comparison of calculator use in eighth-grade mathematics classrooms in the United States, Japan, and Portugal: results from the Third International Mathematics and Science study. *School Science and Mathematics: Official Journal of the School Science and Mathematics Association*, 100(3), 139–150.

Van den Broeck, A., Opdenakker, M. and Van Damme, J. (2006) The effects of students, class and school characteristics on TIMSS 1999 mathematics achievement in Flanders In S. J. Howie and T. J. Plomp (eds), *Contexts of Learning Mathematics and Science: Lessons Learned from TIMSS* (London: Routlegde), 83–96.

*Van den Broeck, A., Opdenakker, M.-C. and 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.

Verhaeghe, G., Vanhoof, J., Valcke, M. and Van Petegem, P. (2010) Using school performance feedback: perceptions of primary school principals. *School Effectiveness and School Improvement*, 21(2), 167–188.

Von Davier, M., Gonzalez, E., and Mislevy, R. (2009) Plausible values: what are they and why do we need them? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 2, 9–36.

*Wilkins, J. L. M., Zembylas, M. and Travers, K. J. (2002) Investigating correlates of mathematics and science literacy in the final year of secondary school. In D. Robitaille and A. Beaton (eds), *Secondary Analysis of the TIMSS Data* (Dordrecht: Kluwer Academic), 291–316.

Xin, T., Xu, Z. and Tatsuoka, K. (2004) Linkage between teacher quality, student achievement, and cognitive skills: a rule-space model. *Studies in Educational Evaluation*, 30(4), 205–223.

*Zhao, H. and Akiba, M. (2009) School expectations for parental involvement and student mathematics achievement: a comparative study of middle schools in the US and South Korea. *Compare*, 39(3), 411–428.

*Zuzovsky, R. (2008a) Capturing the dynamics behind the narrowing achievement gap between Hebrew-speaking and Arabic-speaking schools in Israel: findings from TIMSS 1999 and 2003. *Educational Research and Evaluation*, 14(1), 47–71.

*Zuzovsky, R. (2008b) Closing achievement gaps between Hebrew-speaking and Arabic-speaking students in Israel: Findings from TIMSS 2003. *Studies in Educational Evaluation*, 34(2), 105–117.