Chapter 30
Methods for Studying Mathematics Teaching and Learning Internationally

Mogens Niss, Jonas Emanuelsson, and Peter Nyström

Abstract The focus of this chapter is issues related to methods for studying mathematics teaching and learning internationally. The chapter identifies three sorts of overarching purposes and goals of international studies, namely to uncover and analyze, across a group of countries: differences in students' learning outcomes, achievements and attitudes; differences in curricula, teaching approaches, resources and the environments of mathematics education; and possible links between the latter and the former. The chapter provides detailed accounts of the designs, methods, methodologies, and instruments that have been used in two kinds of studies—large-scale international comparative studies, such as TIMSS and PISA, and so-called focal studies concentrating on more specific *problématiques* or themes. The last part of the chapter offers reflections on the nature of international comparative studies with an emphasis on their strengths and potentials as well as on their challenges and limitations. One fundamental question in this context is the extent to which the results of such studies can be meaningfully interpreted, especially in view of the massive interest amongst politicians, administrators, media, and the general public, who often do not pay sufficient attention to the characteristics and conditions of the studies.

M. Niss (✉)
Roskilde University, Roskilde, Denmark
e-mail: mn@ruc.dk

J. Emanuelsson
University of Gothenburg, Gothenburg, Sweden

P. Nyström
Umeå University, Umeå, Sweden

M. A. (Ken) Clements et al. (Eds.), *Third International Handbook of Mathematics Education*, Springer International Handbooks of Education 27, DOI 10.1007/978-1-4614-4684-2_30,
© Springer Science+Business Media New York 2013

Introduction: The Relationship Between Study Issues and Methodology

Since the creation of the International Commission on Mathematical Instruction (ICMI) in 1908 (Schubring, 2008), there has been an interest in considering mathematics teaching and learning from an international perspective. Until the 1960s, the focus was on describing and comparing mathematics curricula across different countries, or on proposing—from normative points of view—new curriculum approaches or components (such as the notion of function in the early decades of the 20th century or the so-called new math or modern mathematics movement from the mid-1950s to the mid-1970s). When the international congresses on mathematical education (the ICMEs) came into being (the first one was held in Lyon, France, in 1969), the majority of the contributions in the early ICMEs were designed to exchange information, views, and experiences amongst delegates from different countries about the actual or potential structures of mathematics curricula, the orchestration of teaching, teaching materials and resources, teaching experiments, and—to a lesser extent—student reactions to the "diets" they were offered.

Even though it dates back to the beginning of the 20th century, the sharing of information, ideas, and experiences has never ceased to be of interest. For example, the so-called International Seminar at the Park City Mathematics Institute (PCMI), held under the auspices of the Princeton Institute for Advanced Study every summer in Park City, Utah, USA, has provided a platform for such exchange since 2001.

The goal of all these endeavours has been to allow participants to learn from each other in terms of ideas, approaches, materials for teaching, and the reported outcomes thereof. Even though selecting, collecting, and presenting the factual information involved in these activities may well have been difficult and time consuming in places, it would not be reasonable to say that these endeavours amount to *studying* mathematics teaching and learning internationally in a scholarly or scientific sense. Studying something is closely linked to trying to come to grips with essential features of or issues related to the objects, situations, or systems to be studied; in other words, seeking answers to pertinent questions by way of some investigation, a disciplined inquiry. Studying something is usually focussed on uncovering and explaining relationships, with particular regard to mechanisms, correlations, and causalities. Therefore, any discussion of the choice and implementation of the methods to be put to use in a study must take its point of departure in the issues and questions that the investigation is designed to address. So, what are the issues addressed and the questions asked in studying mathematics teaching and learning internationally? And what are individuals' and agencies' (or even countries') purposes of engaging in such studies? This is related to the question asked by Clarke (2003) with regard to *comparative research*: "Who are the stakeholders of international comparative research?" (p. 151).

In the sections that follow, we provide more specific and detailed answers to these questions as far as the most important international studies are concerned, of which the first seems to be the so-called FIMS—First International Mathematics Study—which was carried out in 1964 (see below). However, at an overall level it is

fair to claim that most international studies are designed to deal with three major *problématiques*: The first is to uncover and analyze *differences in students' learning outcomes, achievement, and attitudes* across a group of countries. The second is to uncover and analyze *differences in curricula, teaching approaches, resources for teaching, classroom cultures, teachers' educational and other backgrounds, and more general cultural and socio-economic environments of mathematics education*. The third, and often the most significant, is to *link* the former *problématique* to the latter; in particular, in order to come to understand, if possible, the former as a function of the latter. It goes without saying that the methodological deliberations and issues arising in this context (should) depend heavily on the quantitative and qualitative characteristics of the students considered; on the specific learning outcomes, kinds of achievement, and sorts of attitude in focus; on the cultural, societal, economic, and institutional conditions of the countries involved; and on those aspects of teaching approaches and resources, classroom cultures, and teacher backgrounds that are selected to be of interest in the investigation. Clarke (2003) adds a twist to the third *problématique*; namely, what he calls "evaluative comparisons: not just to document similarities and differences, but attaching value to performances judged as superior by some criterion" (p. 152).

Against this background, one may well raise the more general question of the extent to which it makes sense, and is methodologically feasible, to detect, investigate, and interpret differences and to make comparisons across and among countries with particular regard to mathematics education, when multitudes of cultural, societal, and economic and other factors exert predominant influences on the systems in which mathematics education takes place. We return to this issue later in this chapter.

In dealing with issues concerning study methods, a number of words almost automatically enter the stage: *design, method, methodology, instrument, technique,* and *procedure,* among others. Transparency in deliberations and exposition requires some clarification of what these terms are supposed to mean. If we take our point of departure in the idea that scholarly and scientific studies are undertaken in order to answer certain more or less clearly delineated questions (Niss, 2010), we propose the following definitions in the present context.

By the term *design* of a study, we understand the entire *collection of approaches* (whether conceptual, theoretical, or empirical) employed *to provide answers* to the set of questions that drive the study; in other words, the overall *layout* of the study. Each approach is focussed on answering a subset of the questions (but several approaches may be used, e.g., in combination, to answer the same question) and hence gives rise to issues of *methodology*. By *methodology,* we understand the set of deliberations, reflections, and analyses involved in choosing, implementing, and assessing one or more *methods* with a potential to answer a certain class of questions. Typically this involves comparing, contrasting, and relating different actual and potential methods with particular regard to their potentialities, limitations, and tractability in the given context and under the circumstances present. So, we use the term *method* to designate a package of specific undertakings by which a certain class of questions may be answered, and the term *methodology* to include all meta-level considerations about methods. Adopting a particular method as a means for

answering certain questions presupposes the belief that the method actually can, or at least has the potential to, provide valid answers to the questions. A method may be established and well-described, but it may also be in a process of inception or under construction for a certain purpose. Implementing a method normally involves putting a number of *instruments* to use. Typically an instrument—say, a questionnaire—is not restricted to be part of a particular method but will be available for use in several different methods. Finally, using an instrument often requires the activation of various more or less specific *techniques*, some of which may take the form of standardized *procedures*, whereas others may be more loosely defined. In the following sections, these rather general definitions are given flesh and blood when we deal with concrete studies.

This chapter is structured as follows: In the next two sections, we attempt to provide factual presentations, without much commentary, of the studies under consideration in the chapter, including their goals, designs, and methods. In the last section, we offer our more analytic reflections on key issues related to those and other studies.

Different Kinds of Studies and Their Goals

In gross terms we deal with two kinds of internationally-oriented studies of mathematics teaching and learning. The first kind consists of *large-scale international comparative studies*, where the term *large-scale* refers to at least two features—the involvement of a multitude of countries and of large numbers of students. Sometimes *large-scale* also means "many dimensions," such as student achievement and affect, socio-economic background variables, structure of education systems, curriculum organization, approaches to teaching, and teacher backgrounds. Studies of the second kind, let us agree to call them *focal studies*, have a narrower focus—for example, problem solving, curriculum structure, textbooks, classroom interaction—and typically involve just a few countries. Large-scale studies—which almost by definition require huge efforts and human and material resources, including funding, and are time consuming—tend to attract a lot of public interest and debate, especially if league tables are included in the reporting, whereas focal studies rather attract the attention of mathematics educators and researchers, and occasionally of politicians dealing with education.

Large-Scale Studies

We begin by listing the international large-scale studies that are taken into consideration in this chapter. Because of the resources required to undertake large-scale studies, there are not so many of them. Although comparative international studies of education at large have a long history (Kaiser, 1999a), as previously mentioned

the first large-scale comparative international study of *mathematics* was the FIMS. It was produced and published by the IEA, the International Association for the Evaluation of Educational Achievement, which was created by a group of educationists in 1958 and established as a legal entity based in the Netherlands in 1967. The study was designed and conducted during the years 1961–1964, and students' achievements in mathematics in 12 countries were tested in 1964 (Freudenthal, 1975). The outcomes were reported in 1967 (Husén, 1967). Freudenthal (1975) made the following comments on the aims of FIMS:

> The overall aim is, with the aid of psychometric techniques, to compare outcomes in different educational *systems*. The fact that these comparisons are cross-national should not be taken as an indication that the primary interest was, for instance, national means and dispersions in school achievement at certain age and school levels. …
>
> The main objective of the study is to investigate the "outcomes" of various school systems by relating as many as possible of the relevant input variables (to the extent that they could be assessed) to the output assessed by international test instruments. (p. 131)

Two populations of students took part in the study, one consisting of 13-year-olds, and one consisting of students at the final year of upper secondary school.

It is worth noticing in the above quotation that the ultimate goal of FIMS was to compare different educational systems and that students' achievements in mathematics were used as *the* indicator of the outcomes of these different systems.

The next comparative IEA study, SIMS, the Second International Mathematics Study, was decided upon in 1976 (Travers & Weinzweig, 1999), and data were collected during 1980–1982 (Robitaille & Travers, 1992). The final reports were published some years later (Robitaille & Garden, 1989; Travers & Westbury, 1990). SIMS was considerably more complex than FIMS. First and foremost, the goal was broader: "The overall objective was to produce an international portrait of mathematics education, with a particular emphasis on the mathematics classroom" (Travers & Weinzweig, 1999). More specifically, the emphasis was on an in-depth study of the curriculum:

> The curriculum in many countries is mandated at the national or system level. This is spelled out in curriculum guides and presented in the approved textbooks. Teachers are then expected to translate these guides into actual classroom instruction. There is an implicit assumption that students will learn the material presented in the classroom. How well do teachers translate what has been mandated? How close a match is there between what actually goes on in the classroom and what has been mandated? How much and what do the students learn? (p. 20)

Thus the focus of this study was on mathematics education as an end in itself, not as a means to a different end as was the case with FIMS. Based on the intentions indicated in the quotation, SIMS introduced a distinction which since then has become standard in mathematics education: the distinction between the *intended* curriculum, the *implemented* curriculum and the *attained* curriculum (a curriculum-oriented version of Bauersfeld's (1979) older distinction between the matter "meant," the matter "taught," and the matter "learned"). The student populations targeted in the study were roughly the same as the ones in FIMS; namely, 13-year-olds and those students at the final year of upper secondary school whose program had mathematics

as a substantial component. Seventeen countries took part in SIMS, and also the Canadian provinces Ontario and British Columbia. Of the 17 countries, the French- and Flemish-speaking parts of Belgium entered the study as separate entities.

TIMSS, The Third International Mathematics and Science Study, conducted in 1995 under the auspices of the IEA, represented further growth of scale and complexity in comparison with SIMS. The focus on the intended, the implemented, and the attained curriculum and the relationships between them was maintained in TIMSS. Beaton and Robitaille (1999) listed four "research questions" that underlay the study design. First, as to the intended curriculum, the question concerns the ways in which countries vary in the intended learning goals for mathematics and how these goals are influenced by the characteristics of the educational systems, the schools and the students, the ways in which the curriculum is articulated, and the locus of curricular decision-making. Next, when it comes to the implemented curriculum, the question concerns (possible) differences between the implemented and the intended curriculum and the multitude of factors that may be responsible for observed differences. Factors that influence the attained curriculum form the concern of the third question, including students' homework, investment of effort, classroom behaviour, attitudes and aspirations with regard to education, and self-concept, as well as parents' economic status and expectations for their children. The fourth and final question addresses the relationships between the three curriculum aspects and the social and educational contexts, including "arrangements for teaching and learning, and outcomes of the educational process" (p. 34).

The student populations addressed in TIMSS were three, roughly comprising 9-year-olds, who were not included in FIMS or SIMS, 13-year-olds, and the students in the final year of upper secondary schooling. Forty-five countries took part in the study with at least one of these three populations. A huge body of reports were published about TIMSS in the late 1990s (c.f., http://timss.bc.edu), including one on mathematics achievement in the primary school years (1997), one on mathematics achievement in the middle-school years (1996) and one on mathematics and science achievement in the final year of secondary schooling (1998), in addition to various survey and technical reports (e.g. Martin & Kelly, 1996; Martin, Gregory & Stemler, 2000; and Martin, Mullis & Christowsky, 2004). Moreover, three so-called TIMSS monographs on curriculum frameworks for mathematics and science, research questions and study design, and textbooks, respectively, were published as well.

A follow up on TIMSS, called TIMSS-Repeat (TIMSS-R), was conducted in 1999. It focussed on the 13-year-olds only (Population 2 in TIMSS), but slightly changed the definition of the group. The four general research questions posed in TIMSS (1995) were also in focus in TIMSS-R: What kinds of mathematics and science are students expected to learn? Who provides the instruction? How is instruction organized? What have students learned?

Since then, taking advantage of the fact that the acronym TIMSS has become a brand in itself, IEA decided, rather than to insert still new first letters, to change the acronym to Trends in International Mathematics and Science Study, with the year in which it was conducted added to the acronym. Under that heading, subsequent studies were conducted in 2003, 2007, 2008, and 2011. Accordingly, previous studies were renamed to TIMSS 1995 and TIMSS 1999. The change from *third* to *trends* also reflects a new focus on trends in the IEA studies. The definition of

TIMSS target populations (Populations 1–3) has developed from a focus on age to a focus on grade level. By attempting to compare students' achievements after the same amount of schooling, the researchers assume the results will be directly useful for educational purposes.

In 1964 FIMS targeted not only compulsory schooling but also post-compulsory secondary education. As previously described, TIMSS 1995 contained such an element as well, and around 2005 initiatives were taken to establish a study enabling comparison with upper secondary school results from 1995. These initiatives led to TIMSS Advanced 2008, aimed at assessing the advanced mathematics (and physics) achievement of students in the final year of secondary schooling, which in most countries is the 12th year (Garden et al., 2006). For advanced mathematics, the target population was defined as those students in the final year of secondary schooling who have taken courses in advanced mathematics.

During the writing of this chapter, TIMSS 2011 was well under way. This study aimed at Populations 1 and 2 with similar definitions to those found in TIMSS 2007. A unique characteristic of this TIMSS cycle is that the IEA study PIRLS (Progress in International Reading Literacy Study) was done simultaneously in Grade 4. This created opportunities for research aiming at investigating and understanding relationships between language and mathematics.

TIMSS always took its point of departure in student achievement vis-à-vis school curricula. In contrast, the Organisation for Economic Co-operation and Development (OECD) decided in the late 1990s to mount a series of international comparative studies that focussed on the outcomes of schooling for students leaving compulsory education in most countries, settling on students of age 15, irrespective of the curricula according to which they have been taught. The purpose was to study education systems' ability to equip the youth in the participating countries with the capabilities needed for citizenship in a broad sense, but with particular regard to reading, mathematics, and science. This undertaking was given the name Programme for International Student Assessment, better known as PISA (for an in-depth comparison between TIMSS and PISA, see de Lange, 2007). The first study was to take place in 2000, and then every three years a new study would be conducted. The introduction to the initiating publication of PISA, *Measuring Student Knowledge and Skills: A New Framework for Assessment* (OECD, 1999) reads:

> How well are young adults prepared to meet the challenges of the future? Are they able to analyse, reason and communicate their ideas effectively? Do they have the capacity to continue learning throughout life? Parents, students, the public and those who run education systems need to know. …
>
> OECD/PISA will produce policy-oriented and internationally comparable indicators of student achievement on a regular and timely basis. The assessments will focus on 15-year-olds, and the indicators are designed to contribute an understanding of the extent to which education systems in participating countries are preparing their students to become lifelong learners and to play constructive roles as citizens in society. (p. 9)

Furthermore,

> PISA is the most comprehensive and rigorous international effort to date to assess student performance and to collect data on the student, family and institutional factors that can help to explain differences in performance. (p. 14)

The international consortium chosen by the OECD to be in charge of conducting the study was the Australian Council for Educational Research (ACER). It was decided to adopt a cyclical study structure, such that for each round—cycle-one of the three domains reading, mathematics, and science would be the major domain, and the other two would be minor domains. Thus, reading was the major domain in 2000, mathematics in 2003, science in 2006, reading again in 2009, and so on. Mathematics will be the major domain again in 2012.

The fact that the purpose of PISA is to uncover the capabilities for citizenship and lifelong learning that students gain from schooling in different countries, implies that the focus of the study is, and has been from the very beginning, expressed in terms of *literacy*, including mathematical literacy. The first definition of *mathematical literacy* was as follows:

> Mathematical literacy is an individual's capacity to identify and understand the role that mathematics plays in the world, to make well-founded mathematical judgments and to engage in mathematics in ways that meet the needs of that individual's current and future life as a constructive, concerned and reflective citizen. (OECD, 1999, p. 43)

Very minor changes were made to this definition in the frameworks for PISA 2003, 2006, and 2009. However, as a result of changes in the composition and management of PISA instigated by the OECD in 2009, the U.S. organization Achieve became associated with the consortium with the specific task to oversee the development of a new framework for PISA mathematics in 2012. As part of this process, a new definition of mathematical literacy was agreed upon. Its purpose was to spell out, in an explicit way, the main components involved in identifying and understanding the role of mathematics and in engaging with it:

> *Mathematical literacy* is an individual's capacity to formulate, employ, and interpret mathematics in a variety of contexts. It includes reasoning mathematically and using mathematical concepts, procedures, facts, and tools to describe, explain, and predict phenomena. It assists individuals to recognise the role that mathematics plays in the world and to make the well-founded judgments and decisions needed by constructive, engaged and reflective citizens. (OECD, 2010b)

In 2000 (OECD, 2001), 32 countries participated in PISA, including 28 OECD countries. In 2002, another 13 countries joined the first cycle. In the 2003 round, in which mathematics was the major domain, 30 OECD countries and 11 non-OECD countries participated (OECD, 2004). In 2006, the 30 OECD countries were joined by 27 other countries or "economies" (OECD, 2007), whereas 34 OECD countries and 31 other countries or "economies" took part in PISA 2009 (OECD, 2010a). In addition to the outcomes reports just referenced, OECD PISA has published hosts of other reports, some of which are technical reports, whereas others focus on specific themes or issues (see http://www.pisa.oecd.org).

Focal Studies

When it comes to what we here call international focal studies, there are quite a few of them. Some are accompanying or following up on large-scale studies,

whereas others are independent studies. A study of the former kind is the so-called Survey of Mathematics and Science Opportunities (SMSO), a four-year study on instructional practices in six countries (France, Japan, Norway, Spain, Switzerland, and the USA), "charged with developing the research instruments and procedures that would be used in the Third International Mathematics and Science Study (TIMSS)" (Cogan & Schmidt, 1999, p. 69) with particular regard to 9- and 13-year-old students. Although SMSO was conducted prior to TIMSS itself, the so-called TIMSS Video Study of eighth-grade classrooms in Germany, Japan, and the USA, and the so-called Case Study Project of TIMSS concerning the same three countries, were supplementary additions to TIMSS proper, even though they were funded by the US Department of Education, National Center for Education Statistics (Kawanaka, Stigler, & Hiebert, 1999; Stevenson, 1999). Germany and Japan were chosen because they were, at the time, seen as major economic competitors with the USA, and because Japan was consistently obtaining scores at the top end of international comparison tests (Kawanaka et al., 1999; Stevenson, 1999). Another related study (Schmidt et al., 1997) surveyed the curricular intentions in school mathematics in a number of countries.

One driving force behind the development of the TIMSS Video Study was the ambition to go beyond international comparisons of students' achievements as measured by tests. IEA wanted also to consider so-called contextual factors (Stigler, Gallimore, & Hiebert, 2000). Previously, information on teaching processes had relied solely on the responses of teachers and students to questionnaires.

The overall goal of the Video Study was to provide a rich account of what happens inside Grade 8 classrooms in the three countries, and in that context:

> To develop objective observational measures of classroom instruction to serve as quantitative indicators at a national level of teaching practices in the three countries.
>
> To compare actual mathematics teaching methods in the US and the other countries with those recommended in current reform documents and with teachers' perceptions of those documents.
>
> To assess the feasibility of applying videotape methodology in future wider-scale national and international surveys of classroom instructional practices. (Kawanaka et al., 1999, p. 87)

The Video Study was later extended to include eight countries in the TIMSS-R video survey study.

The Case Study Project was included in TIMSS "in the hope that [the findings] would provide in-depth information about beliefs, attitudes and practices of students, parents and teachers that would complement and amplify information obtained through the questionnaires used in the main TIMSS study" (Stevenson, 1999, p. 106). The research topics chosen were meant to "be of interest to US policymakers who deal with elementary and secondary schooling" (p. 107), and comprised "national standards, teachers' training and working conditions, attitudes towards dealing with differences in ability and the place of school in adolescents' lives" (p. 107).

So, the common task of the Video Study and the Case Study of TIMSS was to zoom in on factors in Germany, Japan, and the USA that might potentially serve to explain the differences in outcomes of mathematics (and science) education, including students' achievements, in these countries.

In the beginning of 2000, the Learner's Perspective Study (LPS) was launched. Initially research groups from four countries—Australia, Germany, Japan and the USA—participated. The study was mainly funded by Australian means (Clarke, Keitel, & Shimizu, 2006). There were different rationales behind the original study. One of the more important ambitions was to be able to situate Australian mathematics teaching in relation to results from the first TIMSS video survey study (Stigler & Hiebert, 1999). Later the study was extended by research groups from several additional countries joining the project. At the time of writing this chapter the number of participating groups amounts to 15 (see the Web site of the project http://www.lps.iccr.edu.au). As a result, the original project has gradually been expanded and can today rather be seen as a network of researchers with a common interest in classrooms studies in an international context.

A broad range of research questions are addressed within the LPS. Since the project is a conglomerate of research groups belonging to different traditions, there is no unifying set of questions. Clarke, Keitel, & Shimizu (2006) put forward a set of seven overarching questions ranging from addressing issues of the presence of coherent and culturally-specific student and teacher practices, over relationships between these practices, to variability within classrooms and countries as well as among classrooms and countries. The questions also reflect ambitions of the project to provide information about the practices studied.

It is also worth mentioning that in comparison with the large-scale international studies described in this chapter, the LPS stands out by not being anchored in an international organization such as IEA, OECD, or ICMI. Instead, it is based on researcher-driven interests. Hence, LPS is an example of scholarly stakeholders working in the field of international comparative studies.

The US–Japan Cross-cultural Research on Students' Problem-Solving Behaviours is an early example of another independent focal study with the researchers themselves as the stakeholders, emphasizing problem solving. It began by joint US–Japan seminars in 1987 instigated by Jerry Becker and T. Miwa, and was subsequently developed into a research project, the purpose of which was "to collect descriptive data pertaining to the performance of Japanese and US students on certain kinds of problem-solving behaviours," and "contrasts in these behaviours between students in the two countries were also sought" (Becker, Sawada, & Shimizu, 1999, p. 121). The students under consideration were 4th, 6th, 8th, and 11th graders in the two countries.

A comparative study—called the Kassel Project—of secondary mathematics teaching in England and Germany was carried out in the 1990s. One of the rationales stated for this study (Kaiser, 1999b) was that European countries will, to an increasing extent, receive each others' students. Therefore it will be important to know what students know and to develop a mutual understanding of the different education systems in the European countries. The goals were to provide

> an examination of the differences in the mathematical achievement of English and German students.
>
> an analysis of the differences in the ways of teaching and learning mathematics in both countries. Based on this, the teaching methods will be questioned, and ideas gathered on how to improve the different ways of teaching mathematics. (p. 141)

An entirely different kind of comparative study is found in the 13th ICMI Study *Mathematics Education in Different Cultural Traditions: A Comparative Study of East Asia and the West* (Leung, Graf, and Lopez-Real, 2006). In this study, which is actually a collection of different theoretical and empirical contributions, numerous aspects of observed differences between the Confucian tradition and approach to mathematics education, which is predominant in East Asia, and the Western traditions are investigated. In contradistinction to what is common to several other international comparative studies, where the overall idea is, in some way or another, to provide lessons for learning from each other, the 13th ICMI study had a different, if not outright opposite, rationale:

> The globalisation processes are producing reactions from mathematics educators in many countries who are concerned that regional and local differences in educational approach are being eradicated. This is not just a mathematical ecology argument, about being concerned that the rich global environment of mathematical practices is becoming quickly impoverished. It is also an argument about education, which recognises the crucial significance of any society's cultural and religious values, socio-historical background and goals for the future, in determining the character of that society's mathematics education. (p. 6).

In other words, this study can be seen as an attempt to counteract (Western) cultural and educational imperialism with regard to mathematics education. It did so by comparing and contrasting the contexts of mathematics education, the curricula, teaching and learning and, finally, values and beliefs in Confucian and Western cultures and traditions.

Several other focal studies might have been mentioned, for example, Collaborative Studies on Innovations for Teaching and Learning Mathematics in Different Cultures in APEC Member Economies (cf. http://www.criced.tsukuba.ac.jp/math/apec and http://www.crmekku.ac.th), but they would not fundamentally expand the set of purposes already encountered in the international studies mentioned.

Designs and Methods Adopted in International Studies

Based on the distinctions introduced in the first section, we concentrate here on presenting and discussing the *designs* (i.e., the set of approaches adopted to answer the questions that drive a given study) and the *methods* chosen and implemented for pursuing these approaches. Moreover, we consider the most important *instruments* involved in these methods.

The IEA Studies

The *design* adopted for FIMS consisted of three approaches to answering the question driving the study (Robitaille & Travers, 1992). As the fundamental idea in FIMS was to measure and compare outcomes of education systems by way of student achievement in mathematics, the overarching and most important approach

was to *construct achievement tests*. This was closely linked to the second approach, *choosing the student populations* in participating countries whose achievements were to represent countries' school achievements at large. That constituted the second approach. The third approach to answering the primary question was to *ask students, parents, and teachers* about attitudes, demographics, socio-economic backgrounds, and so on.

Considerations about which *student populations* to involve in FIMS led to the definition of three student populations to be tested, but results were reported for only two of these: A younger population, consisting of students close to the very end of compulsory schooling in most countries (Postlethwaite, 1971), roughly speaking consisting of 13-year-olds, and an older population, consisting of students at the end of secondary schooling. Both populations were divided into two subpopulations, but the details are omitted here. Methods for identifying samples of these populations in the participating countries were employed nationally according to general guidelines, which included stratified random probability sampling.

As to the *achievement tests*, the method adopted was to construct them in accordance with a matrix structure: "topics" by "cognitive behaviour levels." Although the topics varied across the populations, the five cognitive behaviour levels were the same for all populations (Husén, 1967): (a) knowledge and information: recall of definitions, notation, concepts; (b) techniques and skills: solutions; (c) translation of data into symbols or schema and vice versa; (d) comprehension: capacity to analyze problems, to follow reasoning; and (e) inventiveness: reasoning creatively in mathematics. The sets of test items constructed with this matrix structure in mind were then administered to students in all participating countries after having been filtered through elaborate piloting procedures. More specifically, each student in a given population was required to do the same three-to-four one-hour item booklets—forming the test *instruments*—such that each student had to complete a total of 50 to 70 items (Postlethwaite, 1971; Robitaille & Travers, 1992). Most of the items had a multiple response format, but a couple of open-ended items were included in each booklet. Included in the item booklets were also some scale-based questions concerning student attitudes to mathematics and its learning (Postlethwaite, 1971). More specifically, these questions concerned "mathematics as a process," "difficulties of learning mathematics," "the place of mathematics in society," "school and school learning," and "man and his environment."

Finally, the method to probe into institutional characteristics, socio-economic background variables, career perspectives, teacher backgrounds, and so on, was to make use of four types of questionnaires—each forming a sociological *instrument*—student questionnaires, teacher questionnaires, school questionnaires, and a national case study questionnaire.

Given its focus on portraying mathematics education at large, and curricula in particular, SIMS had a somewhat different *design*, which was based on an overall framework distinguishing between the intended, the implemented, and the attained curriculum. This framework gave rise to three *different approaches* to answering questions concerning the constitution of each type of curriculum across participating countries. However, the basic—and more overarching—approach was to decide on the student populations whose curricula were to be investigated in the study.

Again, as part of the design of SIMS *target populations* had to be chosen. It seems as though the basic approach leading to the selection of these populations was to keep the definitions of FIMS, whenever possible, but also to attempt to solve some of the delineation problems encountered with FIMS, especially with students in the older population. In most countries, the actual samples of students representing each population studied were selected by using probabilistic sampling methods at a national level.

As to the method adopted in the identification of *the intended curriculum* in participating countries, a matrix-based specification in terms of a content dimension and a cognitive behaviour dimension, similar to but not identical with that employed in FIMS, was chosen (Travers & Weinzweig, 1999). Subdivided content strands were identified for the two populations (five for the younger and nine for the older population). As regards the cognitive behaviour dimension, SIMS deviated from FIMS in making use of a more hierarchical classification: computation, comprehension, application, and analysis. Considerable effort was made to avoid ambiguity, for example, by describing the resulting cells in the matrices by detailed examples of what the SIMS committee had in mind such that countries' respondents were able to tell whether a certain cell was part of their curriculum or not. Moreover, countries' respondents were asked to indicate the degree of importance of each cell for the curriculum at issue in their country. In other words, the instruments employed in this method were content-by-cognitive behaviour grids, together with illustrations and comments, which country respondents were asked to fill out and return accompanied by importance degrees assigned to each cell.

When it came to investigating *the implemented curriculum* in the SIMS countries, that is, the second approach in the design of the study, the method employed was to ask teachers to fill in detailed questionnaires—the instruments—about their classrooms, their teaching methods during the school year, their attitudes and beliefs, and the place and role of each cell in the above-mentioned grids. For "each topic, a detailed description of a large variety of teaching methods that could be utilized in the teaching of that topic" was provided (Travers & Weinzweig, 1999, p. 22).

Finally, the core approach in the design was to capture *the attained curriculum* in participating countries. As in FIMS, the method to investigate this curriculum first of all consisted in written student achievement tests containing items referring to the content-by-cognitive behaviour grid mentioned above. The number of items belonging to each cell was determined by the importance assigned to that cell by participating countries. The final pool of items also contained some anchor items in order to detect possible changes for the 11 countries that participated in both FIMS and SIMS. The actual instrument employed consisted of multiple item booklets, such that each student answered one or two booklets, at least one of which was from a set of rotated booklets. This rotation was introduced in order to ensure a broad coverage of grid cells across countries (Travers & Weinzweig, 1999). Moreover, the instrument also included, for each item, a student and a teacher question, asking whether the content implicated in the item had been taught or not, and if so when.

The *design* of the Third International Mathematics and Science Test (TIMSS) was a continuation of that of SIMS. For TIMSS, the design was focussed on answering what Beaton and Robitaille (1999) called Research Questions 1–4, using the

three-part model of intended, implemented, and achieved curricula. Methods used to describe and evaluate the different curriculum levels were similar to those of SIMS, but there were also some differences. In TIMSS 1995, a set of performance items was used as a supplement to the core paper-and-pencil tests given to students. Furthermore, the construction of the tests was based on a framework specifying three dimensions in a mathematics curriculum: content, performance expectations, and perspectives (Robitaille et al., 1993). The content dimension listed the mathematical content areas to be covered, performance expectations defined competencies such as knowing and communicating, and perspectives covered other aspects such as attitudes and habits of mind. The target populations in TIMSS 1995 were similar, though not identical, to those of FIMS and SIMS: Population 1 (9-year-olds), Population 2 (13-year-olds), and Population 3 (students in their final year of secondary schooling). All participating countries were required to enter Population 2, whereas the other two were optional. A two-stage random-sampling procedure was used as the method for identifying samples representing the sample populations in each participating country. In Populations 1 and 2, entire classrooms were sampled, whereas in Population 3, individual students were selected.

TIMSS 1999 is often described as a repetition of TIMSS 1995, using basically the same *design*. The framework for constructing tests in TIMSS 1999 was the same as for TIMSS 1995. Thus the mathematical content covered was the same. The goal with TIMSS 1999 was "more modest in scope, focussing on one target population only." Nevertheless, it "yielded valuable information on the curricular intentions of participating countries" (Martin, Gregory, & Stemler, 2000). Even though the design was essentially unchanged, some important changes in the *methods* employed were introduced in TIMSS 1999, which proved significant for the development of successive TIMSS cycles. As far as the *achievement test* approach is concerned, additional items were developed since two-thirds of the items from TIMSS 1995 had been released and consequently had to be replaced by similar items in order to cover the framework. In so doing, TIMSS 1999 introduced the focus on trends which later became a "trademark" of TIMSS. In earlier studies, some items had been reused, but there had not been a focus on the trend aspect as such. Next, substantial and influential changes in the third approach, *the questionnaires*, were implemented. A curriculum questionnaire to be answered by the National Research Coordinator of each participating country, summarizing features of the school system on a national level, was introduced. Similar questionnaires were used in all subsequent TIMSS cycles. Whereas the TIMSS 1999 school questionnaire was very similar to the 1995 version, several changes were made to the teacher questionnaires for the 1999 cycle, mainly because the previous ones were considered too lengthy. In the student questionnaire, questions dealing with student self-concept in mathematics, Internet access, and its use for mathematical activities were added. It is an interesting fact that outcomes of the TIMSS Video Study helped frame a set of questions about activities in mathematics classes in TIMSS 1999.

TIMSS 2003 confirmed the focus on trends introduced in TIMSS 1999. Furthermore, the transition of definitions of participating populations from age to

years of schooling was taken one step further. In addition to a basic definition based on age, the population definition stated that the identified grade level was intended to represent 4 and 8 years of schooling (Martin et al., 2004). In the first three cycles of TIMSS (1995, 1999, and 2003), *student achievement* in mathematics in addition to an overall result was reported in content domains (e.g., algebra, geometry). At the time, several other international studies (e.g., PIRLS—also conducted by IEA—and PISA) had introduced reporting of student achievement in different cognitive domains. TIMSS participating countries also expressed a need for comparative information about cognitive aspects of how students performed in mathematics (and science). An international group of mathematics experts was gathered to develop categories that could be the basis for meaningful reporting of achievement in cognitive domains. Previous definitions of four cognitive domains had been used in the development of items for the TIMSS assessments, but the existing model led to some overlap across these domains. The expert group worked to develop mutually exclusive cognitive domains for reporting the TIMSS 2003 results (Mullis, Martin, & Foy, 2005) leading to the definition of three cognitive domains: knowing facts, procedures and concepts; applying knowledge and understanding; and reasoning. These domains, supported by categorization of items from TIMSS 2003 and reanalysis of TIMSS 2003 data with respect to these categories, were published in 2005 (Mullis, Martin, & Foy, 2005).

Further refinement of the assessment framework was done in the early stages of TIMSS 2007 as published in the TIMSS 2007 assessment frameworks (Mullis, Martin, Ruddock et al., 2005). Based on the development project mentioned above, the number of content domains and cognitive domains was decreased. The revision of the framework was at least partly a consequence of a decision made that, beginning with TIMSS 2007, frameworks were to be updated with every cycle of the study, thereby permitting the frameworks, the achievement tests, and the procedures to evolve gradually into the future. Another small but still significant change from 2003 to 2007 is found in the definition of the study populations. An important feature of the research design that TIMSS represents is that these populations must be defined rather precisely and can be viewed as "a collection of units to which the survey results apply" (Olson et al., 2008, p. 78). A subset of the target population was sampled for participation in the study, and a lot of effort was put into identifying the sample in such a way that results from the sample can be generalized to the entire target population.

TIMSS Advanced 2008 focussed on a population which had not been targeted in IEA studies since TIMSS 1995—that is, students at the end of upper secondary education (Grade 12) who had taken courses in advanced mathematics. Apart from that, the basic *design* was essentially the same as for TIMSS 2007, the aim being to study the intended, the implemented, and the achieved curriculum. The *methods* used were also similar to those of TIMSS 2007. The assessment framework guiding the development and construction of instruments defined three broad mathematical content domains (algebra, calculus, and geometry) and three cognitive domains (knowing, applying, and reasoning) (Garden et al., 2006).

PISA: Programme for International Student Achievement

PISA 2000 and studies which followed upon it, were not research studies as such even though they have given rise to several research questions, some of which have been pursued in follow-up studies. Instead, PISA is a survey designed to assess students' "ability to complete tasks relating to real life, depending on a broad understanding of key concepts, rather than assessing the possession of specific knowledge" (OECD, 2001, p. 19). Thus the *design* of PISA 2000 was focussed on charting students' performance with regard to reading (the major domain), mathematical and scientific (the minor domains) *literacy* (see the definition of mathematical literacy above), and relating such performance to student and school background factors. Correspondingly, four approaches were pursued: constructing an assessment *framework* for literacy (OECD, 1999), constructing and administering *achievement tests*, and constructing and administering *background questionnaires*. Further, an approach to *ranking participating countries* according to various performance variables was part of the design as well. The basic decision to assess 15-year-olds in participating countries was taken much before the other design decisions.

The *method* undertaken in constructing the framework was to ask an expert group for each domain, to devise such a framework. As far as mathematics is concerned, the framework contained three dimensions: a content dimension, which for PISA 2000 had two components "change and relations" and "space and shape"; a process dimension (called "competency clusters") "reproduction," "connections," and "reflection"; and a situation dimension focussing on the spheres in which students live, that is, private/personal, school, work and sports, local community and society, and scientific spheres of life. These dimensions then formed the platform for constructing the test items. The items were devised to be literacy items and were, moreover, to be cast in one of three paper-and-pencil response formats: multiple choice, closed constructed, and open constructed response. A total of 64 items, chosen as a result of extensive field-testing, comprised the test.

The methods involved in identifying educational background factors and relating them to student performance consisted in devising two questionnaires: a student and a school questionnaire. Responses to those questionnaires were then correlated by way of several statistical analyses to student performance so as to explain a multitude of performance variations. Also, the methods employed in ranking countries by way of certain ranking measures were probabilistic and statistical in nature, based, more specifically, on the so-called Rasch model. In particular, the methods in item response theory were utilized.

The *instruments* adopted consisted of the actual student tests and questionnaire and a school questionnaire to be completed by the principals of the schools whose students were included in the sample. Each student was given one out of nine item booklets, containing items from the three domains (reading, mathematics, science) without any indication of which domain they belonged to. This rotation principle implied that different students were completing different booklets. Each student was given two hours to complete the booklet. The questionnaire that each student was asked to complete was a 30-minutes questionnaire containing questions about

students' and parents' economic, cultural, and social status; student characteristics and family backgrounds; and learning strategies and attitudes (OECD, 2001). The school principals' questionnaire—which also was meant to take 20 to 30 minutes to complete—contained questions concerning school policies and practices, classroom practices, school resources and type of school.

In PISA 2003, mathematics was the major domain, the aims and overall design were not much different from those of PISA 2000, except in one respect: *trends* from PISA 2000 to PISA 2003 were sought. As before, the primary aim of the OECD/PISA assessment was "to determine the extent to which young people have acquired the wider knowledge and skills in reading, mathematical and scientific literacy" that they would need in adult life (OECD, 2003, p. 12).

The *framework* part of the design was unchanged along the main lines. But there were minor changes in the content, process, and situations dimensions. Two new content categories, "overarching ideas," were added to the ones in PISA 2000; namely, "quantity" and "uncertainty," thus forming a total of four. The situation and context categories were slightly modified as well. As to the mathematical process dimension, the notion of eight mathematical competencies as developed in the Danish KOM-project (Niss & Hoejgaard, 2011; Niss & Jensen, 2002) was introduced to underpin the competency clusters that were utilized in PISA 2000.

In the *achievement test*, a rotated design was employed, with a total of 85 mathematics items included in the pool, 20 of which were also used in PISA 2000. These are called "link items." Student and school *questionnaires* were included as in PISA 2000, and also contained questions concerning students' self-concept, learning strategies, and affects specifically concerning mathematics. Again, the items were selected and the questionnaires finalized after substantial field trialling.

The method adopted for *charting trends* in mathematics performance from PISA 2000 to PISA 2003 was to establish common PISA 2000–2003 performance scales. This was done by using the detected changes of difficulty in the 20 link items from 2000 to 2003 to construct a transformation of scores so as to fit a common scale (OECD, 2004), having 500 score points as the OECD average. With that in hand, PISA 2000 and PISA 2003 subscales for the two content categories which were common to both cycles, "space and shape" and "change and relationships," were constructed. It was then possible to see that the OECD average in space and shape grew from 494 to 496 score points, whereas in change and relationships, scores grew from 488 to 499. The 2003 score for quantity was 501, and for uncertainty 502. It did not make sense to make an overall comparison of mathematics performance from 2000 to 2003, since the combined average score was set to be 500.

In PISA 2006 and PISA 2009, mathematics was again a minor domain. Therefore, only minor changes were made to the *design* of the study as far as mathematics and student and school questionnaires are concerned. In 2006 only 48 items were used. As these were also included in 2003, they were all link items. Each participating student received a randomly selected booklet. With regard to detection of trends the PISA 2003 scale with an average OECD score of 500 was used as the benchmark (OECD, 2007), and again the link items were used to create a transformation that allowed for comparison between the two assessments. The OECD mathematics

score for 2006 was 498, which was not significantly different from the 500 in 2003. In 2009 the total testing time in mathematics was reduced and only 35 items were included in the test. The OECD average score in mathematics 2009 was 496, which was not significantly different from 2006.

Various changes were incorporated in PISA 2012, when mathematics was again the major domain, but it is premature to go into details with these changes. For current information consult OECD (2010b). More changes are likely to occur from 2015 as a new contractor will be in charge of the future development of frameworks.

The TIMSS Video and Case Studies

In the TIMSS Video Study, the *design* adopted was chosen so as to reduce the conceptual and terminological ambiguities within and across cultures that could arise from using questionnaires, as well as to avoid dependence on coding schemes fixed beforehand and the impossibility of critical scrutiny of documentation of live observations (Kawanaka et al. 1999):

> We needed data that could be analyzed and re-analyzed objectively by researchers working from a variety of perspectives. The idea of using videotapes began to emerge, and the final decision was made to collect direct information on classroom processes by videotaping instructional practices. (p. 88)

So, approaching the reality of classrooms by *videotaping* them was, of course, the fundamental approach in the study. This decision allowed researchers to engage in many iterations and related discussions between observations and post hoc coding of the observations. Teachers' views of the representativeness of the lessons videotaped and their goals were sought as well, by means of *questionnaires*. The next key approach in the design was *analyzing and coding the data* generated by the videotapes, and the final approach was to devise ways to *represent and depict mathematics classroom reality* in a manner that would make sense to researchers outside the project.

Each of these approaches gave rise to its own set of *methodological issues* and decisions. First, how to *sample the classrooms* that were to be videotaped, and when and for how long should they be videotaped? Another important issue to decide upon was what to aim cameras at and hence what type of classroom activities to document. It was decided to focus on the middle TIMSS population only (eighth grade) in Germany, Japan, and the USA. The classrooms sampled were a subsample of the national random probability samples in TIMSS 1995. Eventually 100 German, 50 Japanese, and 81 US classrooms were included in the study. Classrooms were videotaped in 1994–1995 (Stigler et al., 1999) evenly across the school year in Germany and in the USA, but less so in Japan, where the sample was skewed towards a time of the school year when geometry was predominant in the curriculum (Kawanaka et al., 1999).

When seeking a method for *coding the tapes*, Kawanaka et al. (1999) had three dimensions in focus: the work environment in the classroom, the nature of the work students are engaged in, and the methods teachers use for engaging students in work. The coding schemes were developed with the aim to construct objective and reliable categories and codes that allowed for capturing, representing, and quantifying characteristic features and patterns in the classrooms of the three countries.

In putting the method of videotaping into practice, the actual *instrument* employed was to film one complete lesson per classroom by one camera, representing the perspective of an ideal(ized) student, typically focussing on the teacher. Prior to that event, participating teachers were given a common set of information and instructions, and afterwards they completed the questionnaires mentioned above (Stigler et al., 1999). All videotapes were digitized, and lessons were translated into English and transcribed, linking the transcript to the video by time codes (Kawanaka et al., 1999). The final instrument for coding was very elaborate. It focussed on what was called "lesson tables."

> These lesson tables were skeletons of each lesson that showed, on a time-indexed chart, how the lesson was organized through alternating segments of classwork and seatwork, what pedagogical activities were used …, what tasks were presented and the solution strategies for the tasks that were offered by the teacher and by the students. (p. 96)

The tables included several components: organization of the class; outside interruption; organization of interaction; activity segments; mathematical content referring to units (Stigler et al., 1999) and to mathematical topics (numbers; measurement; geometry; proportionality; functions, relations and equations; data representation, probability and statistics; elementary analysis; validation and structure; other). Also, a very detailed coding of classroom discourse, based on a rather fine-grained division of public talk and private talk, respectively, was undertaken. Coding schemes were refined along the road when warranted by the analysis of the videos and intercoder reliability checks (Kawanaka et al., 1999). In addition to being guides to the entire video of a classroom, the lesson tables also served as separate reporting outcomes which could themselves be coded. Statistical analyses were conducted to capture and describe patterns for comparison across the three countries.

The *design* of the TIMSS Case Study encompassed three approaches to seeking in-depth answers to the initiating question "about the beliefs, attitudes and practices of students, parents and teachers" in Germany, Japan, and the USA (Stevenson, 1999). The *first approach* was to identify the topics on which information was to be sought. The method adopted was to select, after consultation with the funding agencies, four such topics: national standards, teachers' training and working environment, dealing with differences in students' ability, and, finally, the place of secondary school in adolescents' lives. One of the *instruments* put to use in relation to this method was to attach a number (15 to 35) of predetermined tags, in terms of key concepts and words, to each topic so as to facilitate subsequent computerized retrieval of the tagged instances. It was further decided not to form a particular set of hypotheses from the outset but to let them be generated from the data collected. The *second approach* was to identify the units from which information

should be collected. The method then was to concentrate on one primary and two secondary sites in each of the three countries, all chosen to be representative in demographic and socioeconomic terms. Each site would contain several schools. The *third*—key—*approach* concerned the ways in which researchers were to gather information. Here the method was to make each researcher responsible for one of the four topics and to conduct a number of so-called encounters (i.e., interviews, observations, conversations) of a minimum duration of one hour. Moreover, each researcher was to produce and circulate weekly field notes—another instrument—to the other researchers. A total of more than 960 encounters were conducted in the three countries. In addition, 250 hours of observation of mathematics and science classes were carried out. All interviews were to be conducted according to a predetermined semi-structured format, which involved yet another instrument. Whenever possible, the encounters were tape-recorded, which constituted the final instrument involved in implementing the third approach.

The Learner's Perspective Study

The design used within the TIMSS Videotape Study was extended for use in the Learner's Perspective Study (LPS), and measures were taken to improve the possibilities to capture not only teachers' activities during lessons but also the students' learning processes. The capturing of students' learning processes—the *first approach* in the *design*—was operationalized by adding some features to the design of the TIMSS Videotape Study. An important such feature, which differs from earlier major studies with comparative possibilities, was that sequences of lessons rather than singular ones were documented. A minimum of 10 consecutive lessons were recorded at each site. The main characteristic of the method adopted in this approach is the use of video documentation of teachers' and students' work in eighth-grade mathematics classrooms. Three cameras were used in each classroom: one stationary camera equipped with a wide-angle lens capturing as much of the classroom as possible, a second one pointing to a group of so-called focus students, and finally a manually operated camera following and documenting the activities of the teacher. Depending on the seating plan, one to four focus students' work was video- and audio-recorded in each lesson.

In each city, three teachers' classrooms were selected for recording. The relatively small number of classrooms investigated is a trade-off with the comparatively large number of consecutive lessons documented. The sampling of participating teachers, classrooms, and hence students was not made randomly but was based on the selection of "competent" teachers as defined by the local community in each city and country. The focus students were interviewed in a stimulated recall interview—the *second approach* in the design—after the lesson. This decision was informed by the aim to explore learners' practices and allow them to generate reconstructive accounts of classroom events. Three times during a lesson sequence the teachers, too, were interviewed in a subsequent stimulated recall session. The actual recordings of the focus

student and the teacher cameras were used as recall stimulus in the interviews (Clarke, 1998, 2001, 2003, n.d.). The interviewees were invited to comment on each recorded lesson in terms of what they found significant in the classroom activities. They were in control of the replay of the videos and could freely choose when to use the fast forward (or rewind) buttons and when to stop and comment on the recordings.

Documenting sequences of lessons allows for analyses of single lessons but also analyses that stretch beyond those, hence making it possible to address questions on how both teaching and learning unfold over a longer period of time. When it comes to analyzing the data—*the third approach*—there is no framework common to all the participating research groups in the network. However, the overall approach is informed by a Vygotskian point of view where teaching and learning are seen as mutually constitutive processes.

Complementarity is a distinguishing characteristic of the research design on four levels (Clarke, Emanuelsson, Jablonka, & Mok, 2006):

> (a) At the level of data, the accounts of the various classroom participants are juxtaposed; (b) At the level of primary interpretation, complementary interpretations are developed by the research team from the various data sources related to particular incidents, settings, or individuals; (c) At the level of theoretical framework, complementary analyses are generated from a common data set through the application by different members of the research team of distinct analytical frameworks; and (d) At the level of culture, complementary characterizations of practice and meaning are constructed for the classrooms in each culture (and by the researchers from each culture) and these characterizations can then be compared and any similarities or differences identified for further analysis, particularly from the perspective of potential cross-cultural transfer. (pp. 12–13)

All video materials were transcribed and translated into English. The transcripts, together with digitized videos, were included in a database which also contained seating plans describing students' positions during class and so-called lesson plans; that is, rough summaries of each lesson. Survey materials such as short teacher questionnaires, performance tests compiled from released items from TIMSS studies, scanned copies of the focus students' work, and textbooks were also part of the integrated datasets constructed by each participating research group.

The US–Japan Problem-Solving Study

In order to compare and contrast Japanese and US students' abilities, behaviours, and views concerning problem solving in mathematics, the design of the US–Japan Cross-cultural Research on Students' Problem-Solving Behaviors (Becker, 1992; Becker et al., 1999) included the following *four approaches*. First, the subjects to be studied had to be specified. Next, the ways in which they were to be studied had to be determined. More specifically, it was decided to put the students selected to work on certain tasks, and they as well as their teachers were asked to complete questionnaires pertinent to the problems solved and to mathematics at large. Finally, student problem responses were coded by means of certain predetermined categories, and the questionnaire answers were analyzed.

As to the *first approach*, the subjects to be studied formed a number of populations in the two countries. The method was to sample students—with their teachers—in 4th, 6th, 8th, and 11th grades from large rural, small urban and large urban schools in Japan and the USA in the school year 1989–1990. The selection of the schools seems to have been made on pragmatic grounds, namely from districts near the researchers' own institutions. At least two classes participated in each region in each country. Neither the schools nor the classes were randomly selected (Becker, 1992; Becker et al., 1999). The number of students involved in the study was several hundred from each population in both the USA and Japan.

The method employed to implement the *second approach* was to give all but the 11th-grade students two problems to solve. The problems had been used and investigated by researchers in previous studies, and their final formulation and place in problem work booklets—the *instrument* employed—had been tried out in a pilot study (Becker, 1992). The US 11th-graders also got an extra problem to solve. Each student was given exactly 15 minutes to solve each of the two problems, except that the US 11th graders got an additional 10 minutes to solve the third problem. For all problems, students were asked to solve them in as many different ways as possible—on separate answer sheets handed out to them—within the given time frame. This introduces an unusual feature in task-based studies, which usually only ask for single solutions.

As to the *third approach*, students were asked to fill out a questionnaire—forming one *instrument* in this approach—after having worked on the problems. The questionnaire contained questions concerning students' degrees of interest, difficulty, and familiarity with the problems they had just solved, and their attitudes and self-concept with regard to mathematics. Teachers were asked to fill out their questionnaires (another instrument) while the students were doing the problems. These questionnaires, in addition to seeking information about the school and the students, addressed the teacher's view of the problems posed and of the students' reactions to them (Becker et al., 1999).

The *final approach* was to analyze the data collected. Individual or pairs of researchers were responsible for analyzing the data for one problem (Becker et al., 1999). The focus of the analyses, which often made use of categories established by previous Japanese or American research, was on comparison of correctness of responses, solution strategies, and modes of explanation.

The Kassel Project

The so-called Kassel Project (Kaiser, 1999b), aiming at comparing essential features of secondary mathematics teaching and learning in England and Germany and at explaining observed differences, had a *design* which in important respects differed from the designs adopted in most international studies, even though the study—as is often the case—is a combined quantitative and qualitative one. The quantitative part of the study was focussed on longitudinal student achievement,

and the qualitative part concentrated on capturing and charting key features of teaching and learning in the two countries.

In the achievement part of the study, the *first approach* was to identify two comparable lower secondary cohorts in 1993 (a sample of about 800 students in Year 8 in Germany and about 1,000 students in Year 9 in England) who were then followed and tested until Year 10 and, respectively, Year 11. Testing—the *second approach*—was conducted in 1993, 1994, and 1996 (Kaiser, 1999b). Tests were informal, non-standardized, and based on an analysis of curricula. All test rounds covered three large topic areas: number, algebra, and functions and graphs with geometry.

The main difference from other studies lies in the qualitative part of the project. The *third approach* adopted was to conduct participant observer classroom observations in about 240 lessons in 17 schools in England and about 100 lessons in 12 German schools (Kaiser, 1999b). Based on the entire set of observations, idealized descriptions—*constituting the fourth approach*—of typical mathematics teaching in German and English classrooms were constructed so as to encompass the following three foci: mathematical theory (including introduction of new concepts and methods, importance of theory and rules, organization by subject structure or a spiral curriculum, the role of proofs, rules versus examples, the role of precise language and formal notations); the role of real-world examples; and teaching and learning styles (for further details of the method adopted, see Kaiser, 1997).

The 13th ICMI Study

The final study to be considered here is the 13th ICMI Study, *Mathematics Education in Different Cultural Traditions: A Comparative Study of East Asia and the West* (Leung et al., 2006). This study was not a uniform, coherent one, based on one single design, but rather an umbrella overarching a variety of specific studies with different foci and perspectives, all seeking to compare and contrast fundamental features of mathematics education in East Asia and the West.

We confine ourselves to outlining, in an aggregate manner, some of the most significant aspects of methodology involved in this study. One of the pertinent issues dealt with in the study was how it can be that East Asian students excel in international comparative mathematics achievement tests such as TIMSS and PISA while at the same time possessing negative attitudes and low self-concept towards mathematics and its study (Leung et al., 2006). One of the methods adopted to answer this question is to undertake historico-cultural investigations of the origins and development of the fundamental traditions in East Asian and Western countries, in particular with regard to the role of the teacher. This was done, for example, in Hirabayashi's and Ueno's chapters, as far as Japan is concerned, and in Wong's and Li Shiqi chapters concerning China. The Western tradition was depicted in Keitel's chapter. Analytic comparisons between Eastern and Western *curricula* were made in Bessot and Comiti's chapter on French and Vietnamese curricula, and in Wu and Zhang's chapter, whereas comparative analyses of Eastern and Western *textbooks*

were presented in Li Yeping and Ginsburg's chapter and in Park and Leung's chapter. There was a focus on teachers' beliefs and values in the last part of the book. Perry, Wong, and Howard's chapter reported on a questionnaire-based study comparing Australian and Hong Kong primary and secondary teachers' beliefs about mathematics, mathematics learning, and mathematics teaching, whereas middle-school teachers' beliefs in the USA and China were studied through a combined questionnaire–interview–observation approach reported in An, Kulm, Wu, Ma, and Wang's chapter.

The ICMI study book also included chapters which surveyed the other comparative studies referred to in the present chapter. In summary, a fair sample of the spectrum of research methods employed in international studies of mathematics teaching and learning were represented in the ICMI volume.

Reflections on Designs and Methods

Before we summarize, analyze, and reflect on the designs and methods encountered in the studies presented in this chapter, we consider a more fundamental question which has been briefly touched upon above: To what extent are international comparative studies at all possible and meaningful? It goes without saying that the very carrying out of such studies presupposes that they appear as both meaningful and possible to those who conduct them. Otherwise they would not exist. As this is a deep and complex issue, which in a way deserves a chapter of its own, we have to confine ourselves to sketching some basic deliberations.

First, one should bear in mind that the task of this chapter is to present and analyze—from a methodological perspective—studies that actually exist. The primary task is not to assess and judge them. The agencies and people who instigate and conduct the studies—the primary stakeholders—do so for a purpose, and to them the most significant issues therefore are whether a given study serves its purpose and can be said to be methodologically sound so as to produce results that are useful, valid, and reliable relative to that purpose. What is likely to be less important to the stakeholders of a study is whether or not it is useful, valid, and reliable with respect to other sorts of purposes. So, any critique of a study should be more concerned with the extent to which it lives up to what it purports to be, than with its capability of responding adequately to something else.

There are two components involved in "international comparative studies," namely "international" and "comparative study." The fundamental component in the question about the possibility of international comparative studies seems to be the very notion of comparative study. Whenever entities (such as objects, situations, conditions, relationships, mechanisms, phenomena, or categories of contents) are subjected to any form of comparison, certain features of the entities are deemed irrelevant or less important and left out of consideration, yet others are chosen to be in focus. How then can one be sure that the entities left out of consideration do not—behind the curtain, so to speak—exert a significant influence on the features

actually considered in the comparison? In disciplined inquiry in general, and in science in particular, this may well be seen as the most essential question of all. Since it is usually extremely difficult to guarantee that no hidden variables have an impact on the entities being compared, and hence on the outcomes of a comparative study, the most important thing is to subject the study to open discussion, critical scrutiny, alternative studies, methodological debate, and so on, much of which will concentrate on the balance between the factors left out, or kept in, in the comparisons undertaken.

When comparative studies deal with human beings and human behaviour, the issues just mentioned become aggravated. For instance, this is the case when we compare *n*th-grade students in different schools in the same town, in different parts of the same country, in different socioeconomic, ethnic, cultural, or religious groups, and so on. Going beyond the borders of one country to involve other countries, continents, cultural traditions, and so on, implies further complexity. It introduces changes of degree or orders of magnitude, but not fundamental changes. Needless to say, even more openness, care, analysis, scrutiny, and alternative views or investigations are needed in international comparative studies than in other kinds of comparative studies. But it would be unreasonable to claim that whereas comparisons between *n*th-grade students in two schools in neighbourhood *N* of municipality *M* in county *C* in country *S* are perfectly possible and meaningful, the possibility and meaningfulness of comparisons disappear when national, regional, continental, or cultural borders are being crossed.

We now offer a number of more specific observations concerning the designs and methodologies of international studies of the teaching and learning of mathematics. The first observation worth making is that a large fraction of the studies have investigated *student achievement on written tasks* as a key component of their design, not only when assessment of achievement *is* actually the primary subject of study but also when the purpose of the study is to come to grips with something else. Since the time allocated per test item is usually very limited, ranging from 1 to 2 minutes, to 15 minutes, only those kinds of achievement which can come to fruition within such a time frame are represented in the tests. It is remarkable that student achievement on tasks is taken to epitomize *mathematical competence at large*. This fact is indeed worth discussing, not only because of the constrained spectrum of forms of achievement which can find their way into the test but also because mathematical competence possesses many more significant dimensions than the ability to do well in achievement tests. It can, of course, be very well justified to include achievement tests in a given study, and sophisticated test items may have been developed for the study according to the highest international standards. So, achievement tests are not a problem in and of themselves. However, a problem occurs when no other probes into mathematical competence are taken into account and employed.

The problem is aggravated when media, politicians, and other outsiders to mathematics education oversimplify things even further by equating mathematical competence with success on achievement tests. It is not unusual to encounter the following line of argumentation: As the test results are numbers that speak for themselves, you are not allowed to interpret what they tell us, let alone to argue against them.

This way of reasoning is not very different from saying: The thermometer in my hand yields a result you can't argue with. It displays the gravitational force on the spot where I'm standing! It may be seen as surprising that no international studies have devised methods to investigate aspects of students' mathematical competence that cannot be accessed by tightly time-constrained achievement tests. It is conceivable that future international studies would benefit considerably from the development of new kinds of gauges of mathematical competence. There are, however, huge challenges in adopting more complex assessment situations in the wide variety of school contexts found among the many countries participating in large-scale international studies such as TIMSS and PISA.

In cross-national achievement studies, the fact that all the items included in the tasks have to be meaningful and reasonable in participating countries leads to a fair amount of harmonization of items, item types, response formats, and score coding. This is true both of curriculum-referenced studies, such as TIMSS, where items at least to some extent have to be related to the curricula students have been exposed to, and of literacy or competency referenced studies, where some basic degree of familiarity with contexts and situations needs to be ascertained, as is the case in, for example, PISA. It poses particular challenges to test mathematics embedded in extra-mathematical contexts in a manner that is not *too* dependent on cultural, technological, or socio-economic contexts. All this being said, items in international studies are typically highly thoughtfully and carefully constructed, developed, piloted, field-trialled, score coded, and rated, sometimes with an impressive degree of sophistication. Also, sophisticated item analysis methods that allow for studying achievement conditioned on a variety of (sub-)population characteristics and other background variables are put into use in many studies, especially large-scale ones such as TIMSS and PISA. Against this background, the various pools of items from international studies are goldmines for research and practice, as are the multitude of achievement databases, many of which have already been subjected to several correlation studies. However, unfortunately this happens too seldom, and the existing item pools and databases deserve to be put to use in new research.

The next observation is that even though student achievement tests are a major component in several international studies, tests never stand alone. They are *always accompanied by other approaches and instruments*, such as student or teacher questionnaires and interviews, classroom observations, analyses of written materials such as curriculum documents, teacher education programs, textbooks, and assessment instruments. There are three main reasons for making use of such other approaches in relation to achievement studies: to provide a means for interpreting and understanding what students had in mind in their solution processes and how these were related to what and how they had been taught, to provide causal or correlational explanations of students' achievement or of related observed phenomena in terms of background factors and variables, or to provide an entirely different sort of information from that sought in the achievement tests; for instance, about students' attitudes, beliefs, and career perspectives. It goes without saying that the approaches listed above are not only utilized in connection with achievement studies, they also can, and often do, stand alone as independent approaches.

As an independent approach, or in addition to other approaches, *questionnaires* to students, teachers, school principals, or other target groups primarily serve two purposes. Sometimes the primary purpose is to gather information of intrinsic, separate interest in the study. At other times, it is to constitute a platform to follow up on or lead into other approaches, say, classroom observations or interviews. Questionnaire questions come in different types. Some questions ask for factual, unambiguous, objective information such as student sex and age, number of students in a class, types of school programs, and the like. Other questions may ask for multiple-choice responses representing predetermined, but not necessarily well-defined, entities, while other questions may ask the respondent to describe objects, phenomena, or situations in his or her own words, and still others may concern affective or attitudinal matters.

It is generally acknowledged that questionnaires give rise to many methodological issues, at least as far as nontrivial, nonfactual questions are concerned. One such issue is that the response categories offered in multiple-choice questions may often not be understood or accepted by respondents, for instance, because of ambiguity or problematic demarcation lines between response options. This becomes a special concern when questionnaire responses are subjected to subsequent quantitative aggregation. A related issue concerns questions in which respondents are asked to estimate the frequency of the occurrence of certain kinds of experiences or acts, where it may simply be difficult to remember things well enough to provide reliable answers. Another issue is to do with questions that ask respondents to write comments or statements which are likely to be difficult to interpret by researchers. In some contexts, respondents may tend to figure out which answers are "good" or "right," or would impress or please those who administer the questionnaires, and then respond accordingly. Moreover, there may well be socio-cultural biases in the occurrence of this tendency. (Similar arguments are posed in relation to video observation and interviews considered below.) However, designed with reflection and care and treated with caution, questionnaires can be powerful instruments, both in quantitative surveys and in in-depth qualitative investigations.

As with questionnaires, *interviews* can be a method to gather information of independent research interest, and they may constitute an approach accompanying other approaches. To the extent interviews are used in large-scale studies, they are typically used for the latter purpose, as a method to probe deeper into issues or phenomena which have emerged through other means, such as achievement tests, questionnaires, or classroom observations. It may be that students' comments and reflections on their solutions to problems are sought, in order to shed light on their background knowledge, strategies, or solution processes. Or it may be that elaboration on students' responses to attitudinal questions in a questionnaire is needed, either as a means for ascertaining investigators' interpretation of the responses or as a way to resolve possible inconsistencies in the responses. Or it may be that the reasons for teachers' observed acts and decisions in classrooms need or deserve further elucidation. Usually interviews employed in large-scale studies address a much smaller subject sample than does the study itself. Therefore, such interview data are rarely aggregated in a quantitative form but remain qualitative data, possibly

subjected to some sort of classification. Since the interviews typically serve specific purposes, seeking certain kinds of information, they often—but not always—take place according to some protocol, either a completely structured protocol, not admitting deviation from predefined questions, or a semi-structured protocol that admits tangential excursions to follow up on the responses obtained while returning to the main track afterwards.

When interviews are used in focal studies, all the features just mentioned apply as well, but additional features become relevant. Most importantly, in some studies interviews are given the predominant role. This is typically the case when respondents' comments, experiences, or views are sought on a broad spectrum of topics or issues—for instance, when the aim is to obtain a multi-faceted and integral picture of the respondents selected. In such cases, loosely structured interviews may come into play; that is, in the shape of more freely flowing conversations in which the route taken by the interviewer depends on what happens along the road.

The conducting of interviews poses many challenges, as do their recording, registration, analysis, and sometimes coding. It is often demanding to "get what one is after," because interviews are a form of human interaction and hence subjected to implicit or explicit socio-cultural boundary conditions, which are likely to differ from country to country. It may not only be difficult to obtain a fair degree of homogeneity across countries, but also challenging for the interviewer to steer the conversation according to the interview protocol and pose follow-up questions while paying close attention to the social relationship with the interviewee and perhaps managing the equipment, taking field notes, and so on. Recording, registering, transcribing, coding, or otherwise analyzing the interviews conducted are enormously time-consuming and intellectually demanding activities, especially when it comes to selecting what to store and to interpreting what respondents said. No wonder that a huge body of research literature exists on interviews as a research method.

Comparative *classroom studies*, too, have given rise to huge bodies of methodological considerations, many of which pay special attention to the instruments, procedures, and techniques involved in conducting such studies. As is the case with interviews, classroom studies can take place with varying degrees of structuring, ranging from unstructured studies in which observers, whether participant or neutral observers, focus on what appears to them to be significant along the road, to semi-structured studies, in which researchers concentrate their attention (or intervention) on certain predetermined topics or issues but are also ready to follow up on interesting opportunities or sidetracks that emerge during classroom sessions, through to completely structured studies, where researchers record and classify instances of certain sorts of phenomena or situations in predefined categories and neglect everything else. Since the mathematics classroom is an immensely complex organism, the set of potential objects of study is immensely complex as well. Forms and content of classroom interaction and communication between the teachers and the whole class, student groups or individual students, or among students, may be one possible focus point. Student activities and the teacher's role in orchestrating them may be another focus point, as may student behaviour in particular respects, for example problem solving, hypothesis formation, or explanation of solutions.

Also the nature of the mathematics actually being dealt with in the classroom by teacher or students may be of interest to researchers.

The main reason that the technicalities of classroom research preoccupy researchers is that a classroom session is by definition of a transient nature, so measures that make it possible to register and fix the significant components of the session, either for documentation or for later analysis, are crucial for the entire undertaking. Field notes or written forms to be filled out by the researcher during class, audio or video recordings of whole sessions or episodes, are some of the instruments typically used alone or in combination in such research. Providing detailed information about the procedures followed and the circumstances under which the instruments have been employed is an important documentation task. The concurrent or post hoc coding of the classroom entities identified, and the grounds on which the coding has been performed, are equally important tasks, as is the tracing of them in the analysis. This is not the place to go into details. It is worth noting, though, that some of the international methodological and technical standards for classroom study research in mathematics education have to a large extent been established and moved forward by the international comparative studies, especially as regards to the handling of large samples.

Comparative *curriculum and textbook analyses* are conducted on written documents, and the methods employed therefore involve text analysis. However, apart from general aspects of such analysis and analysis of curricula in relation to education systems—what students get what sort of education, where, with whom, and taught by whom—curriculum and textbook analyses in mathematics education have strong mathematical components in terms of content, exposition, processes, competencies, tasks, activities, and so on, which can be analyzed in a multitude of different, and sometimes even conflicting, ways. Therefore, frameworks for curriculum and textbook analyses in mathematics represent important methodological challenges and decisions, the outcomes of which have a decisive impact on the nature and results of the research conducted. Here, too, many of the international comparative studies considered have contributed to setting and improving significant aspects of the standards of research internationally.

Tasks for students are essential in teaching and learning of school mathematics and in international achievement tests, to such a degree that the nature of the tasks given to students to a large extent codetermines the outcomes of international studies. Against this background, task construction and task analysis become key methodological issues. It is interesting to observe that already in FIMS a matrix-based framework (content-by-cognitive behaviour level) for selecting and analyzing test items was put into practice. In other words, test items were classified not according to more or less traditional content strands only, but according to other dimensions as well. This was the case with all subsequent large-scale studies, including TIMSS. The schemes adopted by PISA were the most elaborate of all. Item classification according to different dimensions gives rise to a variety of correlational item analysis studies of a statistical type.

In addition to the tasks employed in comparative studies, it is also interesting to study the kinds of tasks utilized in mathematics education in different countries. It is therefore somewhat surprising that only few publications of this kind exist. An exception is the book by Shimizu, Kaur, and Clarke (2010).

Concluding Comments

This chapter has attempted two things: (a) to provide a detailed account of the purposes and goals of a number of important large-scale or focal studies of mathematics teaching and learning internationally and of the designs and methods employed to conduct these studies; (b) to analyze and reflect on those designs and methods.

We have found that most of the studies have adopted a *multi-faceted design*, in which combinations of *different approaches* have been used to answer different subsets of the set of questions that gave rise to the study. These approaches are as follows: frameworks to conceptualize the domain being studied, especially as regards mathematics as a subject; construction and administering of student achievement tests; analysis of intrinsic item characteristics; analysis of student responses; student, teacher, or school questionnaires; sampling of the populations studied; interviews with students, teachers, or parents, and methods for analyzing the outcomes; observation (participant or neutral) of real classrooms and methods for recording and analyzing the resulting data; analysis of curricula as part of education systems and as separate entities, textbooks, and assessment tasks; and analytic reflections on the traditions and cultural environments of mathematics education.

Together with these approaches comes a variety of different methods, each of which is implemented by the use of various specific instruments. The methods and the instruments in turn involve a multitude of different procedures and techniques that we have had to leave aside in this chapter, even though quite a few of them are interesting in their own right.

It is a remarkable fact that most, if not all, of the studies considered have contributed to substantial progress in the development of the research designs, approaches, methods, and instruments applied in the studies. Among other things, this progress is due to the fact that several studies have had many human and material resources at their disposal, primarily because the stakeholders of the studies often attribute large amounts of prestige and impact to the outcomes and the politico-administrative uses of the studies.

This phenomenon implies that several sorts of research not meant to deal with international comparisons of one kind or the other can benefit greatly from the contributions to research methodology offered by the international studies.

We have found, however, that the studies display certain limitations as well. This is particularly true of the approaches to gauge student achievement in mathematics, where time and format constraints exclude essential aspects of mathematical competence from being taken into consideration in the studies. This is an issue on which substantial new developments are sorely needed.

Another limitation has been that the overall cultural, economic, and structural contexts and boundary conditions of the education systems at large, and of schooling in particular, have only rarely entered the studies in a direct manner. Such factors influence the classroom reality in ways that go beyond the reality being produced by participants in practice only. Here, too, new approaches directly linking classroom reality to the surrounding contexts are needed.

A chapter such as this one cannot end without comments on the fact that international comparative studies attract a massive interest amongst politicians, media, and the general public. There is a clear tendency of these parties to summarize things in a manner that is "clear, brief, and wrong." This is on the boundary of involving misuse of the studies, but it is a misuse that is difficult to counteract by those involved in them. However, it would probably contribute to more balanced and fact-based debates if researchers undertook to engage in them to a larger extent than is typically seen.

Acknowledgments The authors would like to thank the reviewers, Michael Fried and Maitree Inprasitha, and above all the section editor, Jeremy Kilpatrick, for significant observations and constructive suggestions which have helped to improve this chapter.

References

Bauersfeld, H. (1979). Research related to the mathematical learning process. In B. Christiansen & H. G. Steiner (Eds.), *New trends in mathematics teaching* (Vol. 4, pp. 199–213). Paris, France: UNESCO.

Beaton, A. E., & Robitaille, D. F (1999). An overview of the Third International Mathematics and Science Study. In G. Kaiser, E. Luna, & I. Huntley (Eds.), *International comparisons in mathematics education* (Studies in Mathematics Education Series No. 11, pp. 30–47). London, UK: Falmer.

Becker, J. P. (Ed.). (1992). *Report of U.S.-Japan cross-national research on students problem solving behaviors*. Carbondale: Southern Illinois University at Carbondale. http://eric.ed.gov/PDFS/ED351204.pdf.

Becker, J. P., Sawada, T., & Shimizu, Y. (1999). Some findings of the US-Japan cross-cultural research on students' problem solving behaviours. In G. Kaiser, E. Luna, & I. Huntley (Eds.), *International comparisons in mathematics education* (Studies in Mathematics Education Series No. 11, pp. 121–139). London, UK: Falmer.

Clarke, D. J. (1998). Studying the classroom negotiation of meaning: Complementary accounts methodology. In A. Teppo (Ed.), *Qualitative research methods in mathematics education* (Journal for Research in Mathematics Education Monograph No. 9, pp. 98–111). Reston, VA: National Council of Teachers of Mathematics.

Clarke, D. J. (Ed.). (2001). *Perspectives on practice and meaning in mathematics and science classrooms*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Clarke, D. J. (2003). International comparative research in mathematics education. In A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick, & F. K. S. Leung (Eds.), *Second international handbook of mathematics education* (pp. 145–186). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Clarke, D. J. (n.d.). *The learners' perspective study: Research design.* Retrieved from 2011-04-10. http://extranet.edfac.unimelb.edu.au/DSME/lps/assets/lps.pdf.

Clarke, D. J., Emanuelsson, J., Jablonka, E., & Mok, I. A. C. (Eds.). (2006). *Making connections: Comparing mathematics classrooms around the world*. Rotterdam, the Netherlands: Sense Publishers.

Clarke, D. J., Keitel, C., & Shimizu, Y. (2006). The Learner's perspective study. In D. J. Clarke, C. Keitel, & Y. Shimizu (Eds.), *Mathematics classrooms in twelve countries: The insider's perspective* (pp. 1–14). Rotterdam, the Netherlands: Sense Publishers.

Cogan, L. S., & Schmidt, W. H. (1999). An examination of instructional practices in six countries. In G. Kaiser, E. Luna, & I. Huntley (eds.) *International comparisons in mathematics education* (Studies in Mathematics Education Series No. 11, pp. 68–85). London, UK: Falmer.

De Lange, J. (2007). Large-scale assessment and mathematics education. In F. K. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1111–1142). Charlotte, NC: Information Age.

Freudenthal, H. (1975). Pupils' achievement internationally compared: The IEA. *Educational Studies in Mathematics, 6*, 127–186.

Garden, R. A., Lie, S., Robitaille, D. F., Angell, C., Martin, M. O., Mullis, I. V. S., et al. (2006). *TIMSS advanced 2008 assessment frameworks*. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Husén, T. (Ed.). (1967). *International study of achievement in mathematics* (Vols. 1 & 2). Stockholm, Sweden: Almqvist & Wiksell.

Kaiser, G. (1997). Vergleichende Untersuchungen zun mathematikunterricht im englischen und deutschen Schulwesen [Comparative studies of mathematics instruction in English in the German school system]. *Journal für Mathematik-Didaktik*, *18*(2/3), 127–170.

Kaiser, G. (1999a). International comparisons in mathematics education under the perspectives of comparative education. In: G. Kaiser, E. Luna, & I. Huntley (Eds.), *International comparisons in mathematics education* (Studies in Mathematics Education Series No. 11, pp. 3–15). London, UK: Falmer.

Kaiser, G. (1999b). Comparative studies on teaching mathematics in England and Germany. In G. Kaiser, E. Luna, & I. Huntley (Eds.), *International comparisons in mathematics education* (Studies in Mathematics Education Series No. 11, pp. 140–150). London, UK: Falmer.

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* (Studies in Mathematics Education Series No. 11, pp. 86–103). London, UK: Falmer.

Leung, F. K. S., Graf, K. D., & Lopez-Real, F. J. (Eds.). (2006). *Mathematics education in different cultural traditions: A comparative study of East Asia and the West: The 13th ICMI Study*. New York, NY: Springer.

Martin, M. O., Gregory, K. D., & Stemler, S. E. (Eds.). (2000). *TIMSS 1999 technical report*. Boston, MA: International Study Center, Lynch School of Education, Boston College.

Martin, M. O., & Kelly, D. L. (Eds.). (1996). *TIMSS technical report: Vol. I. Design and development*. Boston, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Martin, M. O., Mullis, I. V. S., & Chrostowski, S. J. (Eds.). (2004). *TIMSS 2003 Technical report*. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, 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: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Niss, M. (2010). What is quality in a PhD dissertation in mathematics education? *Nordisk Matematik(k)Didaktik(k)/Nordic Studies in Mathematics Education, 15*(1), 5–23.

Niss, M., & Hoejgaard, T. (Eds). (2011). *Competencies and mathematical learning. Ideas and inspiration for the development of mathematics teaching and learning in Denmark* (English edition, October 2011, IMFUFAtekst no. 485). Roskilde, Denmark: Roskilde University, Department of Science, Systems and Models.

Niss, M., & Jensen, T. H. (Eds.). (2002). *Kompetencer og matematiklæring, Ideer og inspiration til udvikling af matematikundervisning i Danmark* [Competences and mathematics education: Ideas and inspiration for the development of mathematics education in Denmark] (Uddannelsesstyrelsens temahæfteserie nr. 18). Copenhagen, Denmark: Ministry of Education.

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). *TIMSS 2007 technical report*. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Organisation for Economic Co-operation and Development. (1999). *Measuring student knowledge and skills: A new framework for assessment*. Paris, France: Author.

Organisation for Economic Co-operation and Development. (2001). *Knowledge and skills for life: First results from the OECD Programme for International Student Assessment (PISA) 2000*. Paris, France: Author.

Organisation for Economic Co-operation and Development. (2003). *The PISA 2003 assessment framework: Mathematics, reading, science and problem solving knowledge and skills*. Paris, France: Author.

Organisation for Economic Co-operation and Development. (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris, France: Author.

Organisation for Economic Co-operation and Development. (2005). *School factors related to quality and equity. Results from PISA 2000*. Paris, France: Author.

Organisation for Economic Co-operation and Development. (2006). *Assessing scientific, reading and mathematical literacy. A framework for PISA 2006*. Paris, France: Author.

Organisation for Economic Co-operation and Development. (2007). *Science competencies for tomorrow's world*. Paris, France: OECD.

Organisation for Economic Co-operation and Development. (2010a). *What students know and can do: Student performance in reading, mathematics and science*. Paris, France: Author.

Organisation for Economic Co-operation and Development. (2010b). *PISA 2012 mathematics framework. To OECD, November 30, 2010* [Draft subject to possible revisions after the field trial]. Retrieved from http://www.pisa.oecd.org/dataoecd/8/38/46961598.pdf.

Postlethwaite, T. N. (1971). International Association for the Evaluation of Educational Achievement (IEA): The mathematics study. *Journal for Research in Mathematics Education, 2*, 69–103.

Robitaille, D. F., & Garden, R. A. (1989). *The IEA Study of Mathematics II: Contexts and outcomes of school mathematics*. Oxford, UK: Pergamon.

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

Robitaille, D. F., & Travers, K. J. (1992). International studies of achievement in mathematics. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 687–707). New York, NY: Macmillan.

Schmidt, W. H., McKnight, C., Valverde, G. A., Houang, R. T., & Wiley, D. E. (1997). *Many visions, many aims: A cross-national investigation of curricular intentions in school mathematics*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Schubring, G. (2008). The origin and early incarnations of ICMI. In: M. Menghini, F. Furinghetti, L. Giacardi, & F. Arzarello (Eds.), *The first century of the International Commission on Mathematical Instruction (1908–2008). Reflecting and shaping the world of mathematics education* (pp. 113–130). Rome, Italy: Instituto della Enciclopedia Italiana.

Shimizu, Y., Kaur, B., & Clarke, D. J. (2010). *Mathematical tasks around the world*. Rotterdam, The Netherlands: Sense Publishers.

Stevenson, H. W. (1999). The Case Study Project of TIMSS. In G. Kaiser, E. Luna, & I. Huntley (Eds.), *International comparisons in mathematics education* (Studies in Mathematics Education Series No. 11, pp. 104–120). London, UK: Falmer.

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

Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999, February). *The TIMSS Videotape Classroom Study: Methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States* (National

Center for Educational Statistics, Research and Development Report). Washington, DC: U.S. Department of Education. Retrieved from http://nces.ed.gov.

Stigler, J., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York, NY: Free Press.

Travers, K. J., & Weinzweig. A. I. (1999). The Second International Mathematics Study. In G. Kaiser, E. Luna, & I. Huntley (Eds.), *International comparisons in mathematics education* (Studies in Mathematics Education Series No. 11, pp. 19–29). London, UK: Falmer.

Travers, K. J., & Westbury, I. (1990). *The IEA Study of Mathematics II: Analysis of mathematics curricula*. Oxford, UK: Pergamon.