Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

Handbook of International Large-Scale Assessment

Background, Technical Issues, and Methods of Data Analysis

Edited by
Leslie Rutkowski
Indiana University
Bloomington, USA

Matthias von Davier
Educational Testing Service
Princeton, New Jersey, USA

David Rutkowski
Indiana University
Bloomington, USA

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20130923

International Standard Book Number-13: 978-1-4398-9512-2 (Hardback)

Library of Congress Cataloging-in-Publication Data

Handbook of International large-scale assessment : background, technical issues, and methods of data analysis / [edited by] Leslie Rutkowski, Matthias von Davier, David Rutkowski.
pages cm. -- (Statistics in the social and behavioral sciences series)
Includes bibliographical references and index.
ISBN 978-1-4398-9512-2 (hardback)
1. Educational tests and measurements--Cross-cultural studies. 2. Academic achievement--Cross-cultural studies. 3. Educational tests and measurements--Methodology. 4. Educational tests and measurements--Statistics. I. Rutkowski, Leslie, editor of compilation. II. Davier, Matthias von, editor of compilation. III. Rutkowski, David, editor of compilation.

LB3051.H31983 2013
371.26--dc23

2013025261

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Editors ix
Contributors xi

Section I Policy and Research Relevance of International Large-Scale Assessment Data

1. **A Brief Introduction to Modern International Large-Scale Assessment** 3
   David Rutkowski, Leslie Rutkowski, and Matthias von Davier

2. **International Large-Scale Assessments: From Research to Policy** 11
   Hans Wagemaker

3. **The Impact of International Studies of Academic Achievement on Policy and Research** 37
   Stephen P. Heyneman and Bommi Lee

Section II Analytic Processes and Technical Issues around International Large-Scale Assessment Data

4. **Assessment Design for International Large-Scale Assessments** 75
   Leslie Rutkowski, Eugene Gonzalez, Matthias von Davier, and Yan Zhou

5. **Modeling Country-Specific Differential Item Functioning** 97
   Cees Glas and Khurrem Jehangir

6. **Sampling, Weighting, and Variance Estimation in International Large-Scale Assessments** 117
   Keith Rust

7. **Analytics in International Large-Scale Assessments: Item Response Theory and Population Models** 155
   Matthias von Davier and Sandip Sinharay

8. **Imputing Proficiency Data under Planned Missingness in Population Models** 175
   Matthias von Davier

9. **Population Model Size, Bias, and Variance in Educational Survey Assessments** 203
Andreas Oranje and Lei Ye

10. **Linking Scales in International Large-Scale Assessments** 229
John Mazzeo and Matthias von Davier

11. **Design Considerations for the Program for International Student Assessment** 259
Jonathan P. Weeks, Matthias von Davier, and Kentaro Yamamoto

12. **Innovative Questionnaire Assessment Methods to Increase Cross-Country Comparability** 277
Patrick C. Kyllonen and Jonas P. Bertling

13. **Relationship between Computer Use and Educational Achievement** 287
Martin Senkbeil and Jörg Wittwer

14. **Context Questionnaire Scales in TIMSS and PIRLS 2011** 299
Michael O. Martin, Ina V. S. Mullis, Alka Arora, and Corinna Preuschoff

15. **Motivation and Engagement in Science around the Globe: Testing Measurement Invariance with Multigroup Structural Equation Models across 57 Countries Using PISA 2006** 317
Benjamin Nagengast and Herbert W. Marsh

16. **Contextual Indicators in Adult Literacy Studies: The Case of PIAAC** 345
Jim Allen and Rolf van der Velden

Section III Advanced Analytic Methods for Analyzing International Large-Scale Assessment Data

17. **Incorporating Sampling Weights into Single- and Multilevel Analyses** 363
Laura M. Stapleton

18. **Multilevel Analysis of Assessment Data** 389
Jee-Seon Kim, Carolyn J. Anderson, and Bryan Keller

19. **Using Structural Equation Models to Analyze ILSA Data** 425
Leslie Rutkowski and Yan Zhou

20. **Efficient Handling of Predictors and Outcomes Having Missing Values** 451
Yongyun Shin

21. **Multilevel Modeling of Categorical Response Variables** 481
Carolyn J. Anderson, Jee-Seon Kim, and Bryan Keller

22. **Causal Inference and Comparative Analysis with Large-Scale Assessment Data** 521
Joseph P. Robinson

23. **Analyzing International Large-Scale Assessment Data within a Bayesian Framework** 547
David Kaplan and Soojin Park

24. **A General Psychometric Approach for Educational Survey Assessments: Flexible Statistical Models and Efficient Estimation Methods** 583
Frank Rijmen, Minjeong Jeon, Matthias von Davier, and Sophia Rabe-Hesketh

Index 607

Tatto, M. T., Schwillie, J., Senk, S. L., Ingvarson, L., Rowley, G., Peck, R., Bankov, K., Rodirguez, M., Rekase, and Rekase, M. 2012. Policy, Practice and Readiness to Teach Primary and Secondary Mathematics in 17 Countries: Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-M) IEA, Amsterdam, The Netherlands.

Tuijnman, A. C. and Postlethwaite, N. 1994. *Monitoring the Standards of Education.* Oxford: Pergamon.

United States National Commisson on Excellence in Education. 1983. *A nation at Risk: The Imperative for Educational Reform: A Report to the Nation and the Secretary of Education.* Washington, DC: United States Department of Education.

Wagemaker, H. 2010. IEA: Globalization and assessment. In: *International Encyclopaedia of Education*, ed. E. B. Penelope Peterson, Vol. 4, 663–668. Oxford: Elsevier.

Wagemaker, H. 2011. IEA: International studies, impact and transition. In: *IEA 1958–2008:50 Years of Experiences and Memories*, eds. C. Papanastasiou, T. Plomp and E. Papanstasiou, 253–273. Nicosia: Cultural Center of the Kykkos Monastery.

Weiss, C. H. 1999, October. The interface between evaluation and public policy. *Evaluation* 5(2): 468–486.

World Bank. 2011a. *Learning for all: Investing in People's Knowledge and Skills to Promote Development: Education Sector Strategy 2020.* Washington, DC: Author.

World Bank. 2011b, May 18. *The Arab Regional Agenda on Improving Education Quality (ARIEQ)* (working paper): Tunis: Author.

Wyckoff, P. 2009. *Policy and Evidence in a Partisan Age: The Great Disconnect.* Washington, DC: Urban Institute Press.

3

The Impact of International Studies of Academic Achievement on Policy and Research

Stephen P. Heyneman
Vanderbilt University Peabody College

Bommi Lee
Vanderbilt University Peabody College

CONTENTS

Introduction 38
Debates over Utility and Feasibility 40
Impact on Practice 42
Pedagogy 42
Teacher Training 43
Class Size 43
Hours of Instruction 44
Use of Technology 44
Subject-Specific Findings: Math 45
Subject-Specific Findings: Science 46
Subject-Specific Findings: Reading and Writing 47
Subject-Specific Finding: Civic Education 48
Impact on Specific Regions 49
Sub-Saharan Africa 49
Latin America 49
Impact on Specific Countries 51
Use of Surveys to Generalize about School System Structures 56
Use of Surveys to Uncover More Specific Areas for Investigation 56
Use of Surveys to Develop New Techniques and Discover New Variables 57
Smaller, Quicker, and Cheaper: Improving Learning Assessments for Developing Countries 58
Impact on Research Findings 59
Appendix 64
References 65

Introduction

It began as an educational experiment. In the late 1950s, Torsten Husen from the University of Stockholm was visiting his friends Benjamin Bloom and C. Arnold Anderson at the University of Chicago. "Why don't we test for academic achievement internationally," he was reported to have asked, "The world could be our laboratory" (Heyneman and Lykins 2008, p. 106). This was the origin of the International Association for the Evaluation of Education Achievement (IEA),* an organization that now includes 68 countries (see the chapter appendix) and assists in the testing of a half-dozen academic subjects including foreign languages, reading literacy, mathematics, science, civics, and writing. The IEA commenced as a loose association of university-based personalities interested in finding solutions to pedagogical and other problems. It sought solutions that could not be found locally and it expanded to include international agencies and a dozen different sources of funding. From the first "mom-and-pop shop" studies in the 1960s, international testing is now conducted not only by the IEA, which operates globally by regional organizations in the industrial democracies (the Organization for Economic Cooperation and Development, OECD), and others in Africa and Latin America. A list of these international testing efforts can be found in Table 3.1. As one can see, they have become more frequent and more global.

Beginning with the wealthier countries, the IEA now includes many middle- and even low-income countries. In the period from 1960 to 1989 there were 43 international surveys of academic achievement. In the period between 1990 and 1999 there were 66 international surveys, 49 regional surveys, and 205 local or national surveys. Between 2000 and 2009 there were 152 international surveys, 47 regional surveys, and 324 national or local surveys. The percentage of countries participating in at least one of the three types of academic achievement surveys includes 33% of countries in Europe and Central Asia, 50% in Sub-Saharan Africa and in the Arab States, 60% in East Asia and the Pacific, and 74% in Latin America and the Caribbean (Kamens 2013).

One observer concludes:

> The search for best practices is now an international mantra with extraordinary legitimacy and funding ... Comparison is viewed not only as possible but required for advancing knowledge (Kamens, forthcoming).

In the new education sector policy paper, the World Bank has called for funding to be used to support the systematic testing of children in all parts of the world (World Bank 2011). How did the world get from the tiny

* Ben Bloom was once asked: "Why was the long title reduced to just three letters—IEA." He responded, "Why not? Where is the rule that an acronym has to be the exact replica of an organization's title?"

TABLE 3.1

International Tests of Educational Achievement: Scope and Timing

Sponsor	Description	Countries	Year(s) Conducted
IEA	First International Mathematics Study (FIMS)	12 countries	1964
IEA	Six Subjects Study		1970–1971
	Science	19 systems	
	Reading	15 countries	
	Literature	10 countries	
	French as a foreign language	8 countries	
	English as a foreign language	10 countries	
	Civic Education	10 countries	
IEA	First International Science Study (FISS; part of Six Subjects Study)	19 countries	1970–1971
IEA	Second International Mathematics Study (SIMS)	10 countries	1982
IEA	Second International Science Study (SISS)	19 systems	1983–1984
ETS	First International Assessment of Educational Progress (IAEP-I, Mathematics Study and Science)	6 countries (12 systems)	1988
ETS	Second International Assessment of Educational Progress (IAEP-II, Mathematics and Science)	20 countries	1991
IEA	Reading Literacy (RL)	32 countries	1990–1991
IEA	Computers in Education	22 countries 12 countries	1988–1989
Statistics Canada	International Adult Literacy Survey (IALS)	7 countries	1994
IEA	Preprimary Project:		
	Phase I	11 countries	1989–1991
	Phase II	15 countries	1991–1993
	Phase III (longitudinal follow-up of Phase II sample)	15 countries	1994–1996
IEA	Language Education Study	25 interested countries	1997
IEA	Third International Mathematics and Science Study (TIMSS)	45 countries About 40	1994–1995 1997–1998
	Phase I		
	Phase II (TIMSS-R)		
IEA	Civics Education Study	28 countries	1999
OECD	Program for International Student Assessment (PISA)	43 countries 41 countries 57 countries 65 countries	2000 (reading) 2003 (math) 2006 (science) 2009 (reading)
IEA	Progress in International Reading Literacy Study (PIRLS)	34 countries 41 countries 48 countries	2001 2006 2011
IEA	Trends in International Mathematics and Science Study (TIMSS)	45 countries 48 countries 63 countries	2003 2007 2011

Source: Adapted from Chromy, R.R. 2002. *Methodological Advances in Cross-National Surveys of Educational Achievement* (pp. 80–117). Washington, DC: National Academy Press.

experiment in the 1960s to today's "mantra"? What have the objections been to these surveys? To what extend have they had an effect on education policy and research? That is the subject to which we now turn.

Debates over Utility and Feasibility

In 1933, an argument broke out between M. Dottrens and Jean Piaget at a board meeting of the International Bureau of Education (IBE) in Geneva.* Dotterns proposed an international survey to record what countries were doing in education. Piaget was against it. He said: *"L'experience nous a montre qu'il est extremenent difficile d'estabir des tableaux statistiques comparables"* (In our experience it is extremely difficult to establish a table of comparable statistics; Smyth 1996, p. 4). Piaget had a point. At that time there was no common definition on what education meant, on how schooling might differ from ad hoc learning, or on how to distinguish education levels. The meaning of vocational and general education varied both between and within nations. There were 115 different ways to define literacy and 133 different ways to classify educational attainment by age group (Smyth 2005, p. 15). Dotterns, however, eventually won the argument. In spite of the many procedural complexities and the very real danger of receiving misleading results, the demand to know "what countries were doing" in the field of education was simply irresistible.

These same arguments have reoccurred with regularity in the interim. Though those who agree with Dottrens have lost many battles, they have all but won the war. The record of advancement in geographical coverage and qualitative depth in comparative education statistics has been unidirectional. From defining and counting schools in 1933 to videotaping teaching styles and capturing unit expenditures, the story of educational measurement and the resulting debates over its unprecedented findings is one true sign that there has been progress in education research. This extraordinary growth, both in the quantity and the quality of education data has brought fresh—and sometimes contentious—insights into perennial questions concerning educational achievement (Heyneman and Lykins 2008).

The modern-day equivalent of Piaget's doubt about the feasibility of counting schools in different countries is the criticism of international achievement testing. In the 1970s and 1980s objections included the characterization of cross-national achievement tests in developing countries as tantamount to a form of neocolonialism (Weiler 1994). The notion that developing countries were inherently inferior in education achievement helped stimulate regional efforts in Southern Africa and Latin America. Recent studies, however, have suggested that the results from some developing countries are better than

* The International Bureau of Education was established in 1925 as a nongovernmental education organization. In 1929, it allowed countries to join as members. Both Dotterns and Piaget were educational psychologists. Dotterns served on the Board of Directors; Piaget served as IBE's director for 40 years. Today, the IBE is a specialized agency within the United Nations Educational, Scientific, and Cultural Organization (UNESCO).

wealthier countries in both average achievement and in the efficiency of delivery (Heyneman 2004).

In the early 1990s, criticisms of international testing contained the argument that countries sent lower portions of their age cohorts to secondary school; hence higher scores in Germany, for instance, were based on biased samples (Rotberg 1990). This argument precipitated a formal reply from the members of the Board on International and Comparative Studies in Education of the National Academy of Sciences (BICSE). The members of the BICSE committee pointed out that while countries indeed select some to attend specialized secondary schools in preparation for university, sampling was purposefully random across all categories of schools; hence higher scores from countries such as Germany could not be attributed to a bias in sampling (Bradburn et al. 1991).

BICSE then laid out the virtues of international achievement testing:

- It may provide a wider laboratory on which to observe the consequences of different domestic policies and practices.
- That by studying the challenges and successes of other parts of the world, international test information helps define what is realistic in terms of domestic education policy.
- It may introduce concepts that have been overlooked domestically.
- It may raise important questions or challenge long-held assumptions that may not have been challenged using domestic sources of information alone.
- It may elicit results that were not anticipated but nevertheless have high value (Bradburn and Gilford 1990).

Once arguments over the technical methods of test administration declined, others quickly took their place. These included the probability that the public's expectations about what can be learned from international tests are unrealistic and the notion that other countries are better is based on a set of myths. These myths include whether tests provide valid measures of quality; whether the quality of education in the United States has declined; whether the problem in schools can be fixed with the introduction of new tests; and whether new testing can compensate for inadequate resources (Rotberg 1996, 2006, 2007, 2008; Ravitch 2010).

In addition to these doubts as to the utility of testing to inform policy, there are more specific objections to the nature of the test items, whether they reasonably guarantee equivalency across countries when students speak, read, write, and listen using entirely different systems of communication (Holliday and Holliday 2003); or whether local schooling situations can actually be compared (Fensham 2007). The assets and drawbacks of international assessments have been well summarized by Rowan (2002) and by Braun and Kanjee (2006).

Impact on Practice

Despite the debates over the utility of international assessments, experience suggests that they have often had an effect on domestic education policies, although the magnitude of impact differs. The range of these policies is wide, from pedagogy to national standards of assessment. In this section, we discuss some of these findings and how they influenced specific policies.

Pedagogy

The U.S. TIMSS 1999 Video Study, which examined instructional practices across countries, revealed that classroom instructions in Japan were significantly different from those in the United States (Stigler and Hiebert 1999; Wagemaker 2004). The findings suggested that Japan had the most skillful and purposeful teaching, with students being asked to solve challenging problems. In the United States, teaching tended to focus on learning terms and demonstrating procedures (Stigler and Hiebert 1999). Also, the curriculum content in Japan was more coherent. In the United States teachers had fewer opportunities to work with other teachers to improve teaching (Wagemaker 2004), whereas in Japan teachers were more collaborative and saw themselves as contributing to the knowledge of teaching practices and their own professional development (Stigler and Hiebert 1999). This result grabbed attention in the mathematics education community and the public at large (Park 2004). On the other hand, what the TIMSS videos did not reveal was the effect of Japanese Juku (private tutoring) schools either on achievement or on the pedagogy in public schools. Because the pedagogy of private tutoring is rigorously focused on the content of the university entrance examination, it is quite possible that, relieved of being solely responsible for student success on the examinations, public school teachers could "afford" to teach with more creativity. American teachers, by contrast, do not have such "luxury."

In England, the results of TIMSS 1995 showed that Grade 4 students used calculators in mathematics lessons more frequently than those in other countries. Although there was no association between frequency of calculator use and mean mathematics scores at Grade 4, this result from the teacher survey drew attention to the frequency of calculator use. Consequently, England began to emphasize the importance of mental calculation (Keys 2000).

Policy makers in some developing countries began to model teaching practices on developed countries after TIMSS. Eight among eighteen developing countries that participated in TIMSS-Repeat planned to change the direction of classroom instructions to emphasize activity-based learning, problem solving in mathematics, and more critical thinking in science (Elley 2005). In two countries, policy makers planned to reduce the amount of teacher lecturing and increase student engagement in lessons, with more discussion, questioning, experimenting in class, critical thinking, valuing of student opinion, and exploring students' curiosity (Elley 2005).

Not only teaching practices, but also national curricula were changed as results of TIMSS and PISA. In many countries, new topics were added because findings showed that higher-performing countries generally had greater levels of coverage of the intended curricula (Mullis and Martin 2007). Romania and the Slovak Republic added new topics in mathematics and science curricula as a result of TIMSS (Noveanu and Noveanu 2000; Berova and Matusova 2000). In Sweden, PISA results contributed to the introduction of national tests in biology, chemistry, and physics for 12- and 15-year-olds, as well as the development of diagnostic material in science for the younger ages (Breakspear 2012). In the Slovak Republic, key competencies characterized by PISA, which were not included in the previous curriculum, were incorporated into national curriculum standards (Breakspear 2012).

Teacher Training

TIMSS results led to changes in the existing preservice and in-service teacher training programs as well as development of instructional resources in some countries. In Canada (Ontario), instructional materials were developed as a direct outcome of TIMSS (Taylor 2000). New Zealand also developed resources materials and created professional development programs for mathematics and science teachers to address the areas of relative weakness (Chamberlain and Aalst 2000). In Macedonia, changes were made to in-service training, which encouraged the teachers to change from a teacher-centered to a more student-centered style (Elley 2005). In Israel, findings indicated that they should focus on teacher preparation programs instead of in-service training programs (Mevavech 2000). In Malaysia, the government paid more attention to teaching practices and teacher training after the TIMSS-R study (Elley 2005).

Class Size

Contrary to common belief, findings from ILSA have consistently shown that class size had little association with higher achievement. The IEA's First Mathematics Study and the Second Science Study showed that class size was not related to achievement (Medrich and Griffith 1992). More recent studies using PISA and TIMSS with school-fixed effects and instrumental variables also presented no statistically significant class size effect in most of the countries (Fertig and Wright 2005; Woessmann 2005a,b; Ammermueller et al. 2005; Woessmann and West 2006). While it is irrational to suggest that larger classes are better, these findings suggest that simply lowering class size without making other changes may not result in achievement gains.

Hours of Instruction

TIMSS results were not conclusive about whether more instruction time is related to higher achievement. Early IEA studies showed that the students had higher achievement when there was a greater amount of time spent on teaching and learning (Husen 1967; Anderson and Postlethwaite 1989; Keeves 1995). However, the Second Science Study showed that hours of school each week, hours of mathematics instruction each week, and hours of homework had virtually no relationship to the achievement (Medrich and Griffith 1992). In addition to this finding, TIMSS 2003 results showed that countries that provided more instruction time (the Philippines, Indonesia, and Chile) had relatively low achievement (Mullis and Martin 2007), a conclusion that can sometimes be a topic of local political interest (Heyneman 2009a). Related to the number of school days/year and hours/day is the issue of how much classroom time is actually devoted to the task of learning. Since the First Mathematics Study, the "opportunity to learn" variable emerged as an important indicator of performance, especially at the secondary level (Medrich and Griffith 1992). These findings provided insight into the relationship between the "intended curriculum," the "implemented curriculum," and the "achieved curriculum" (Wagemaker 2004).

Use of Technology

Both the IEA and OECD PISA study provide information on Information Communication Technology (ICT) use in education. The IEA has a long tradition of studying ICT in education (Pelgrum and Plomp 2008). The first one was the so-called Computer in Education (CompEd) study that was conducted in 1989 and 1992. The findings implied that greater attention should be given to how computers can be used in schools. The second wave of ICT large-scale international assessment was the IEA's Second Information Technology in Education Studies (SITES), which started with Module One in 1998 to 1999. The SITES Module One examined how ICT affects teaching and learning processes in different countries. Wagemaker (2004) notes that the first school survey, Module One of SITES, raised three issues—(i) the challenge of making teachers ready-to-use technology in their instruction; (ii) over-expectations suggesting that computers would transform the curriculum and pedagogy; and (iii) the difficulties that schools have in internet access and protecting children from inappropriate materials. Module One was followed by Module Two, which is a collection of qualitative case studies on ICT-supported pedagogical innovations, and findings were used in ongoing discussions in several countries (Kozma 2003; Wagemaker 2004). SITES 2006 is the most recent survey on ICT conducted by the IEA.

Nations such as Austria that introduced computer technologies in a systematic way with universal access, teacher training, and curriculum integration out-performed countries that added computers in an ad-hoc fashion without regard to standardization and planning (Keeves 1995). Perhaps as a result, the main area of interest in the studies of CompEd and three SITES studies shows a shift of focus from computer counts, access rates, and obstacles to how pedagogical practices are changing and how IT supports these practices (Pelgrum and Plomp 2008). TIMSS 2003 and PISA 2003 also have indicators of ICT availability and ICT use on core school subjects; however, these assessments cover only a small number of indicators (Pelgrum and Plomp 2008).

Despite the rising significance of using technology in education, research shows inconsistent but often negative findings on the use of technology and achievement. Although some IEA reports show mathematics achievement to be positively associated with computer usage (Mullis and Martin 2007), later investigations found a negative or no relationship (Fuchs and Woessmann 2004; Papanastasiou et al. 2004; Wittwer and Senkbeil 2008). Still, others found mixed results within the same study (Papanastasiou et al. 2003; Antonijevic 2007) and sometimes results differed according to the kind of software students used. More recent studies have suggested that it is not the computer that influences problem-solving skills, but the purpose to which the computer is put. Using computers to download games and videos, for instance, may adversely affect problem-solving skills. On the other hand, using computers to seek and analyze new information augments problem-solving skills (de Boer 2012).

Subject-Specific Findings: Math

As fluency in Latin was the mark of an educated person and a prerequisite for employment throughout Europe in the past, mathematics is in many key respects the new Latin (Plank and Johnson 2011). Mathematics has become the key indicator for educational success (Plank and Johnson 2011), as a country's best students in mathematics have implications for the role that country will play in the future advanced technology sector and for its overall international competitiveness (OECD, 2004a). Mathematics is also straightforward in comparing across countries, while comparing performance in other subject matters (literacy, history, science) is complicated by cross-national differences in language and culture. Mathematics provides a ready yardstick for international comparisons (Plank and Johnson 2011). Most findings from international studies use mathematics achievement as a dependent variable.

The focus of earlier IEA studies was observing whether differences in selective and comprehensive systems lead to different outcomes (Walker 1976). The Six Subjects Study found that students in academic courses in selective schools attained higher levels in mathematics than students following similar courses in comprehensive schools (Walker 1976). This suggests that selective school systems may have higher outcomes than comprehensive school systems.

Results from the TIMSS 2003 and 2007 supported some common beliefs about factors for high achievement. Findings indicated that high-performing

students had parents with higher levels of education, and had more books at home. High achievers also had computers at home. Findings from TIMSS 2007 also indicated that most high-performing students spoke at home the same language used for the test, and generally had positive attitudes and self-confidence. School factors were also important for high achievers. Generally, high-achieving students were in schools with fewer economically disadvantaged students and the school climate was better. These schools had fewer problems with absenteeism, missing classes, and students arriving late. Both TIMSS 2003 and 2007 findings show that being safe in school was positively associated with math achievement.

The findings from PISA are similar to the findings from TIMSS. In terms of learning behavior, the general learning outcomes from PISA 2000 presented that those who process and elaborate what they learn do better than those who memorize information (OECD and the United Nations Educational, Scientific, and Cultural Organization [UNESCO] 2003). The results also showed that cooperative learning is not necessarily superior to competitive learning, and evidence suggests that these strategies are complementary. For achievement among subgroups, PISA 2000 results showed that there was a significant number of minority students in all countries who displayed negative attitudes toward learning and lack of engagement with school, which is associated with poor performance (OECD and UNESCO 2003). PISA 2000 results also supported the fact that spending is positively associated with mean student performance. However, spending alone is not sufficient to achieve high levels of outcomes and other factors play a crucial role (OECD and UNESCO 2003).

From the results of PISA 2003, two major findings emerged. Some of the best-performing countries showed only a modest gap between high and low performers, suggesting that wide disparities are not a necessary condition for a high level of overall performance (OECD 2004a). Additionally, student performance varied widely between different areas of mathematical content (OECD 2004a), implying that balance and sequence of the national curriculum may be independently important to high achievement.

Subject-Specific Findings: Science

Science is another important area of concern for policy makers, as it is a subject associated with national economic growth in an information, communication, and knowledge-based society. Early studies showed the need for highly qualified teachers in science (Comber and Keeves 1973; Walker 1976). The findings from the first IEA study showed that when teachers specialized in science, had received more post-secondary education, or had participated in science curricular reforms, their students tended to perform better on science achievement tests (Comber and Keeves 1973; Walker 1976). Implications from TIMSS studies were that competent and committed science teachers are

needed for successful teaching of science in secondary schools (Walker 1976; Vlaardingerbroek and Taylor 2003).

Another finding from the IEA Six Subject Study was that achievement was closely linked to opportunity to learn, because students in developing countries found the tests to be much more difficult than their peers elsewhere (Walker 1976). In fact, PISA 2000 results also showed that students' performance in low- and middle-income countries was lower than that of high-income countries (OECD and UNESCO 2003). This raises attention to the gross differences in educational quality between OECD and developing countries (Heyneman 2004).

The gender gap in science was reduced over time. TIMSS 1995 results showed that, in most countries, boys performed significantly higher in science, particularly in physical science, than girls in both seventh and eighth grades (Mullis et al. 1996). Not only did boys perform better, but they also expressed a liking for this content area more often than girls (Mullis et al. 1996). However, TIMSS 2003 results showed that in the eighth grade there was greater improvement for girls than boys since 1999 (Mullis et al. 2004). This suggests a closed gender gap in science achievement. Average science achievement also improved for both boys and girls since 1995 (Mullis et al. 2004).

TIMSS 2003 results also presented some factors that were associated with high performance across countries. Higher science achievement was associated with having at least moderate coverage of the science topics in the curriculum, although high coverage in the intended curriculum itself does not necessarily lead to high student achievement. School climate and safety was also strongly associated with higher science achievement. Principal's perception of school climate was also highly related to average science achievement (Mullis et al. 2004).

Access to computers in science classrooms remained a challenge in many countries. Teachers reported that, on average, computers were not available for 62% of eighth grade students and 54% of fourth grade students internationally (Mullis et al. 2004). Even in countries with computers available in schools, using computers in science class was rare in eighth grade (Mullis et al. 2004). This finding raises concerns about how well schools and teachers are prepared in using computers for pedagogical purposes, particularly in science.

Subject-Specific Findings: Reading and Writing

Perhaps because families influence their children through language, the findings from studies of reading achievement showed stronger home effects. There is a lack of explanatory power of school and classroom-based measures in accounting for the differences in reading achievement between students (Keeves 1995). Mother-tongue instruction beyond an early grade did not seem to advance reading comprehension skills.

There were some significant changes in PISA results since 2000. First, reading scores rose significantly in 13 countries and fell in four others, indicating that countries, in general, are increasing students' achievement. Second, improvement in results is largely driven by improvements at the bottom end of distribution, indicating that there was progress in improving equity (OECD 2010). The percentage of immigrant students increased, but the reading performance gap between students with and without an immigrant background narrowed (OECD 2010). On the other hand, gender differences stayed or widened (OECD 2010). One interesting finding among the countries that improved fastest—Chile, Peru, Albania, and Indonesia—is that the relationship between socioeconomic background and learning outcomes has weakened (OECD 2010).

Gender differences were larger in reading than in mathematics achievement. The PISA 2000 findings indicated a growing problem for males, particularly in reading literacy. In mathematics, females on average remained at a disadvantage in many countries, but this was due to high levels of performance of a comparatively small number of males (OECD and UNESCO 2003). Education systems have made significant efforts toward closing the gender gap; however, much remains to be done. Improving reading skills for males and stimulating self-concept among females in mathematics need to be major policy objectives if gender equality in educational outcomes is to be achieved (OECD and UNESCO 2003).

There are not many large-scale assessments of writing, and perhaps the IEA study of written composition is the only one. The study found that there was differential treatment in most school systems between boys and girls in the provision of writing instruction and in the rating of writing performance. Girls profited from differential treatment in most school systems, where mostly women provide instruction (Purves 1992). This, again, shows that there is a gender gap in writing performance as in reading performance. The problem with international writing tests is that the definition of excellence is heavily influenced by culture. This makes it difficult to compare achievement across cultures.

Subject-Specific Finding: Civic Education

The IEA conducted its first study on civic education in 1971. The civic education measures civic knowledge, support for democratic values, support for the national and local government, and participation in political activities. The Second Civic Education Study (CIVED) was carried out in 1999, when many countries were experiencing political, economic, and social transitions. An additional survey of upper secondary students was undertaken in 2000. The findings indicated that schools appear to be an important factor responsible for civic knowledge and engagement. Schools are also responsible for the content of the curriculum and offer places for students to practice democracy as much as they offer places for learning facts (Amadeo et al. 2002).

In 2009 the IEA conducted a study of civic and citizenship education (ICCS) in 38 countries that built upon previous IEA studies of civic education. The study took place because of significant societal changes including rapid development of new communication technologies, increased mobility of people across countries, and the growth of supranational organizations (Schulz et al. 2010). There was a decline in civic content knowledge since 1999, students were more interested in domestic political issues rather than foreign issues, and the levels of interest were greatly influenced by parents (Schulz et al. 2010). The findings from the ICCS also provide evidence that although students' home backgrounds largely influence students' civic knowledge, schools also can contribute to civic knowledge and intentions to vote in adulthood (Schulz et al. 2010).

Impact on Specific Regions

Sub-Saharan Africa

The Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) began in 1995 with a meeting of 15 Ministers of Education.* Coordinated by the International Institute of Education Planning (UNESCO/Paris) data on mathematics and reading were collected in 2000 and 2007. Publications have included national policy briefs, national reports, working documents, working papers, policy issues, and technical reports. The Programme d'analyse des systems educatifs des pays del la CONFEMEN (PASEC) is roughly the equivalent in 17 French-speaking countries.† Tests have been applied in French, mathematics, and occasionally in the national language at the beginning and end of the school year so as to capture value-added information. More than 500 reports are available online, which detail the methods and the results.

Latin America

UNESCO and UNICEF help coordinate the regional Laboratorio Latin americano de Evaluation de la Calidad de la Educación and its Monitoring of Learning Achievement (MLA) studies, an illustration of how much the region has changed. In the 1960s and 1970s international tests of academic achievement were considered to be "northern" threats to Latin American autonomy. Today, Latin American plays a leading role in the use and interpretation of

* Members include Botswana, Lesotho, Kenya, Malawi, Mauritius, Seychelles, South Africa, Swaziland, Tanzania, Tanzania-Zanzibar, Uganda, Zambia, and Zimbabwe.

† Members include Mauritania, Cape Verde, Senegal, Guinea Bissau, Guinea, the Ivory Coast, Togo, Benin, Burkina-Faso, Niger, Central African Republic, Cameroon, Congo (Brazzaville), Gabon, Madagascar, Comoros, the Seychelles, Mauritius, Djibouti, and Burundi.

international tests. Perhaps the most important stimulant has been the influence of the Partnership for Educational Revitalization in the Americas (PREAL). Since 1995, PREAL has worked to improve the quality and equity of public education in the region. It explains itself as

> A network of private institutions and individuals dedicated to changing the way public and private leaders in Latin America think about schools and education. It seeks to improve learning and equity by helping to promote informed debate on education policy, identify and disseminate best practice, and monitor progress toward improvement. PREAL encourages business and civil society to work with governments in a common effort to improve schools and strengthen their capacity to do so. (PREAL 2011)

PREAL has been responsible for publishing "Report Cards," which give grades to countries for their educational performance. As of 2011, 4 regional, 21 national, and 6 sub-national report cards have been issued; each has stimulated alterations in education policy and practice. Initiatives have stimulated the minister of education in Peru and Honduras to establish national education standards and have been used as a guide to new standards in Chile. Recommendations from PREAL's report cards and "Policy Audits" have been incorporated into education policy in El Salvador, Panama, and the Dominican Republic. PREAL's demand for accountability has directly influenced the orientations of education ministers in Jamaica, Brazil, and Mexico City. To promote informed debate on education policy, PREAL published a report on how Latin America and the Caribbean performed on the Second Regional Student Achievement Test (SERCE; Ganimian 2009) and the regional results from PISA 2009 (Ganimian and Rocha 2011). The results of PISA 2009 were used in spurring public discussion of education challenges in the region, and informed policy makers of specific issues that require attention. PREAL also strongly urged countries to participate in global achievement tests. As a result, El Salvador participated in the TIMSS 2007 for the first time, and the Panamanian minister of education was also convinced to participate in the 2009 PISA exam for the first time (PREAL 2011). Thus, the PREAL used the PISA results to bring educational issues to policy makers and also urged them to participate in the global tests. Since 1995, PREAL has been responsible for publishing three books on testing, 58 working papers, 28 educational synopses (short policy summaries), 29 policy series, and 32 best-practice documents in English, Spanish, or Portuguese. The Working Group on Standards and Assessment (GTEE) alone has been responsible for 16 publications. No organization in any other region seems to have had the impact of PREAL on both policy and practice. Today the Latin American region is in an advanced state of assessment use, innovation, and implementation. This use of education data to underpin transparency and accountability has occurred in spite of the fact that students in the region tend to learn less and its systems appear more inefficient than many other parts of the world. It is not irrelevant to note that the emphasis on educational accountability has occurred in parallel fashion with the shift from authoritarian dictatorships to democracy. Latin America, and the influence of PREAL specifically, is an illustration of the power voters have over the existence of information on which to judge their school systems. Without assessments, political leaders can claim excellence without reference to the facts; with assessments, claims of excellence require proof based on international standards of evidence.

Impact on Specific Countries

In England, the first TIMSS results were announced at a time when there were concerns about standards in education, and it was in 1997 that the policy makers had the incentives to make change because of the new government. Shortly after the results were announced, what was then the Schools Curriculum and Assessment Authority (SCAA), former Qualifications and Curriculum Development Agency (QCDA), investigated the TIMSS data to provide more information about students' strengths and weaknesses in relation to the national curriculum and its assessment (Shorrocks-Taylor et al. 1998).

Whereas TIMSS results stimulated changes in assessment in England, PIRLS 2001 results confirmed that their national education strategy worked well. England introduced the National Literacy Project in 1996, which was a large-scale pilot of a new approach to literacy teaching in the first 6 years of schooling (Twist 2007). One example of this new project was whole-class and highly interactive teaching (Twist 2007). However, this national strategy was not evaluated internationally until PIRLS 2001 was conducted. PIRLS 2001 was the first international survey to include children in England who spent most of their primary education being taught using the approaches detailed in the National Literacy Strategy (Twist 2007). Results showed that their performance was significantly better than other participating English-speaking countries (Twist 2007). England's relative success in PIRLS 2001 was seen as endorsement of an innovative national strategy (Twist 2007).

Both the government and the unions in England used PISA 2006 to support their arguments. The unions lobbied against the government's argument favoring the expansion of different types of schooling by using PISA findings, which suggested that countries with a school system segmented by vocational and academic institutions do not perform better. The government also used the PISA findings to create the Masters in Teaching and Learning (MTL) program, which is to be tested in three universities in the North West of England (Figazzolo 2009). Today, interpreting the results of international tests are a daily activity in the National Education Ministry (Heyneman 2009b).

In Iceland, it was already decided to increase the number of national examinations at the elementary level, but after the first results of TIMSS were made public, they decided to test for science achievement in Grade 10 (Gudmundsson 2000).

In Switzerland, prior to PISA there were efforts to harmonize the different education standards in each canton and test student performance on a regular basis (Bieber and Martens 2011). However, only after PISA results came out did they start to concretize the reform. PISA revealed unexpected shortcomings in reading competencies and also showed that Switzerland was one of the OECD countries where the influence of students' socioeconomic backgrounds on their reading skills was the most pronounced (OECD 2002). Education became top of the political agenda and reform pressures increased. PISA played a vital role in the proceeding high degree of convergence of policies (Bieber and Martens 2011).

In Ireland, PISA 2003 demonstrated that the national standard of mathematics education had to improve. With its focus on science, PISA 2006 served to expedite changes in the curriculum for lower secondary education that had been initiated in 2003 (Figazzolo 2009).

In Iran, changes were initiated by the Mathematics Curriculum Committee on the basis of the TIMSS results. TIMSS have been used in the development of the first national item bank for primary education in Iran. TIMSS items formed the framework for the development of tables of specification and items. Another new initiative was that they sponsored a national research project where they used TIMSS curriculum framework in developing test items (Kiamanesh and Kheirieh 2000).

Compared to TIMSS, the PIRLS 2001 results were more quickly disseminated and acted on, as the students did not perform well overall in PIRLS. The policy makers emphasized the ongoing effort, which should be directed toward sound language learning, especially for those whose mother tongue is not Farsi. Consequently, several long-term programs were developed for primary education. Also, the Organization for Research and Educational Planning has incorporated suggestions from the analysis of the PIRLS 2001 results for the future policies and programs (Karimi and Daeipour 2007).

In Latvia, the TIMSS result stimulated a significant step to improve their educational curriculum. The TIMSS results showed that Latvian students did not know how to apply acquired knowledge to real-life situations, which was a rather surprising result to them as the existing school examinations did not inform them of this fact. Thus, in 1997 they formed the new national compulsory education standard and established a centralized examination system for secondary education graduation examinations (Geske and Kangro 2000).

In Russia, TIMSS results contributed to the development of new educational standards in mathematics and science. Russia had no tradition in using standardized assessments in school practice; however, as a consequence of TIMSS, Russia began to use standardized tests for assessing student achievement (Kovalyova 2000).

In Germany, the mediocre performance in the 1995 TIMSS study came as a shock to German teachers, researchers, and policy makers. The results were confirmed by the PISA 2000, which revealed average performance and a large disparity between the federal states within Germany (Neumann et al. 2010). Before the PISA study, the German education system did not traditionally rely on standardized testing. However, after the low performance, particularly in science achievement, German policy makers decided for a major reform, which included the introduction of national education standards (NES). The NES were framed by the PISA framework and the particular deficits of German science education identified by PISA (Neumann et al. 2010).

PISA attracted more interest than PIRLS in Germany because the results were unexpected. However, PIRLS also indicated a clear picture of what is going on in German schools (Schwippert 2007). For example, the PIRLS data showed that one in every five of the German students participating in the study was from a family where at least one parent was not born in Germany, and they often performed poorly in school. This raised some concerns about the immigrant populations that had been growing steadily since the late 1960s due to Germany's booming economy and demands additional labor. Thus, researchers argued the need for remedial programs for the children from immigrant families (Schwippert 2007). The most significant change after PISA and PIRLS results in Germany is the shift of academic discourse on education in Germany toward a more empirical and practice-focused framework of research from general and didactic research (Bohl 2004). This might have been possible with German teacher unions who began to support large-scale national and international surveys, and began to build strong relationships with educational researchers since the first presentations of PIRLS results (Schwippert 2007).

In Mexico, the OECD was directly involved in the formulation and implementation of reform. The poor results of Mexico in PISA 2006 were more than once used as a justification for the implementation of the reform by the government, as well as by the media and by the OECD itself. The reform especially focused on "improving teacher performance" and "consolidating evaluation" (Figazzolo 2009).

In Denmark, the result of PISA 2000 raised doubts about the efficiency of the Danish education system. Compared to their well-funded education system, the results showed only middle-range outcomes. The results also questioned why social equity continued to be a problem despite the significant investment in social welfare programs (Egelund, 2008). After the PISA 2000 result, Denmark subsequently implemented a range of reform policies, including increased national assessment and evaluation, and strategies to target socioeconomically disadvantaged and immigrant students (Egelund 2008).

In Japan, the national curriculum was revised after the release of the PISA 2003 results in 2004 to incorporate the PISA-type competencies. Implementation began in 2007 (Breakspear 2012).

In the Czech Republic, there was a long-standing belief that its system was homogenous and provided an equal opportunity to all students. However, findings from TIMSS 1995 raised an important debate over social equity. TIMSS also accelerated changes in strengthening technical and vocational schools, which had already been planned (Strakova et al. 2000). In terms of

an assessment system, TIMSS played another important role in the Czech Republic, as it was the first time that Czech students had encountered multiple-choice items, and the first time that standardized tests had been administered in Czech schools (Strakova et al. 2000). As a result, the ministry of education created national standardized examinations for students in their final year of secondary school. Also, in order to improve the quality of tests developed within the Czech Republic, they used TIMSS as an exemplar and as a methodological guide. TIMSS also gave Czech scholars their first opportunity to explore the relationship between home background and achievement (Strakova et al. 2000).

In Italy, the government used PISA results to advocate a rather neo-liberal reform of the country's education system. The government justified reductions in spending and the promotion of evaluation systems to increase efficiency by referring to PISA results. The reform introduced a yearly bonus for "good" teachers in order to foster the quality of the Italian school system. Interestingly, PISA results were employed by unions and opposition parties to criticize this very same reform. They also rejected the government's recommendations by arguing that PISA did not take into account the peculiarity of the Italian system (Figazzolo 2009).

In Australia, PISA results have been used to support preexisting positions. For example, the government has used it to sustain an increasing focus on testing and evaluation. The Federal Education Minister has quoted Australia's position on the PISA ranking in support for her own already existing agenda (Figazzolo 2009).

In New Zealand, high performance on PISA seemed to reinforce existing policies and thus there was no impetus for substantial change (Dobbins 2010).

Some countries emphasized that PISA results complement national data derived from national/federal assessments. For high-performing countries, PISA results are used to compare and validate data from their own national assessments (Breakspear 2012).

In Canada, PISA is used as a complementary indicator to monitor overall system performance. They even redesigned the national assessment program in 2007 to better complement PISA in terms of schedule, target population, and subject areas covered (Breakspear 2012).

In Singapore, PISA data complements national assessment data to inform about the effectiveness of the education system (Breakspear 2012).

In Spain, PISA is used as a means of comparing and contrasting the country's own data generated by national and regional assessment studies. They mention that PISA is an important referent (Breakspear 2012).

In Kuwait, the principal goal of participation in TIMSS was to evaluate the standard of the curriculum with reference to international benchmarks. As a direct result of TIMSS 1995, committees were formed to revise the mathematics and science curricula. In addition, some aspects of the teacher preparation programs were being questioned. Also, examination regulations were revised, particularly at the secondary school level. The impact of TIMSS was not limited to only Kuwait, but throughout a number of Arab countries. As a result, the Kuwait Society for the Advancement of Arab children has begun another study similar to TIMSS for the Gulf Arab countries (Hussein and Hussain 2000).

In Romania, the TIMSS-R results came as a "shock," because they fell below the international average. This finding was surprising because the past success of Romania's top students in Olympiad competitions left educators and the public with the impression that Romanian standards of achievement were high.* With this "wake'up call," Romania made important changes in curriculum guides and the new textbooks reflecting the impact of TIMSS studies. The curriculum guides referred to the TIMSS findings and presented examples of new test items following the TIMSS models. Other changes have occurred in the time devoted to mathematics and science, sequence of topics, emphasis given to the nature of science, experimentation in science, and statistics and probability (Noveanu and Noveanu 2000). Another effect of international studies was on Romanian national assessments. As a result of international assessments, local assessments began to use a variety of item types, particularly multiple-choice and performance items, and to emphasize links between mental processes and test outcomes in teacher seminars. In conclusion, there was a widespread consensus that participation in TIMSS-R was "extremely important" for Romania (Elley, 2005).

In Macedonia, the results of TIMSS-R were a surprise, as they were in Romania, since elite students performed well in Olympiad competitions. Changes were made in curricula, national assessments, and in-service training. New topics have been added to the curriculum. For in-service training, the findings of TIMSS-R were used as a lever to get teachers to take seriously the need to change their teaching styles from predominantly lecturing style to one of interaction (Elley 2005).

Findings from PIRLS also had significant influence in Macedonia. The finding that early literacy activities are very important for further student performance in reading contributed to the government's decision to make the school starting age 6 instead of 7 years old (Naceva and Mickovska 2007). Another main impact of both PIRLS and TIMSS is that educational policy makers are now aware that empirical data provide very important indicators of the quality of education systems. This awareness led to the government establishing a National Center for Assessment and Examinations, which aims to monitor and raise standards in pre-university education by employing large-scale assessment and formal examinations (Naceva and Mickovska 2007). TIMSS-R methodology was used when the national "Matura" examinations began (Elley 2005).

* In the former Soviet Union and throughout Eastern and Central Europe, it was common to assume that the quality of a school system was defined by performance in the international academic Olympics. One important impact of international surveys of academic achievement was to instill the notion of having a result that represents a representative sample of students at large.

Use of Surveys to Generalize about School System Structures

ILSA results show that, although achievement was higher in countries with selective school systems, the effect of socioeconomic status was stronger than the factors associated with achievement. The Six Subject Study found that students in academic courses in countries with a selective education system exhibited higher levels of mathematics achievement than students who had similar courses in countries with a comprehensive school system (Walker 1976). This finding suggests that selective school systems may have higher outcomes than comprehensive school systems. However, the First IEA study also found that family background strongly influences the selection when it occurs early in age (Husen 1967).

Noting these observations, recent studies using TIMSS and PISA data examined the relationship between tracking systems and social inequalities. Studies found that there was a significant effect of early tracking on inequality, whereas there was no clear effect of tracking on mean performance (e.g., Hanushek and Woessmann 2006). Studies also found that later tracking systems are associated with the reduced effect of family background (Brunello and Checchi 2007; Schuetz et al. 2008; Woessmann et al. 2009), although one study by Waldinger (2006) found that tracking does not increase the impact of family background after controlling for pretracking differences.

Thus, one of the recommendations for equity from the OECD is that school structures should be dissuaded from selective models in favor of integrated ones, because selection and tracking reinforce socioeconomic disparities (OECD 2004b). Germany used the PISA 2006 findings to advocate against early tracking in school and for a more socially equitable school system (Figazzolo 2009). To accommodate the differences in abilities of students in comprehensive school systems, the OECD requires teacher training be combined with an integrated school system, as teachers should individually promote students from different backgrounds (OECD 2009).

Use of Surveys to Uncover More Specific Areas for Investigation

Jimenez and Lockheed's (1995) study provides a good example of how surveys can be used as a follow-up of findings from large-scale data. Their earlier work (1986, 1988, 1989) used SIMS data for analysis at the national level. Recognizing that simple findings from large-scale data cannot easily explain what occurs within a classroom, they conducted a "mini-survey," in addition to using SIMS data. The study (1995) shows how student characteristics are different in public and private schools, and what the private school effect

is on achievement. Students in private schools come from more privileged families than those in public school, but, with student background and selection bias held constant, students in private schools outperform students in public schools on a variety of achievement tests.

Having found that private schools are more effective, Jimenez and Lockheed wanted to find out why. They explored what characteristics of private schools contribute to the effectiveness. Using the same large-scale data, they further examined the contribution of student selectivity, peer effects, and some school inputs, such as teacher training and experience and student–teacher ratios. However, they were particularly interested in how private schools allocate their resources and inputs and manage school systems more efficiently. Their hypothesis is that private schools may be more effective than public schools because they invest more of their resources in instructional materials and in improving teacher performance than do public schools. Another hypothesis is that private schools are more flexible in terms of organizational structure because they are more responsive to the demands of parents, students, and educational professionals than are public schools.

Therefore, using a mini-survey they examined the two aspects of public and private schools that they were unable to explore with the large-scale data: school-level resources and school-level management (Jimenez and Lockheed 1995). They invited a senior researcher in each country to gather systematic data about a variety of institutional practices in public and private schools, using a survey instrument that Lockheed and her colleagues provided (Jimenez and Lockheed 1995). The results from the mini-survey confirm what they had observed with the large-scale data. More importantly, it also supports their hypotheses and provides evidence on why private schools are more effective and efficient than public schools. In addition, it shows that simple resource availability cannot explain the differences in effectiveness because the public and private schools in the sample were very similar with respect to their overall resources (Jimenez and Lockheed 1995). Furthermore, the results support the fact that it is resource allocation that makes private schools different. The lesson from this innovative methodology that Lockheed and her colleagues adopted is that the large-scale survey was used not to find out what works in terms of predicting academic achievement, but as a question filtering mechanism for generating feasible hypotheses to be followed-up by small-scale and qualitative work.

Use of Surveys to Develop New Techniques and Discover New Variables

One of the largest limitations of large-scale international data from IEA and OECD is that the data are available only for the countries that participate in the

achievement tests. Barro and Lee (2001) have covered a broad group of countries by using census and survey observations to measure educational attainment. Barro and Lee (2001) have recognized the importance of gathering data on educational attainment across the world as an important indicator of human capital. For measuring quality of education, they compiled test scores on examinations in science, mathematics, and reading that have been conducted in various years for up to 58 countries by the IEA and the International Assessment of Educational Progress (IAEP). At the same time, these measures were restricted by the limited sample, which consists mostly of OECD countries (Barro and Lee 2000). Thus, Barro and Lee explain that educational attainment still provides the best available information to cover a broad number of countries (Barro and Lee 2000). They expanded their data set on educational attainment from 1950 to 2010 and constructed a new panel data set that is disaggregated by sex and age. The data are broken down into 5-year age intervals, and by adding former Soviet republics the coverage has now expanded to 146 countries. The data are more accurate than before because it incorporates recently available census/survey observations (Barro and Lee 2010).

Smaller, Quicker, and Cheaper: Improving Learning Assessments for Developing Countries

LSEAs increasingly became a key tool for meeting the demand of accountability and systematic evaluation in Least Developed Countries (LDCs), particularly after the "Education for All" initiative. However, the complexity and expense of LSEAs have led some to question the utility of conducting LSEAs in LDCs (Wagner 2011). LSEAs are complex data because the number of participating countries and population samples are large and testing instruments must be vetted by experts. A more serious limitation of LSEAs is that they are also costly in terms of money and time. It often costs millions of dollars and takes multiple years to achieve closure, which is a long turnaround time (Wagner 2011).

Recently, there is a new hybrid methodology termed the "smaller, quicker, and cheaper" (SQC) approach, which Wagner (2011) describes. This new approach seeks to focus more directly on the needs of LDC contexts. SQC pays close attention to a variety of factors such as population diversity, linguistic and orthographic diversity, individual differences in learning, and timeliness of analysis (Wagner 2011). One well-known current example of recent hybrid assessment in reading is the Early Grading Reading Assessment (EGRA). The largest advantage of this method is its modest size in assessment. Because the scale is small, one can implement it with a much smaller budget and more quickly assess achievement than with LSEAs. Because of its frequency, this approach can also be more useful

for policy makers if they are interested in early detection and near-term impact. Another advantage of this method is that researchers can explore more deeply the factors that affect learning outcomes, such as language of instruction, language of assessment, and opportunity to learn (Wagner 2011). In sum, SQC assessments can better track learning over time, can better adapt to local linguistic contexts, and can be better designed to understand poor-performing children (Wagner 2011).

Impact on Research Findings

The first study of American schools using a representative national sample generated a debate that, 4 decades later, continues to dominate discussions of education policy. This was the first attempt to summarize the many different influences on learning—neighborhood, socioeconomic status, curriculum, student attitudes, teacher training, and the like. Once summarized, the Equality of Educational Opportunity Report (EEOR; Coleman et al. 1966) concluded that the predominant influence on academic achievement came not from the school but from the characteristics of the home, a set of influences over which education policy was either ineffective or immaterial. Reanalyses of the EEOR (Coleman et al. 1966) data challenged some of the findings, but the main conclusions were reenforced (Mosteller and Moynihan 1972).* The implication was that because the influence of school quality on school achievement in comparison to home background was so weak, equality of opportunity could not occur through changes in education policy alone; it required changes in social policy affecting neighborhoods and the patterns of school attendance.†

In the spring of 1971, Stephen Heyneman was offered the opportunity to do a similar study of the primary schools in Uganda as had been done in the United States. The "Coleman format" was used. This format included separate questionnaires for students, teachers, and headmasters, a lengthy counting of the physical equipment in each school, and achievement tests of math, general knowledge, and English from the primary school leaving examination, the test used to determine entry to secondary school. In 1972, Coleman himself had moved from Johns Hopkins to the University of Chicago and served on the committee of the resulting dissertation (Heyneman 1975a). The irony was that the results from Uganda diverged from the results of the original Coleman Report. Little correlation could be found between a student's

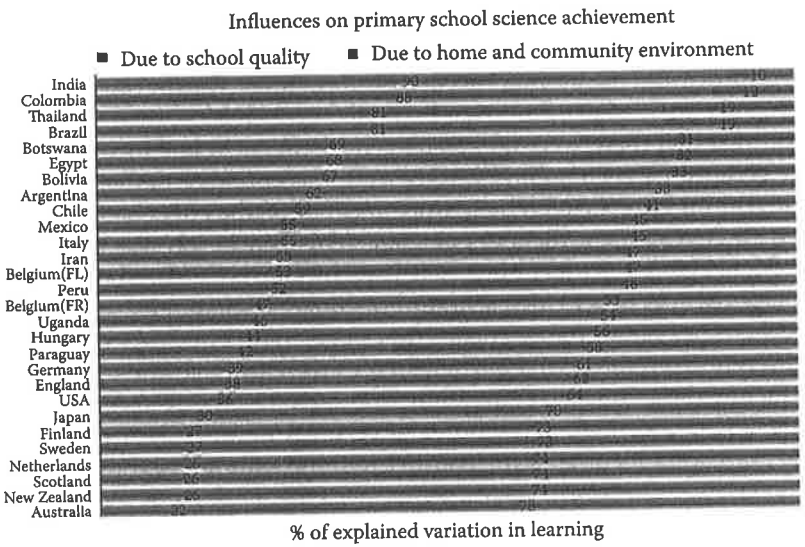
* The Coleman report was cited in 132 publications in 1975; an average of 71/year during the 1980s; 48/year during the 1980s and 50/year after 2000. It is one of the most influential studies in the history of the social sciences.

† Coleman used ordinary least squares (OLS) as the analytical tool. Today OLS has largely been replaced by HLM, which is able to more accurately capture different levels of institutional effects at the classroom, school, and school district level. Using HLM to analyze Coleman's data, school quality appears to have a greater effect than student background (Borman and Dowling 2010).

socioeconomic status and school achievement, and the predominant influence on achievement was the quality of the school, not the home background of the student (Heyneman 1976, 1977a, 1979).

The Uganda study was the first survey of primary schools in Sub-Saharan Africa and led to better understanding of the school system in many ways. It helped isolate the importance of particular teacher characteristics, particularly what they actually know about the subject they are teaching (Heyneman 1975b). It helped to better understand the influences on student learning from school community (Heyneman 1977b); school construction, facilities, and equipment (Heyneman 1977c); language of instruction (Heyneman 1980c); and school administration (Heyneman 1975c). In fact, because the data were representative of the nation's schools, it allowed a comparison of the methods of distribution of school supplies in which it was found that under the government monopoly, supplies were distributed more inequitably than when they were purchased on the open market by the school's parent/teacher committee (Heyneman 1975c). However, perhaps the most important utility from this new category of research came with the identification of textbooks as the single most important contributor to academic achievement, an influence that was all but ignored by the original Coleman Report (Heyneman 1980a). This led the way to a line of investigation on textbook effects, beginning with a meta-analysis (Heyneman et al. 1978) and moving on to experiments comparing the intervention effects of textbooks and education radio (Jamison et al. 1981), and to a nationwide experiment in which the ratio of textbooks-to-child was reduced from 10:1 to 2:1. This generated national gains in academic achievement of unprecedented magnitude (Heyneman et al. 1984). The original Uganda survey was also useful as a baseline to compare changes in the patterns of academic achievement before and after the political catastrophe of Idi Amin (Heyneman and Loxley 1983a).

The Coleman Report's findings, however, had been so powerful that many were convinced that they were universal. If so, it would imply that investments in schools and in school quality would only reinforce preexisting patterns of social stratification; that schooling was in fact harmful to the social opportunities of the poor. This interpretation became popular with the generalization of the Coleman findings outside the United States (Simmons and Alexander 1978). This interpretation, however, had used only those low-income countries that had participated in the IEA studies (Chile and Hungary). They had not included the results from Uganda and elsewhere. When these other studies were included, the pattern of achievement seemed to diverge from the "norm" typical of high-income countries (Heyneman 1980b; Heyneman and Loxley, 1982). The result of a 6-year meta-analysis on this question, sponsored by the World Bank, resulted in the comparison of achievement data from 29 countries, including 15 low- and middle-income countries, which suggested that the pattern was in fact linear: the lower the income of the country, the greater influence school quality



* Technical details in "The effect of primary school quality on academic achievement across 29 high and low income countries," *American Journal of Sociology*, (May 1983).
* Correlation between the influence of school quality and national GNP per capita [$R = -0.72$ ($p < 0.001$)].

Source: Heyneman and Loxely (1983).

FIGURE 3.1
Influences on academic achievement in 29 countries.

had on academic achievement (Heyneman and Loxley, 1983a).* The pattern looked like this (Figure 3.1).

Opposition to the findings and generally to the use of achievement tests across countries rose very quickly in the 1980s. The first set of challenges concerned sampling strategies, test reliability, and measures of socioeconomic status, which, it was alleged, were not sufficiently sensitive to local culture (Theisen et al. 1983). A separate set of issues arose with the use of new regression technologies. Hierarchical linear modeling (HLM) was said to be more accurate in capturing the effects when variables—classroom, school, system—were nested within one another. At first, it was alleged that HLM made the use of OLS simply outdated and the findings derived from OLS unreliable (Riddell 1989a, 1997). While HLM did seem to provide a new look at the model, it appeared unfair to criticize previous analyses as being inadequate. It was as though one criticized Charles Lindberg for flying across the Atlantic without using radar, when radar had not been invented at the time of his flight (Heyneman 1989; Riddell 1989b).

* The pattern of distribution of school quality within countries did not differ between high- and low-income countries, suggesting that the pattern of distribution was not responsible for the differences in the achievement findings (Heyneman and Loxley 1983b).

The real question of whether the Heyneman/Loxley effect was valid came once the TIMSS results were released.* TIMSS allowed the comparison of many new countries and increasingly accurate information. The first attempt to retest the Heyneman/Loxley effect used both OLS and HLM as methods. The conclusion was that there were no greater school effects in low-income countries, either because Heyneman and Loxley were inaccurate or perhaps because in the 20-year interim between their work and TIMSS the effect had been attenuated with development (Baker et al. 2002). This conclusion was supported by the work of Hanushek and Luque (2003) and by Harris (2007), who found insufficient school effects or evidence of diminishing marginal returns (Table 3.2).

On the other hand, the countries that participated in TIMSS were quite a bit different from those that participated in the earlier studies. TIMSS countries included the Russian Federation, Romania, Lithuania, Slovakia, Greece, Portugal, and middle-income countries (Table 3.3).

The per-capita income of the countries in the sample used by Baker et al. was 300% over the world mean, while the Heyneman/Loxley sample was only 50% over the world mean. Did that mean that the two samples were systematically different, and did this difference explain the differences in the results between Heyneman and Loxley versus Baker et al.? Work by Gameron and Long tested this by taking the countries used in the Baker et al. sample and adding 10 countries from Latin America, including Honduras, Bolivia, the Dominican Republic, Paraguay, and Mexico. The per capita income of their sample was $3409, comparable to that of Heyneman and Loxley ($2896) and considerably less than that of Baker et al. ($17,429; Table 3.4).

TABLE 3.2

Summary of Re-Analyses of the Heyneman/Loxley Effect

Source	Findings
Baker et al.	No greater school effect in low-income countries. Hypothesis: due to (i) economic development or (ii) threshold effect.
Hanushek and Luque	No effect of school resources (anywhere) after controlling for SES.
Harris	Search for diminishing marginal returns (DMR) to school quality (SQ): no solid conclusion.
Gameron and Long	School effects greater in low-income countries perhaps because of differences in gross national income (GNI) and differences in the distribution of SQ.
Chudgar and Luschei	School effects greater in low-income countries perhaps due to the inequality in income and the distribution of SQ.
Chiu and Chow	Poorer countries have higher classroom discipline (due to high value of education), hence good teachers have more effect in high-income countries.

* The theory of higher school effects in low-income countries was given the label of the Heyneman/Loxley effect by Baker et al. (2002).

TABLE 3.3

Countries in Different Samples

Heyneman/Loxley (1983)	Baker et al. (2002)	Long (2006)
Uganda	Russia	Honduras
Bolivia	Romania	Bolivia
Egypt	Thailand	Dominican Republic
Iran	Columbia	Paraguay
El Salvador	Latvia	Columbia
Thailand	Lithuania	Brazil
Peru	Slovakia	Venezuela
Paraguay	Hungary	Chile
Columbia	Czech Republic	Mexico
Brazil	Korea	Argentina
Botswana	Slovenia	(plus all countries from Baker et al.)
Chile	Greece	
Mexico	Portugal	
Hungary	Cyprus	
Argentina	New Zealand	
New Zealand	Spain	
Australia	Israel	
Italy	Australia	
United Kingdom	Canada	
Belgium	Hong Kong	
Singapore	France	
Netherlands	United Kingdom	
Finland	Belgium	
Germany	Singapore	
Sweden	Netherlands	
United States	Ireland	
Japan	Austria	
	Germany	
	Iceland	
	Denmark	
	United States	
	Norway	
	Switzerland	

Their conclusions support the original Heyneman/Loxley effect suggesting that the theory is supportable (including the use of HLM) and true over time (Gameron and Long 2007). Their results were also supported by the reanalyses of Chudgar and Luschei, who find that school effects are greater in low-income countries (Chudgar and Luschei, 2009), and the work of Chiu and Chow who find that in low-income countries teachers have more effect, perhaps on grounds that school discipline is higher due to the high value placed on education (Chiu and Chow 2010). Furthermore, a recent reanalysis of PISA data suggests that poor school performance does not necessarily follow from a student's disadvantaged background (Schleicher 2009).

TABLE 3.4

Reanalyses of the Heyneman–Loxley Effect: Comparison of Samples Per-Capita Income and School Effects

Sample	PCI (Per-Capita Income)	School Effects (% of Variance Explained)
Heyneman/Loxley (1983a)	2896	50.5
Baker et al. (2002)	17,429	34.4
Gameron and Long (2007)	3409	56.7

The sum result of this important debate would not have been possible without the ability to compare the influences on academic achievement across nations. Though the methods and measures have many caveats and problems, the fact is that these international surveys allow us to inquire as to the product and purposes of education itself and to either challenge or reinforce the reasons for the public investment. As to whether the Heyneman/Loxley effect is, in the end, supported, the jury is still out. The theory has helped stimulate debate for three decades, and may continue to stimulate debate for three more decades. What is clear is that the patterns of influences on academic achievement are not uniform around the world; they vary by gender, dependent variable (math is more subject to school effects; reading more influenced by home background), and by student age. What we do know is that reasons for disappointing results in the United States, particularly in urban areas, may be informed by the results from these studies conducted in very different environments, in which all children—including all poor children—have a high desire to learn.

Appendix

List of the Members of the International Association for the Evaluation of Education Achievement

Africa

Botswana
Kenya
Nigeria
South Africa

Asia

Armenia
China, People's Republic of
Chinese Taipei
Georgia
Hong Kong SAR
Indonesia
Japan
Kazakhstan
Korea, Republic of
Malaysia
Philippines
Singapore
Thailand

Australasia

Australia
New Zealand

Europe

Austria
Belgium (Flemish)
Belgium (French)
Bosnia and Herzegovina
Bulgaria
Croatia
Cyprus
Czech Republic
Denmark
England
Estonia
Finland
France
Germany
Greece
Hungary
Iceland
Ireland
Italy
Latvia
Lithuania
Luxembourg
Macedonia
Netherlands
Norway
Portugal
Romania
Russian Federation
Scotland
Slovak Republic
Slovenia
Spain
Sweden
Turkey

North Africa and Middle East

Egypt
Iran
Israel
Jordan
Kuwait
Morocco
Palestinian National Authority
Qatar
United Arab Emirates

The Americas

Brazil
Canada
Chile
Colombia
Mexico
United States

References

Amadeo, J., Purta, J.T., Lehmann, R., Husfeldt, V., and Nikolova, R. (2002). *Civic Knowledge and Engagement: An IEA Study of Upper Secondary Students in Sixteen Countries*. Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Ammermueller, A., Heijke, H., and Woessmann, L. 2005. Schooling quality in Eastern Europe: Educational production during transition. *Economics of Education Review*, 24(5), 579–599.

Anderson, L.W. and Postlethwaite, T.N. 1989. What IEA studies say about teachers and teaching. In: Purves, A.C. (Ed.), *International Comparisons and Educational Reform*. Alexandria, VA: Association for Supervision and Curriculum Development. (ERIC Document Reproduction Service No. ED 316 494).

Antonijevic, R. 2007. *Usage of computers and calculators and students' achievement: Results from TIMSS 2003*. Paper presented at the International Conference on Informatics, Educational Technology and New Media in Education, March 31–April 1, Sombor, Serbia.

Baker, D.P., Goesling, B., and Letendre, G.K. 2002. Socioeconomic status, school quality, and national economic development: A cross-national analysis of the 'Heyneman–Loxley Effect' on mathematics and science achievement. *Comparative Education Review*, 46(3), 291–312.

Barro, R.J. and Lee, J.W. 2000. International data on educational attainment: Updates and implications. NBER Working Paper N. 7911.

Barro, R.J. and Lee, J.W. 2001. International data on educational attainment: Updates and implications. *Oxford Economic Papers*, 53(3), 541–563.

Barro, R.J. and Lee, J.W. 2010. A new data set of educational attainment in the world, 1950–2010. NBER Working Paper N° 15902.

Berova, M. and Matusova, S. 2000. Slovak Republic. In: Robitaille, D.F., Beaton, A.E., and Plomp, T. (Eds.), *The Impact of TIMSS on the Teaching & Learning of Mathematics & Science* (pp. 133–138). Vancouver, Canada: Pacific Educational Press.

Bieber, T. and Martens, K. 2011. The OECD PISA study as a soft power in education? Lessons from Switzerland and the U.S. *European Journal of Education*, 46(1), 101–116.

Bohl, T. 2004. Empirische Unterrichcsforschung und Allgemeine Didaktik, Entstehung, Situation und Konsequenzen eines prekaren Spannungsverhaltnisses im Kontext der PISA-Studie. *Die Deutsche Schule*, 96, Ig/Heft 4, 414–425.

Borman, G. and Dowling, M. 2010. Schools and inequality: A multilevel analysis of Coleman's equality of educational opportunity data. *Teachers College Record* (on line version) January 28, 2010.

Bradburn, N. and Gilford, D.M (Eds.) 1990. *A Framework and Principles for International Comparative Studies in Education.* Washington, DC: National Academy Press.

Bradburn, N., Haertel, E. Schwille, J., and Torney-Purta, J. 1991. A rejoinder to 'I never promised you first place.' *Phi Delta Kappan*, 72(10), 774–777.

Braun, H. and Kanjee, A. 2006 Using assessment to improve education in developing nations. In: Cohen, J.E., Bloom, D.E., and Malin, M.B. (Eds.), *Educating All Children: A Global Agenda* (pp. 303–53). Cambridge, MA: American Academy Press.

Breakspear, S. 2012. The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance. *OECD Education Working Papers*, No. 71, OECD Publishing. http://dx.doi.org/10.1787/5k9fdfqffr28-en.

Brunello, G. and Checchi, D. 2007. Does school tracking affect equality of opportunity? New international evidence. *Economic Policy*, 22(52), 781–861.

Chamberlain, M. and van Aalst, I. 2000. New Zealand. In: Robitaille, D.F., Beaton, A.E., and Plomp, T. (Eds.), *The Impact of TIMSS on the Teaching & Learning of Mathematics & Science* (pp. 98–103). Vancouver, Canada: Pacific Educational Press.

Chiu, M.M. and Chow, B.W. 2010. Classroom discipline across 41 countries: School, economic and cultural differences. *Journal of Cross-Cultural Psychology*, 20(10), 1–18.

Chromy, R.R. 2002. Sampling issues in design, conduct and interpretation of international comparative studies of school achievement. In: National Research Council (Ed.), *Methodological Advances in Cross-National Surveys of Educational Achievement* (pp. 80–117). Washington, DC: National Academy Press.

Chudgar, A. and Luschei, T.F. 2009. National income, income inequality, and the importance of schools: A hierarchical cross-national comparison. *American Education Research Journal*, 46(3), 626–658.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfall, F.D., and York, R.L. 1966. *The Equality of Educational Opportunity* (two volumes). Washington, DC: United States Department of Health, Education and Welfare.

Comber, L.C. and Keeves, J.P. 1973. *Science Education in Nineteen Countries.* Stockholm: Almqvist and Wiksell, and New York: Wiley.

De Boer, J. 2012. The role of engineering skills in development. Doctoral dissertation. Vanderbilt University, Nashville, Tennessee.

Dobbins, M. 2010. Education policy in New Zealand—Successfully navigating the international market for education. In: Martens, K., Kenneth, N.A., Windzio, M., and Weymann, A. (Eds.), *Transformation of Education Policy*. Basingstoke: Palgrave.

Egelund, N. 2008. The value of international comparative studies of achievement—A Danish perspective. *Assessment in Education: Principles, Policy & Practice*, 15(3), 245–251.

Fensham, P.J. 2007. Context or culture: Can TIMSS and PISA teach us about what determines educational achievement in science?. In: Atweh, B., Barton, A.C., Borba, M.C., Gough, N., Keitel, C., Vistro-Yu, C., and Vithal, R. (Eds.), *Internationalization and Globalization in Mathematics and Science Education* (pp. 151–172). Dordrecht, Netherlands: Springer.

Fertig, M. and Wright, R.E. 2005. School quality, educational attainment and aggregation bias. *Economics Letters*, 88(1), 109–114.

Figazzolo, L. 2009. Impact of PISA 2006 on the education policy debate. *Education International.* Retrieved from http://www.ei-ie.org/research/en/documentation.php.

Fuchs, T. and Woessmann, L. 2004. *Computers and Student Learning: Bivariate and Multivariate Evidence on the Availability and Use of Computers at Home and at School.* Munich, Germany: IFO Institute for Economic Research at the University of Munich.

Gameron, A. and Long, D. 2007. Equality of educational opportunity: A 40 year retrospective. In: Teese, R., Lamb, S., and Duru-Bellat, M. (Eds.), *International Studies in Educational Inequality, Theory and Policy* (pp. 23–48). Dordrecht, Netherlands: Springer.

Ganimian, A. 2009. *How Much Are Latin American Children Learning?: Highlights from the Second Regional Student Achievement Test (SERCE).* Washington, DC: PREAL.

Ganimian, A. and Rocha, A.S. 2011. *Measuring Up? How did American and the Caribbean perform on the 2009 Programme for International Student Assessment (PISA)?* Programa de Promocion de la Reforma Educativa en America Latina y el Caribe (PREAL) Retrieved from Inter American Dialogue website: http://www.thedialogue.org/PublicationFiles/Preal_PISA_ENGLowres.pdf.

Geske, A. and Kangro, A. 2000. Latvia. In: Robitaille, D.F., Beaton, A.E., and Plomp, T. (Eds.), *The Impact of TIMSS on the Teaching & Learning of Mathematics & Science* (pp. 85–88). Vancouver, Canada: Pacific Educational Press.

Gudmundsson, E. 2000. Iceland. In: Robitaille, D.F., Beaton, A.E., and Plomp, T. (Eds.), *The Impact of TIMSS on the Teaching & Learning of Mathematics & Science* (pp. 56–60). Vancouver, Canada: Pacific Educational Press.

Hanushek, E.A. and Luque, J.A. 2003. Efficiency and equity in schools around the world. *Economics of Education Review*, 22, 481–502.

Hanushek, E.A. and Woessmann, L. 2006. Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *Economic Journal*, 116(510), 63–76.

Harris, D.N. 2007. Diminishing marginal returns and the production of education: An international analysis. *Education Economics*, 15(1), 31–53.

Heyneman, S.P. 1975a. Influences on academic achievement in Uganda: A "Coleman Report" from a non-industrial society. Doctoral dissertation. University of Chicago, Chicago, IL.

Heyneman, S.P. 1975b. Changes in efficiency and in equity accruing from government involvement in Ugandan primary education. *African Studies Review*, 18(1), 51–60.

Heyneman, S.P. 1975c. Relationships between teachers' characteristics and differences in academic achievement among Ugandan primary schools. *Education in Eastern Africa*, 6(1), 41–51.

Heyneman, S.P. 1976a. A brief note on the relationship between socioeconomic status and test performance among Ugandan primary school children. *Comparative Education Review*, 20(1), 42–47.
Heyneman, S.P. 1977a. Influences on academic achievement: A comparison of results from Uganda and more industrial societies. *Sociology of Education*, 11(2), 245–259.
Heyneman, S.P. 1977b. Relationships between the primary school community and academic achievement in Uganda. *Journal of Developing Areas*, 11(2), 245–259.
Heyneman, S.P. 1977c. Differences in construction, facilities, equipment and academic achievement among Ugandan primary schools. *International Review of Education*, 23, 35–46.
Heyneman, S.P. 1979. Why impoverished children do well in Ugandan schools. *Comparative Education*, 15(2), 175–185.
Heyneman, S.P. 1980a. Differences between developed and developing countries: Comment on Simmons and Alexander's determinants of school achievement. *Economic Development and Cultural Change*, 28(2), 403–406.
Heyneman, S.P. 1980b. Student learning in Uganda: textbook availability and other determinants. *Comparative Education Review*, 24(2) (June) 108–118 (coauthored with Dean Jamison).
Heyneman, S.P. 1980c. Instruction in the mother tongue: The question of logistics. *Journal of Canadian and International Education*, 9(2), 88–94.
Heyneman, S.P. 1983. Education during a period of austerity: Uganda, 1971–1981. *Comparative Education Review*, 27(3), 403–413.
Heyneman, S.P. 1989. Multilevel methods for analyzing school effects in developing countries. *Comparative Education Review*, 33(4), 498–504.
Heyneman, S.P. 2004. International education quality. *Economics of Education Review*, 23, 441–52.
Heyneman, S.P. 2009a. "Should school time be extended each day and in the summer?" *The Tennessean* (October 6).
Heyneman, S.P. 2009b. An education bureaucracy that works. *Education Next* (November) http://educationnext.org/an-education-bureaucracy-that-works/
Heyneman, S.P., Farrell, J.P., and Sepulveda-Stuardo. 1978. *Textbooks and Achievement: What We Know.* Washington, DC: World Bank Staff Working Paper No. 298 (October) (available in English, French, and Spanish).
Heyneman, S.P. and Loxley, W. 1982. Influences on academic achievement across high and low-income countries: A re-analysis of IEA data. *Sociology of Education*, 55(1), 13–21.
Heyneman, S.P. and Loxley, W. 1983a. The effect of primary school quality on academic achievement across twenty-nine high- and low-income countries, *American Journal of Sociology*, 88(6), 1162–1194.
Heyneman, S.P. and Loxley, W. 1983b. The distribution of primary school quality within high- and low-income countries. *Comparative Education Review*, 27(1), 108–118.
Heyneman, S.P., Jamison, D., and Montenegro, X. 1984. Textbooks in the Philippines: Evaluation of the pedagogical impact of a nationwide investment. *Educational Evaluation and Policy Analysis*, 6(2), 139–150.
Heyneman, S.P. and Lykins, C. 2008. The evolution of comparative and international education statistics. In: Ladd, H.F. and Fiske E.B. (Eds.), *Handbook of Research in Education Finance and Policy* (pp. 105–127). New York: Routledge.

Holliday, W.G. and Holliday, B.W. 2003. Why using international comparative math and science achievement data from TIMSS is not helpful. *The Education Forum* 63(3), 250–257.
Husen, T., (Ed). 1967. *International Study of Achievement in Mathematics: A Comparison of Twelve Countries.* New York: John Wiley and Sons.
Hussein, M.G.A. and Hussain, A.A. 2000. Kuwait. In: Robitaille, D.F., Beaton, A.E., and Plomp, T. (Eds.), *The Impact of TIMSS on the Teaching & Learning of Mathematics & Science* (pp. 82–84). Vancouver, Canada: Pacific Educational Press.
Jamison, D.T., Heyneman, S.P., Searle, B., and Galda, K. 1981. Improving elementary mathematics education in Nicaragua: An experimental study of the impact of textbooks and radio on achievement. *Journal of Educational Psychology*, 73(4), 556–567.
Jimenez, E. and Lockheed, M.E. 1995. Public and private secondary education in developing countries. Discussion Paper No. 309. The World Bank. Washington, DC.
Jimmenez, E., Lockheed, M., and Wattanawaha, N. 1988. The relative efficiency of public and private school: The case of Thailand. The World Bank.
Kamens, D.H. 2013. Globalization and the emergence of an audit culture: PISA and the search for 'best practice' and magic bullets. In: Benavot, A. and M. Heinz-Dieter (Eds.), *PISA, Power and Policy: the Emergence of Global Educational Governance* (Oxford Studies in Comparative Education). Oxford, UK: Symposium Books.
Karimi, A. and Daeipour, P. 2007. The impact of PIRLS in the Islamic Republic of Iran. In: Schwippert, K. (Ed.), *Progress in Reading Literacy: The Impact of PIRLS 2001 in 13 Countries.* New York, NY: Munster Waxmann.
Keeves, J.P. 1995. *The World of School Learning: Selected Key Findings from 35 Years of IEA Research.* The Hague: International Association for the Evaluation of Educational Achievement.
Keys, W. 2000. England. In: Robitaille, D.F., Beaton, A.E., and Plomp, T. (Eds.), *The Impact of TIMSS on the Teaching & Learning of Mathematics & Science.* Vancouver, Canada: Pacific Educational Press.
Kiamanesh, A.R. and Kheirieh, M. 2000. Iran. In Robitaille, D.F., Beaton, A.E., and Plomp, T. (Eds.), *The Impact of TIMSS on the Teaching & Learning of Mathematics & Science* (pp. 61–65). Vancouver, Canada: Pacific Educational Press.
Kovalyova, G. 2000. Russia. In: Robitaille, D.F., Beaton, A.E., and Plomp, T. (Eds.), *The Impact of TIMSS on the Teaching & Learning of Mathematics & Science* (pp. 120–124). Vancouver, Canada: Pacific Educational Press.
Kozma, R.B. (Ed.) 2003. *Technology, Innovation and Educational Change: A Global Perspective.* Eugene, OR: International Society for Technology in Education.
Lockheed, M.E., Vail, S.C., and Fuller, B. 1986. How textbooks affect achievement in developing countries: Evidence from Thailand. *Educational Evaluation and Policy Analysis Winter*, 8(4), 379–392.
Lockheed, M.E., Fuller, B., and Nyrongo, R. 1989. Family effects on students' achievement in Thailand and Malawi. *Sociology of Education*, 62, 239–256.
Medrich, E.A. and Griffith, J.E. 1992. International Mathematics and Science Assessments: What Have We Learned? U.S. Department of Education, Office of Educational Research and Improvement, NCES-92–011, January.
Mevavech, Z.R. 2000. Israel. In: Robitaille, D.F., Beaton, A.E., and Plomp, T. (Eds.), *The Impact of TIMSS on the Teaching & Learning of Mathematics & Science* (pp. 66–70). Vancouver, Canada: Pacific Educational Press.

Mosteller, F. and Moynihan, D. (Eds.) 1972. *On Equality of Educational Opportunity: Papers Deriving from the Harvard Faculty Seminar on the Coleman Report*. New York: Vintage Books.

Mullis, I.V.S. and Martin, M.O. 2007. Lessons learned: What international assessments tell us about math achievement. In: Loveless, T. (Ed.), *TIMSS in Perspective: Lessons Learned from IEA's Four Decades of International Mathematics Assessments*. Washington, DC: Brookings Institution Press.

Mullis, I.V., Martin, M.O., Gonzalez, E.J., and Chrostowski, S.J. 2004. TIMSS 2003 International Mathematics Report.

Mullis, I.V., Martin, M.O., Gonzales, E.J., Kelly, D.L., and Smith, T.A. 1996. *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Naceva, B. and Mickovska, G. 2007. The impact of PIRLS in the Republic of Macedonia. In: Schwippert, K. (Ed.), *Progress in Reading Literacy: The Impact of PIRLS 2001 in 13 Countries*. New York, NY: Munster Waxmann.

Neumann, K., Fischer, H.E., and Kauertz, A. 2010. From PISA to educational standards: The impact of large-scale assessments on science education in Germany. *International Journal of Science and Mathematics Education*, 8, 545–563.

Noveanu, G. and Noveanu, D. 2000. Romania. In Robitaille, D.F., Beaton, A.E., and Plomp, T. (Eds.), *The Impact of TIMSS on the Teaching & Learning of Mathematics & Science* (pp. 117–119). Vancouver, Canada: Pacific Educational Press.

OECD. 2002. *Lesen Kann Die Welt verandern: Leistung und Engagement im Landervergleich*. Paris: OECD.

OECD. 2004a. *Learning for Tomorrow's World. First Results from PISA 2003*. Paris: OECD.

OECD. 2004b. *What Makes School Systems Perform. Seeing School Systems Through the Prism of PISA*. Paris: OECD.

OECD. 2009. *Education Today. The OECD Perspective*. Paris: OECD.

OECD and UNESCO. 2003. *Literacy Skills for the World of Tomorrow—Further Results from PISA 2000*. Paris and Montréal: OECD and UNESCO Institute for Statistics.

OECD. 2010. *PISA 2009 Results: Executive Summary*. Paris: OECD.

Papanastasiou, E.C., Zembylas, M., and Vrasidas, C. 2003. Can computer use hurt science achievement? The U.S.A results from PISA. *Journal of Science Education and Technology*, 12(3), 325–332.

Papanastasiou, E.C., Zembylas, M., and Vrasidas, C. 2004. Reexamining patterns of negative computer-use and achievement relationships. Where and why do they exist? In: Papanastasiou, C. (Ed.), *Proceedings of the IRC-2004. TIMSS. Volume 1*. Lefkosia, Cyprus: Cyprus University Press.

Park, K. 2004. Factors contributing to East Asian students' high achievement: Focusing on East Asian teachers and their teaching. Paper presented at the APEC Educational Reform Summit, 12, January 2004.

Pelgrum, W. and Plomp, T. 2008. Methods for large-scale international studies on ICT in education. In: Voogt, J., Knezek, G. (Eds.), *International Handbook of Information Technology in Primary and Secondary Education* (pp. 1053–1066). Springer Science + Business Media, LLC, New York, 2008.

Plank, D.N. and Johnson, Jr. B.L. 2011. Curriculum policy and educational productivity. In: Mitchell, D.E., Crowson, R.L. and Shipps, D. (Eds.), *Shaping Education Policy: Power and Process* (pp. 167–188). New York, NY: Routledge Taylor & Francis Group.

PREAL. 2011. *Better Schools through Better Policy*. Washington, DC: PREAL.

Purves, A.C. (Ed.). 1992. *The IEA Study of Written Composition II: Education and Performance in Fourteen Countries*. Oxford: Pergamon Press.

Ravitch, D. 2010. *The Death and Life of the Great American School System: How Testing and Choice are Undermining Education*. New York: Basic Books.

Riddell, A.R. 1989a. An alternative approach to the study of school effectiveness in third world countries. *Comparative Education Review*, 33(4), 481–497.

Riddell, A.R. 1989b. Response to Heyneman. *Comparative Education Review*, 33(4), 505–506.

Riddell, A.R. 1997. Assessing designs for school effectiveness research in developing countries. *Comparative Education Review*, 41(2), 178–204.

Rotberg I.C. 1990. I never promised you first place. *Phi Delta Kappan*, 72, 296–303.

Rotberg I.C. 1996. Five myths about test score comparisons. *School Administrator*, 53, 30–31.

Rotberg I.C. 2006. Assessment around the world. *Educational Leadership*, 64(3), 58–63.

Rotberg I.C. 2007. Why do our myths matter? *School Administrator*, 64(4), 6.

Rotberg I.C. 2008. Quick fixes, test scores and the global economy: Myths that continue to confound us. *Education Week*, 27(41), 32.

Rowan, B. 2002. Large scale cross-national surveys of educational achievement: Pitfalls and possibilities. In: Porter, A.C. and Gamoran, A. (Eds.), *Methodological Advances in Cross-National Surveys of Educational Achievement* (pp. 321–349). *Board on Comparative Studies in Education*. Washington, DC: National Academy Press.

Schleicher, A. 2009. Securing quality and equity in education: Lessons from PISA. *UNESCO Prospects*, 39, 251–263.

Schuetz, G., Ursprung, H.W., and Woessmann, L. 2008. Education policy and equality of opportunity. *Kyklos*, 61(2), 279–308.

Schulz, W., Ainley, J., Fraillon, J., Kerr, D., and Losito, B. 2010. *ICCS 2009 International Report: Civic Knowledge, Attitudes, and Engagement among Lower-Secondary School Students in 38 Countries*. Amsterdam: International Association for the Evaluation of Educational Achievement.

Schwippert, K. 2007. The impact of PIRLS in Germany. In Schwippert, K. (Ed.), *Progress in Reading Literacy: The Impact of PIRLS 2001 in 13 Countries*. New York, NY: Munster Waxmann.

Shorrocks-Taylor, D., Jenkins, E., Curry, J., Swinnerton, B., Laws, P., Hargreaves, M., and Nelson, N. 1998. *An Investigation of the Performance of English Pupils in the Third International Mathematics and Science Study (TIMSS)*. Leeds: Leeds University Press.

Simmons, J. and Alexander, L. 1978. The determinants of school achievement in developing countries: A review of research. *Economic Development and Cultural Change* (January), 341–358.

Smyth, J. 1996. The origins, purposes and scope of the international standard classification of education. Paper submitted to the ISCED revision task force. Paris: UNESCO (February) (mimeographed).

Smyth, J. 2005. *International Literacy Statistics 1950–2000*. Paris: UNESCO (mimeographed).

Stigler, J.W. and Hiebert, J. 1999. *The Teaching Gap*. New York: Free Press.

Strakova, J., Paleckova, J., and Tomasek, V. 2000. Czech Republic. In Robitaille, D.F. Beaton, A.E., and Plomp, T. (Eds.), *The Impact of TIMSS on the Teaching & Learning of Mathematics & Science* (pp. 41–44). Vancouver, Canada: Pacific Educational Press

Taylor, A.R. 2000. Canada. In: Robitaille, D.F., Beaton, A.E., and Plomp, T. (Eds.), *The Impact of TIMSS on the Teaching & Learning of Mathematics & Science* (pp. 50–55) Vancouver, Canada: Pacific Educational Press.

Theisen, G.L., Achola, P.W., and Boakari, F.M. 1983. The underachievement of cross-national studies of academic achievement. *Comparative Education Review*, 27(1), 46–68.

Twist, L. 2007. The impact of PIRLS in England. In: Schwippert, K. (Ed.), *Progress in Reading Literacy: The Impact of PIRLS 2001 in 13 Countries*. New York, NY: Munster Waxmann.

Vlaardingerbroek, B. and Taylor, T.G.N. 2003. Teacher education variables as correlates of primary science ratings in thirteen TIMSS systems. *International Journal of Educational Development*, 23, 429–438.

Wagemaker, H. 2004. IEA: International studies, impact and transition. Speech given on the occasion of the *1st IEA International Research Conference*, University of Cyprus, Lefkosia, Cyprus, 11–13, May 2004.

Wagner, D.A. 2011. *Smaller, Quicker, Cheaper. Improving Learning Assessments for Developing Countries*. Paris, France: International Institute for Educational Planning. UNESCO.

Waldinger, F. 2006. *Does Tracking Affect the Importance of Family Background on Students' Test Scores*? Mimeo: London School of Economics.

Walker, D.A. (Ed.). 1976. *The IEA Six-Subject Survey: An Empirical Study of Education in Twenty-One Countries*. New York: Wiley.

Weiler, H. 1994. The failure of reform and the macro-politics of education: Notes on a theoretical challenge. In: Val Rust (Ed.), *Educational Reform in International Perspective* (pp. 43–54). Greenwich, CT: JAI Press.

Wittwer, J. and Senkbeil, M. 2008. Is students' computer use at home related to their mathematical performance at school? *Computers & Education*, 50(4), 1558–1571.

Woessmann, L. 2005a. Educational production in East Asia: The impact of family background and schooling policies on student performance. *German Economic Review*, 6(3), 331–353.

Woessmann, L. 2005b. Educational production in Europe. *Economic Policy*, 20(43), 446–504.

Woessmann, L. and West, M.R. 2006. Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review*, 50(3), 695–736.

Woessmann, L., Luedemann, E., Schuetz, G., and West, M.R. 2009. *School Accountability, Autonomy, and Choice Around the World*. Cheltenham, UK: Edward Elgar.

World Bank. 2011. *Learning for All: Investing in People's Knowledge and Skills to Promote Development. Education Sector Strategy, 2020*. Washington, DC: The World Bank.

Section II

Analytic Processes and Technical Issues Around International Large-Scale Assessment Data

4

Assessment Design for International Large-Scale Assessments

Leslie Rutkowski
Indiana University

Eugene Gonzalez
IEA-ETS Research Institute and Educational Testing Service

Matthias von Davier
Educational Testing Service

Yan Zhou
Indiana University

CONTENTS

Introduction 75
Background 76
What Is MMS? 77
Why Is MMS Useful? 79
Evolution of a Method 80
Item Sampling 80
Balanced Incomplete Block Designs 81
Current Applications of MMS in International Assessment 83
PISA 2009 83
TIMSS 2011 87
PIRLS 2011 89
How Can MMS Be Used in the Future? 90
References 92

Introduction

As an invariant constraint to producing high-quality measurement instruments, questionnaire length is a significant factor in the development process. Owing to issues of fatigue, attrition, or the logistics of available study