IEA Research for Education

A Series of In-depth Analyses Based on Data of the International Association for the Evaluation of Educational Achievement (IEA)

Hans Wagemaker *Editor*

Reliability and Validity of International Large-Scale Assessment

Understanding IEA's Comparative Studies of Student Achievement

IEA

OPEN ACCESS

Springer

IEA Research for Education

A Series of In-depth Analyses Based on Data of the International Association for the Evaluation of Educational Achievement (IEA)

Volume 10

The International Association for the Evaluation of Educational Achievement (IEA) is an independent nongovernmental nonprofit cooperative of national research institutions and governmental research agencies that originated in Hamburg, Germany in 1958. For over 60 years, IEA has developed and conducted high-quality, large-scale comparative studies in education to support countries' efforts to engage in national strategies for educational monitoring and improvement.

IEA continues to promote capacity building and knowledge sharing to foster innovation and quality in education, proudly uniting more than 60 member institutions, with studies conducted in more than 100 countries worldwide.

IEA's comprehensive data provide an unparalleled longitudinal resource for researchers, and this series of in-depth peer-reviewed thematic reports can be used to shed light on critical questions concerning educational policies and educational research. The goal is to encourage international dialogue focusing on policy matters and technical evaluation procedures. The resulting debate integrates powerful conceptual frameworks, comprehensive datasets and rigorous analysis, thus enhancing understanding of diverse education systems worldwide.

More information about this series at http://www.springer.com/series/14293

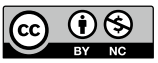
Hans Wagemaker
Editor

Reliability and Validity of International Large-Scale Assessment

Understanding IEA's Comparative Studies of Student Achievement

Springer

Editor
Hans Wagemaker
Newlands, Wellington, New Zealand



ISSN 2366-1631 ISSN 2366-164X (electronic)
IEA Research for Education
ISBN 978-3-030-53080-8 ISBN 978-3-030-53081-5 (eBook)
https://doi.org/10.1007/978-3-030-53081-5

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Founded in 1958, IEA (International Association for the Evaluation of Educational Achievement) pioneered international large-scale assessments in education (ILSAs). IEA today is an international cooperative of national research institutions and governmental research agencies representing more than 60 countries, working together with scholars and analysts to research, understand, and improve education worldwide. Over 100 education systems participate in IEA studies. IEA has conducted more than 30 international large-scale assessments.

There are four core IEA studies. The Trends in International Mathematics and Science Study (TIMSS) has assessed grade 4 and grade 8 students' achievement in mathematics and science every four years since 1995, while IEA's Progress in International Reading Literacy Study (PIRLS) has investigated grade 4 students' reading abilities every five years since 2001. The International Civic and Citizenship Education Study (ICCS) reports on grade 8 students' knowledge and understanding of concepts and issues related to civics and citizenship, as well as their beliefs, attitudes, and behaviors with respect to this domain. Lastly, the International Computer and Information Literacy Study (ICILS) researches grade 8 students' ability to use computers to investigate, create, and communicate in order to participate effectively at home, at school, in the workplace, and in the community. All these studies also collect rich background information from the assessed students and their teachers, school principals, and, in case of grade 4 student assessments, also from the assessed students' parents.

Over the last 60 years, IEA studies have not only seen an increase in the number of countries participating but have also attracted growing interest from both policymakers and the general public. An increasing number of researchers from many different fields are using data from IEA studies to inform their inquiries and micro-level data from all IEA studies can be downloaded freely from the IEA website (https://www.iea.nl/data-tools/repository). The data includes thorough documentation and dedicated software that enables researchers to analyze the data correctly (https://www.iea.nl/data-tools/tools). In the early years of ILSAs, the data

was used mostly by the researchers involved in the studies, but, as the studies have become established and the utility of the data better understood, interest has extended beyond those involved in the design and implementation of these studies and more and more researchers are making use of the data, a development that is highly appreciated by all of us involved with IEA. But, although there is thorough documentation of the data and extensive reports explaining how each study is conducted, it is essential to ensure that IEA's procedures are transparent to researchers not directly involved in these studies and to others who might be generally interested in how ILSAs work.

This book aims to shed more light on various aspects of IEA studies. It outlines how IEA studies are conducted, from developing the research questions to constructing study frameworks, to designing the instruments (assessment materials, background questionnaires, and manuals), establishing the important aspects when conducting a study, and reporting on the results. This book shows the rigor of these scientific endeavors and the quality control mechanisms that are required to ensure high quality data and high quality study reports.

This book also reveals how ILSAs have evolved. From an outside perspective, ILSAs may seem to have a well-established formula, with data collected in the same way over years and for decades to come; however, as this book explains, this is a misunderstanding. Procedures developed based on previous experiences and especially in response to developments in the digital world offer new opportunities but also pose new challenges to this field of science. I am certain that this book will provide a wealth of relevant information for anyone interested in ILSAs.

As with every IEA project, this book is the result of a strong collaborative effort. I am most grateful to Hans Wagemaker for his willingness to organize this book, bringing together a team of distinguished experts to share their knowledge and experiences. Editing a book with so many authors is always a challenging task, and I heartily congratulate Hans for creating such a coherent and informative volume to grace this series.

I also must thank the authors of all the chapters for their willingness to contribute both their time and knowledge to support this endeavor. All have excellent reputations in their respective fields and extensive demands on their time, and thus I truly appreciate their enthusiastic responses to the invitation to contribute to this book, given all additional work this entailed.

No acknowledgment would be complete without also recognizing IEA's study centers and all the researchers involved in our studies not only for their work on the studies but also for their continuous efforts to improve IEA studies further.

I especially thank IEA's Publications and Editorial Committee for ensuring the quality of this book, as they do for all IEA publications, and in particular, I am grateful to Valena Plisko, who, because the series editors and many committee colleagues were already involved as authors, coordinated the independent peer

review process for this special volume. My final thanks go to the IEA staff who helped in coordinating, editing, and publishing this book, in particular Gillian Wilson, IEA's Senior Publications Officer.

Hamburg, Germany

Dirk Hastedt
IEA Executive Director

Contents

About the Editor

Dr. Hans Wagemaker was the executive director of the International Association for the Evaluation of Educational Achievement (IEA) for 17 years, responsible for the management of all IEA international research and assessment projects and activities. He helped develop IEA's Progress in International Reading Literacy Study (PIRLS) and oversaw the development and expansion of IEA's training and capacity building activities in low to middle income countries, and IEA's educational consultancy services. Together with Educational Testing Services (ETS), he established the IEA-ETS Research Institute (IERI), where he continues to serve as a Board member.

Dr. Wagemaker was a Senior Manager Research and International with the Ministry of Education, New Zealand, and represented New Zealand's interests in the APEC Education Forum, UNESCO's commissions, and the OECD, CERI, and the Education Governing Board. He has consulted for the Inter American Development Bank and UNESCO and worked extensively with the World Bank to advance a common interest in the uses of assessment for improving educational systems in developing countries. Most recently Dr. Wagemaker served as an advisor to the Minister of Education for the Sultanate of Oman. He is also a member of the Advisory Board for the Center for Education Statistics and Evaluation (CESE) for the government of New South Wales, Australia, the H Institute, Beirut, Lebanon, and continues in an advisory role with IEA.

Dr. Wagemaker holds BA and MA degrees from the University of Otago, New Zealand, and a PhD from the University of Illinois, where he was awarded a University Fellowship and, in 2009, the College of Education's Distinguished Alumni Award.

Chapter 1
Introduction to Reliability and Validity of International Large-Scale Assessment

Hans Wagemaker

Abstract Although international large-scale assessment of education is now a well-established science, non-practitioners and many users often substantially misunderstand how large-scale assessments are conducted, what questions and challenges they are designed to address, and how technologies have evolved to achieve their stated goals. This book focuses on the work of the International Association for the Evaluation of Educational Achievement (IEA), with a particular emphasis on the methodologies and technologies that IEA employs to address issues related to the validity and reliability (quality) of its data. The context in which large-scale assessments operate has changed significantly since the early 1960s when IEA first developed its program of research. The last 60 years has seen an increase in the number of countries participating, with a concomitant expansion in the cultural, socioeconomic, and linguistic heterogeneity of participants. These quantitative and qualitative changes mean that the methodologies and assessment strategies have to evolve continuously to ensure the quality of data is not compromised. This chapter provides an introductory overview of the chronology and development of IEA's international large-scale assessments.

Keywords International Association for the Evaluation of Educational Achievement (IEA) · International large-scale assessment (ILSA) · Large-scale comparative assessment

1.1 Introduction

Since its founding in 1958, IEA has conducted a significant program of research on understanding student achievement in an international context. Its large-scale, comparative studies of education systems are firmly embedded in the educational policy landscape in many countries and sub-national educational jurisdictions. These

H. Wagemaker (✉)
International Association for the Evaluation of Educational Achievement (IEA), Amsterdam, The Netherlands
e-mail: hanswagemaker@compuserve.com

H. Wagemaker (ed.), *Reliability and Validity of International Large-Scale Assessment*, IEA Research for Education 10, https://doi.org/10.1007/978-3-030-53081-5_1

studies of education and schooling have made a significant contribution, not only to advancing understanding of learning outcomes and their antecedents but also to the development of the methodologies that have advanced the science of this field. Furthermore, in keeping with its stated aims, IEA has contributed significantly to the development of the wider research community and to capacity building, particularly with regard to those nations, institutions, and individuals charged with and committed to enhancing the performance of educational systems (Aggarwala 2004; Elley 2005; Gilmore 2005; Lockheed and Wagemaker 2013; Wagemaker 2011, 2013, 2014).

Throughout what is now the considerable history of IEA's international large-scale assessments (ILSAs), a constant concern has been to provide data of the highest possible quality. Researching the complex multi-national context in which IEA studies operate imposes significant burdens and challenges in terms of the methodologies and technologies that are required to achieve the stated study goals to a standard where analysts and policymakers alike can be confident of the interpretations and comparisons that may be made. The demands of the twin imperatives of validity and reliability, tempered by a concern for fairness, must be satisfied in the context of multiple, and diverse cultures, languages, scripts, educational structures, educational histories, and traditions.

Furthermore, as data become more generally accessible and the use of published analyses become more intimately woven into the fabric of the educational reform and policy process, greater attention needs to be paid to more nuanced interpretations of study outcomes.

Each stage of the development and execution of ILSAs, from the framework development that outlines the fundamental research questions to be addressed and the parameters of the study in question (subject, population, and design), to the methodologies used in the conduct of the study (sampling, field operations, quality control, and analysis), to reporting, raise important challenges and questions with respect to ensuring the ultimate reliability and validity of the assessment.

1.2 Outline of This Book

This book offers a comprehensive analysis of the science underpinning all IEA's ILSAs. The content bridges gaps in the general knowledge of consumers of reported outcomes of IEA studies, providing readers with the understanding necessary to properly interpret the results as well as critical insight into the way in which IEA, through its expressed methodologies, has addressed concerns related to quality, with particular reference to issues of reliability and validity.

To a large extent, the chapters that follow reflect the chronology of the development and execution of a large-scale assessment and provide a thorough overview of some of the key challenges that confront developers of IEA assessments and the strategies adopted to address them. While the issues and strategies outlined may also be common to other ILSAs, the focus of this volume is on the work of IEA; the chapters present examples that reflect some of the problems common to

most ILSAs, but include those that are unique to particular IEA studies. As it is not possible to cover every aspect of every study, readers interested in the detail of a particular study are encouraged to consult the related technical reports and documentation that are available on the IEA website (see www.iea.nl).

Following this introductory chapter, Chap. 2 discusses the related concepts of reliability and validity, and provides an overview of the research landscape as it applies to the work of IEA. It reflects on the continued evolution of the research landscape in terms of its magnitude and complexity and how these changes have impacted the process associated with the conduct of IEA's assessments.

The starting point of all assessment is the assessment framework, and Chap. 3 describes the purpose and rationale for developing assessment frameworks for large-scale assessments. As well as differentiating the different types of frameworks, Chap. 3 identifies the steps required for framework development, examining their coverage, publication, and dissemination, and how, procedurally, input from both participants and experts informs their development.

The way in which assessment frameworks are realized in terms of test content is the focus of Chap. 4. Issues related to construct validity, appropriateness of stimulus material in an international context (content), item format, item types, fairness, ability to be translated, sourcing items, and item review processes are considered. Elements like field testing, scorer training, and ultimately, the interpretation of field test outcomes and item adjudication are also addressed.

Chapter 5 embraces the various issues and aspects driving the development of background questionnaires. As for the cognitive tests, item formats, development of indicators, item response coding, measurement of latent constructs, quality control procedures, delivery format options, and future developments are all addressed in this section.

In the multicultural, multilingual, multi-ethnic environment in which ILSAs operate, issues related to translation from the source language to the national test language(s) have fundamental effects on validity and reliability. Chapter 6 addresses the procedures and approaches that IEA has adopted to secure high-quality test instruments in the participants' language of instruction.

For each study, IEA aims to derive reliable and valid population estimates that take into account the complex sample and assessment designs that characterize most studies. Chapter 7 addresses the sampling, weighting, and variance estimation strategies that IEA has developed to take into account and minimize potential sources of error in the calculation of such things as country rankings and subpopulation differences, describing the quality control and adjudication procedures that are key to the construction of validity arguments.

While Chap. 7 focuses on the processes associated with the population definitions and sampling, Chap. 8 focuses on the methodological concerns related to data collection, examining the perspectives of the national study centers for each participating country as well as the views of the international study center. The challenges of monitoring and ensuring data quality are fundamental to ensuring the overall quality of study outcomes and satisfying the imperatives of reliability and validity.

Concerns about data quality are not confined to the methods used to capture data in the field, but extend to the scoring of test booklets and the processing of the questionnaire data. Chapter 9 outlines some of the challenges of quality control, examining issues of comparability and describing the reliability checks that IEA employs to ensure data quality.

New e-technologies for assessment and data collection play an increasingly important role in ILSAs. Chapter 10 examines how those technologies are used to enhance the quality of assessments, their potential for improving measurement and data collection, and the advent of the new knowledge and skill domains related to information technology. In addition to establishing what is best practice in contemporary IEA studies, Chap. 10 explores not only the challenges of developing and delivering technology-based assessments in an international comparative context but also reflects on the potential that e-technologies offer in solving some of the ongoing challenges surrounding ILSA.

Chapter 11 explores the statistical foundations of ILSA and provides an explanation of how the methodologies adopted by the IEA can be used to describe and model individual and system level differences in performance in large-scale surveys of learning outcomes. Beginning with a brief historical overview of the development of latent variable models, this chapter stitches together a critical component of the validity argument as it relates to measurement and analysis in ILSA.

Whereas Chaps. 2 through 11 are intended to address and establish the reliability and validity of the data from a technical point of view, Chaps. 12 through 15 are concerned with aspects related to the issue of consequential validity, namely, the steps taken to enhance impact and mitigate the worst excesses related to misinterpretation, or overinterpretation of IEA's data.

Chapter 12 provides a brief overview of IEA's publication and dissemination strategy, including the quality control procedures for IEA publications. Chapter 13 describes the training and capacity building opportunities that IEA provides to study participants and interested researchers. IEA's investment in skill development aims to provide researchers with sufficient understanding to define good research questions and produce analyses that address issues of interest that are both methodologically and analytically defensible. In keeping with this theme of consequential validity, Chap. 14 provides a case study of best practice from Singapore, demonstrating how data from ILSAs can be used to inform educational policy development and reform.

Finally, Chap. 15 explores the limits and possibilities of using ILSA data to inform that policy process, reminding readers of the need for continual development of the science and proposing a novel way forward in the search for enhancing impact.

References

Aggarwala, N. (2004). *Quality assessment of primary and middle education in mathematics and science (TIMSS).* Evaluation report RAB/01/005/A/01/31. Takoma Park, MD: Eaton-Batson International. https://www.iea.nl/publications/technical-reports/evaluation-report.

Elley, W. B. (2005). How TIMSS-R contributed to education in eighteen developing countries. *Prospects, 35,* 199–212. https://doi.org/10.1007/s11125-005-1822-6.

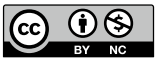
Gilmore, A. (2005). *The impact of PIRLS (2001) and TIMSS (2003) in low- and middle-income countries: An evaluation of the value of World Bank support for international surveys of reading literacy (PIRLS) and mathematics and science (TIMSS).* Washington, DC: World Bank. https://www.iea.nl/publications/study-reports/international-reports-iea-studies/impact-pirls-2001-and-timss-2003-low.

Lockheed, M. E., & Wagemaker, H. (2013). International large-scale assessments, thermometers, whips or useful policy tools? *Research in Comparative and International Education, 8*(3), 296–306. https://doi.org/10.2304/rcie.2013.8.3.296.

Wagemaker, H. (2011). IEA: International studies, impact and transition. In C. Papanastasiou, T. Plomp, & E. Papanastasiou (Eds.), *IEA 1958–2008: 50 years of experiences and memories* (pp. 253–273). Nicosia, Cyprus: Cultural Center of the Kykkos Monastery. https://www.iea.nl/publications/iea-reference/iea-1958-2008.

Wagemaker, H. (2013). International large-scale assessment (ILSA) programs and the challenges of consequential validity. In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability and equity* (pp. 217–233). Bingley, UK: Emerald Publishing.

Wagemaker, H. (2014). International large-scale assessments: From research to policy. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *A handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 11–37). London, UK: Chapman & Hall/CRC Press.



Chapter 2
Study Design and Evolution, and the Imperatives of Reliability and Validity

Hans Wagemaker

Abstract During the six decades since the International Association for the Evaluation of Educational Achievement (IEA) began its program of studies, an ever-growing political impetus worldwide for improved educational provision has stimulated countries' willingness to participate in international large-scale comparative assessments of learning outcomes. However, research within the complex multinational context that IEA operates in has resulted in significant methodological and technological challenges for the researchers and practitioners endeavoring to meet the goals of IEA studies. Such studies must satisfy the twin imperatives of validity and reliability, rarely an easy task given the multiple and diverse cultures, languages, scripts, educational structures, educational histories, and traditions of the countries and jurisdictions that participate. An appreciation of IEA's underlying assessment philosophy is fundamental to understanding the Association's assessment goals and the key design features of its studies, and what these mean with respect to ensuring that the studies satisfactorily address the demands of reliability and validity.

Keywords Assessment · Comparative assessment · International association for the evaluation of educational achievement (IEA) · International large-scale assessment (ILSA) · Reliability · Validity

2.1 Introduction

The assessment landscape in which IEA conducts its studies imposes some key contextual factors that have influenced IEA's research designs and execution. The science (and its evolution) underlying the construction and implementation of IEA's international large-scale assessments (ILSAs) must be understood in the context of the Association's philosophical and theoretical underpinnings and its clear focus on

H. Wagemaker (✉)
International Association for the Evaluation of Educational Achievement (IEA), Amsterdam, The Netherlands
e-mail: hanswagemaker@compuserve.com

H. Wagemaker (ed.), *Reliability and Validity of International Large-Scale Assessment*, IEA Research for Education 10, https://doi.org/10.1007/978-3-030-53081-5_2

school-based learning. Concerns with fairness and adherence to the imperatives of reliability and validity have also shaped the development of all IEA assessments.

2.2 Decisions Informing the Design of IEA Studies and Test Development

IEA's primary mission is to provide educational policymakers and researchers in the participating countries with an understanding of the factors associated with the quality of teaching and learning processes. Rather than focusing its assessments on the performance of a particular age cohort, IEA measures what students have learned after a fixed period of schooling, and seeks to understand the linkages between (a) the intended curriculum (the curriculum dictated by and described according to policy), (b) the implemented curriculum (that which is taught in schools), and (c) the achieved curriculum (what students actually learn). IEA accordingly centers its studies on classrooms and curricula (i.e., grade-appropriate content knowledge, skills, attitudes, and dispositions). However, because IEA studies operate within a comparative, multicultural context, the people responsible for developing and implementing them always face the challenge of how to prepare assessments that can be used cross-nationally and do not unfairly favor one country's curriculum.

Furthermore, the broad international space in which IEA operates, and the diversity this represents in terms of culture, languages, economic development, and educational development, inevitably places constraints on what can be assessed and how concerns about validity and reliability for such assessments can be satisfied. Although IEA's early studies featured English, French as a foreign language, written composition, and pre-primary education, IEA's long-term central emphasis has been on the foundational skills of literacy, mathematics and science, and civic and citizenship education. In the 1980s, the introduction of computer technologies in schools saw a new assessment centered on information and computer technologies. These four subject-matter areas are now regularly assessed (usually every four or five years), with each iteration featuring design changes that make it possible to produce trend data; they continue to shape IEA's core assessment strategies. The core studies are augmented from time to time with studies in areas such as early childhood education and the preparation of mathematics teachers.

The distinguishing feature of current IEA studies, and one that has been adopted by all major ILSAs, with the exception of the assessments carried out by the Organisation for Economic Co-operation and Development (OECD), is the focus on grade rather than an age cohort as the unit of analysis. Developers of ILSAs are confronted with a somewhat intractable challenge when determining how to sample populations of interest. Because age is approximately normally distributed and because countries operate different enrollment policies with respect to school starting age, an age-based sample means that the students selected will have different amounts of schooling, a variable that can compromise analyses because it presents a potentially significant grade effect (Wagemaker 2008). Similarly, a sample based on grade is subject to a

potential maturational effect because differences in school entry policies can result in students of different ages being placed in the same grade.

However, careful population definitions and sampling procedures avoid the most egregious errors related to maturational effects. As an organization focused on determining school-based learning outcomes, IEA's position is that students acquire the knowledge, skills, and dispositions that are the focus of the school curriculum through attending school rather than by getting older.

In addition, because schooling is organized on the basis of grades, successive periods of instruction reflect a progressively advanced (and in some cases hierarchical) organization of subject-matter, learning of which is generally not easily acquired outside of schooling and is certainly not a simple outcome of maturation. As Cliffordson (2010) found in a study of Swedish data from IEA's Trends in International Mathematics and Science Study (TIMSS), the effects of schooling on learning are twice as large as maturational effects.

Grade-based sampling has other analytical benefits from a research perspective. Careful design allows linkages between classes, teachers, teachers' instructional practices, and students' learning outcomes. The utility of the data gained from this approach is largely lost though when the linkage between teacher and students is severed, as is largely the case with an age-based approach.

Having decided to select grade and classroom as the unit of analysis, IEA needed to address the question of which grades to study and why. The Association's answer to this question is fundamental to understanding the design of its ILSAs. The studies typically focus on grades 4 and 8, and also 12 in the case of TIMSS. IEA considers grade 4 to be the point at which large-scale assessment methodologies can be reliably employed because students have acquired sufficient reading fluency to be able to read and answer written questionnaires. In many countries, grade 8 represents the transition out of lower primary school (and the end of compulsory schooling in some countries), while grade 12 (or equivalent) typically marks the end of secondary schooling.

The four-year intervals, as with TIMSS in particular, permits analysis of relative change in educational outcomes over time for the same cohort of students as they progress from grade 4 through to 8 or from grades 4 and 8 through to grade 12. For IEA's assessment of reading literacy, the Progress in International Reading Literacy Study (PIRLS), grade 4 represents the point at which students in most countries are making the transition from learning to read to reading to learn. At this stage, students who have not acquired basic reading skills will struggle with the other subject-matter areas they are required to master.

2.3 Addressing the Challenges

Designing a research project that attempts to provide understanding of the linkages described above presents many challenges, particularly given the interest in observing changes over time in both the antecedent factors that are implicated in achieve-

ment and in educational achievement itself. The different aspects of curriculum that interest IEA are assessed through two types of instrument: cognitive tests that are based on careful analysis of the curricula of participating countries and are designed to measure student knowledge of the target populations, and extensive questionnaires that capture information about student and teacher attitudes and dispositions, instructional practices, and more general background information related to schools and teachers as well as students and their homes. Information related to national or jurisdictional educational policy is also gathered from the studies' national research coordinators.

2.3.1 Governance and Representation (Validity and Fairness)

Concerns relating to both fairness and validity require IEA studies to be supported by a governance and management structure that secures national or local fidelity and ensures the assessments do not privilege one participating country over others. As a non-governmental organization (NGO), IEA works through a governance structure where study selection is determined through the Association's General Assembly, the policy body that represents the membership of the organization. However, IEA also enfranchises all study participants (including non-members) by including them in decisions relating to study design, content, execution, and reporting.

All IEA studies are coordinated through international study centers. The centers typically are responsible for overall management of studies and for orchestrating the contributions from collaborating centers of excellence. Unlike the early IEA studies, such as Reading Literacy, where one center conducted all aspects of the project, operational responsibilities since 1990 have usually been distributed among several research organizations.

This change reflected not only the development of specialist expertise within IEA but also the desire to ensure study centers would have access to the best expertise available. In the case of TIMSS and PIRLS, for example, the management group today consists of the following: Boston College as the international study center, which is responsible for the overall study design and test development; IEA responsible for most of the field operations, sampling, data management, translation services, study participation, country recruitment, and dissemination of results; the Educational Testing Service, which brings scaling support and expertise; and Statistics Canada, which provides sampling expertise and adjudication not found elsewhere. Similar structures now exist for all studies, including IEA's International Civic and Citizenship Education Study (ICCS), and its International Computer and Information Literacy Study (ICILS).

Participation in studies is open to any country and/or subnational entity that has responsibility for the administration of education. The latter typically includes major municipalities, states, or regions from within countries that may or may not participate as a nation. Participants appoint national research coordinators (NRCs) to represent their interests, contribute to study decision making, and take responsibility for all

aspects of study execution for the nation or region that they represent. This aspect of the IEA studies denotes an important governance distinction between the IEA and the OECD. The national project managers for OECD studies function solely as project managers; policy decisions are made by the Board of Participating Countries.

IEA also appoints groups of people expert in subject-matter content and/or item development and questionnaire design. The purpose of these representative groups is twofold: to ensure input from countries is as wide as possible, and that the pool of expertise is broader than what any one country, linguistic community, or educational tradition can provide. All participating countries have opportunity to review, elaborate, and approve the study frameworks drafted by subject-matter experts. Similarly, the expert groups for both subject matter and instrument development send, via the international study centers, their assessment items and questions to all study participants for their input, review, and approval. Item preparation is likewise the product of a collaborative process that includes item-development workshops. Finally, all participants review, agree on, and approve the content of the studies' international reports,

2.3.2 Reliability and Validity

The related constructs of reliability and validity constitute the foundational pillars of research. For those developing ILSAs, the need to ensure that these "pillars" are integral to each stage of the ILSA research process presents an ongoing methodological challenge.

Reliability is one of the key metrics for judging data quality. Reliability can be defined here as the degree to which a measurement or calculation can be considered accurate. Unlike validity, where evaluative judgments are based on the accumulation of evidence, reliability requires test developers to employ multiple measures to reassure the research analysts, policymakers, and other ILSA stakeholders that the assessment remains reliable throughout the various stages of the study's execution.

Later chapters in this book document and discuss the variety of measures that can be and are used to assess the reliability of the various components of a large-scale assessment. However, one example from classical test theory is useful here. Classical test theory is based on the premise that a test-taker's observed or attained score is the sum of a true score and an error score. As such, the precision of measurement cannot be deemed uniform across a measurement scale: the best estimates available are for the test-takers who exhibit moderate score levels, while less accurate measures pertain to those individuals who attain the high or low scores. Therefore, a primary goal of classical test theory is to estimate errors in measurement so as to improve measurement reliability and thus support the appropriate interpretation of test scores. Standard errors, the mean measure of dispersion of sample means around the population mean, constitute the "device" commonly used to guide score interpretation. For IEA, estimating and then publishing standard errors is a crucial

aspect of interpreting and understanding differences in achievement outcomes and in the proportions of responses to questionnaires among the participating populations.

Item response theory (IRT) extends this notion of reliability from a single index to an information function based on item difficulty and person ability. The item response function (IRF) provides the probability that a person with a given ability level will answer the item correctly. The ability to independently estimate item difficulty and person abilities is fundamental to trend analysis because scales are linked through secure items common to each assessment iteration. More recently, Rasch technology has been used to analyze questionnaire data during creation of latent constructs and subsequent modeling. A fuller exploration of the measurement procedures can be found in Chap. 11.

In simple terms, validity gives meaning to test scores. Validity evidence provides the reassurance that the assessment measures what it purports to measure. It describes the degree to which someone using an assessment can draw specific, realistic conclusions about individuals or populations from their test scores. However, as the literature on validity makes apparent, no single summary statistic or procedure can adequately satisfy validity-related concerns. Rather, as Messick (1989, p. 13) pointed out, "Validity is an integrated evaluative judgment of the degree to which empirical evidence and the theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment."

Crooks et al. (1996) argued that validity, which is enshrined in the standards of the American Educational Research Association (https://www.aera.net/), the National Council on Measurement in Education (https://www.ncme.org/), and the American Psychological Association (https://www.apa.org/), is the most important consideration informing the use and evaluation of assessment procedures. Despite this consensus on the importance of these considerations, the authors also observed that those developing and critiquing an assessment often neglect to evaluate the validity of the assessment.

Messick (1989) and Shepard (1993) opened up a major debate about the nature of validity when they proposed that assessment developers need to concern themselves with both the potential positive and negative consequences of testing and assessment (consequential validity). Messick, in particular, argued for appropriate labelling of tests to aid interpretation, a practice that is well established in IEA's reporting and publication strategies. Some of the key arguments and concerns surrounding the question of consequential validity are well reviewed in a monograph edited by Singer et al. (2018). The monograph featured the proceedings and conclusions of workshops sponsored by the National Academy of Education and shed light on the often misleading interpretation and use of ILSA data. In similar vein, Lockheed and Wagemaker (2013), in addressing the impact that participation in ILSAs has on developing countries highlighted some of the potential perverse effects on policy when data are misinterpreted.

IEA's consideration of these arguments and concerns are evident in the approach that the Association takes when preparing the reports that are the products of its ILSAs. For example, annotations to tables and graphs provide information that

readers and users of the reports need to take into account to avoid misinterpretation of the data, particularly when they are using that data to compare learning outcomes across countries or populations. Because of its complexity and breadth, validity remains a concept that is difficult to work with in practice and continues to challenge test developers and users alike. However, through its focus on fairness and impact in terms of educational reform and improvement, IEA seeks to ensure that the assessments do not become the goal in and by themselves. As the studies evolve, IEA is using the learning accruing from the experience of earlier cycles together with strong theoretical arguments to enhance construct validity.

2.3.3 Changing Contexts

Although the IEA's theoretical and philosophical underpinnings provide the basis for defining the broad parameters of each of its studies, those underpinnings continue to raise fundamental methodological challenges because of the comparative, international context in which the studies are conducted. These challenges are further heightened by the fact that this context is continually evolving. In short, these changes have placed significant demands on many aspects of assessment design and operations (translation, sampling) and test fidelity (validity).

As already noted, IEA's large-scale assessments have been administered for more than 60 years. This sustained period of endeavor has often produced the belief that this field of research, its design and methodology, is, and has been, relatively settled. However, when we view this period retrospectively, we can see the dramatic changes that have occurred during this period. As those who have had a protracted involvement with this research and its related developments know, change over the last six decades has been the only constant. Understanding how these changes have affected and shaped IEA's ILSAs is fundamental to understanding the challenges associated with conducting high-quality comparative international assessments of learning outcomes.

2.3.4 Compositional Changes

The comparative assessments that IEA began in the 1950s were characterized by participation by some of what are now OECD countries, primarily European and North American. For example, the First International Mathematics Study (FIMS) included 12 countries: Australia, Belgium, England, Finland, France, Germany (at that time the Federal Republic of Germany), Israel, Japan, the Netherlands, Scotland, Sweden, and the United States (IEA 2020a). What began as investigations of the naturally occurring variation among countries' educational policies and practices became the subject of increasing interest from policymakers and researchers

alike because of several important influences, including changes in the global policy discourse on education.

The period of reform that followed the release of studies such as the Second International Mathematics Study (SIMS; Robitaille and Garden 1989), particularly in the more economically and educationally advanced countries, prompted a growing international interest in ILSAs. A key element in this growth was the belief that the performance of a country's educational system could advantage or disadvantage that country's ability to successfully compete in an increasingly globalized economy. Education was essentially viewed as one of the main means whereby social and economic inequities could be mitigated.

One of the more dramatic expressions of this belief was evident in a report from the United States National Commission on Excellence in Education (USNCEE), published in 1983. *A Nation at Risk* (USNCEE 1983) suggested that the threat of United States (US) economic decline was of greater importance than perceived threats from aggressor nations. The authors of the report cited the apparent decline in US educational standards, as evidenced by US students' achievement scores in studies such as SIMS, as the cause of economic decline in the face of intensified global competition.

Globally, the ILSA reform agenda began to reflect a gradual shift in focus from concerns with participation in the studies and universal access to their results to a greater emphasis on equity, efficiency, and quality. Debates at the international level reflected an increasing willingness to tackle the issue of quality of assessment outcomes, and brought an even greater interest in and intensity to participation in ILSAs. While, in 2000, UNESCO through Goal 2 of the Millennium Development Goals (MDGs) focused on the achievement of universal primary education by 2015, the Sustainable Development Goals (SDGs) acknowledged that the quality of learning as much as full participation in it was vital to ensuring strong national development (UN [United Nations] 2015).

2.3.5 Financial Support

At the global level, the change in discourse from one focused on universal primary education to one centered on the quality of learning outcomes contributed to how funding for ILSAs was secured. Among the OECD countries that had achieved universal primary education, the concern was now less about "how many" than about "how well." However, in time, the less advanced economies also embraced quality of learning outcomes, and even more so when global development agencies also endorsed this change in emphasis. From this point on, these agencies helped fund the assessments.

The US Department of Education through the National Center for Education Statistics (NCES) came to play a leading role in funding ILSAs. Unique among statistical agencies, the NCES has a congressional mandate to collect, collate, analyze, and report complete statistics on the condition of American education. The mandate

also requires the center to conduct and publish reports and to review and report on education activities internationally (NCES 2020). It was through this vehicle that funding was secured for TIMSS 1995 and its iterations.

This support was critical at a time when the scale of operations was such that ILSAs could no longer rely on the goodwill of researchers and universities. This funding and the implementation of fees paid by study participants were subsequently leveraged through assistance from global development agencies such as the World Bank. The Bank was able to use its development grant facility to support the participation of up to 20 low- to middle-income countries. The United Nations Development Program (UNDP), in turn, directly supported a number of Middle Eastern and North African (MENA) countries. Both agencies saw education as key to national development, particularly in the low- to middle-income countries. Funding from these agencies not only brought a more secure financial foundation to the conducting of ILSAs but also gave further impetus to ILSA programs.

The changing demand for and the increased availability of financial support on participation in IEA large-scale assessments meant that, by 1990, country participation had grown in numbers and was being augmented by "benchmarking participants," that is, subnational entities such as states/provinces or major metropolitan areas (e.g., Buenos Aires) that were interested in securing information about the performance of their students within their respective administrative levels (Fig. 2.1).

In keeping with the focus on national development, much of which originated with organizations like the World Bank and the United Nations (through the UNDP), support for the low- to middle-income countries in Central and Eastern Europe and those from the MENA region was predicated on measuring and understanding performance in the foundational skills of reading, mathematics, and science, leading to a growth in participation in IEA's TIMSS (Fig. 2.2). A similar trend was experienced with enrollments for IEA's PIRLS.

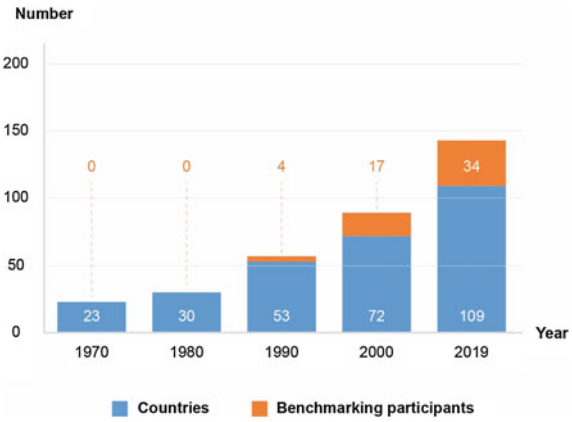


Fig. 2.1 Cumulative growth in unique participants in IEA studies over time. *Note* totals represent unique participants across all studies. The Second International Mathematics Study, for example, had only 20 participants, with outcome data collected from just 15 participants for the cross-sectional part of the assessment (see Robitaille and Garden 1989)

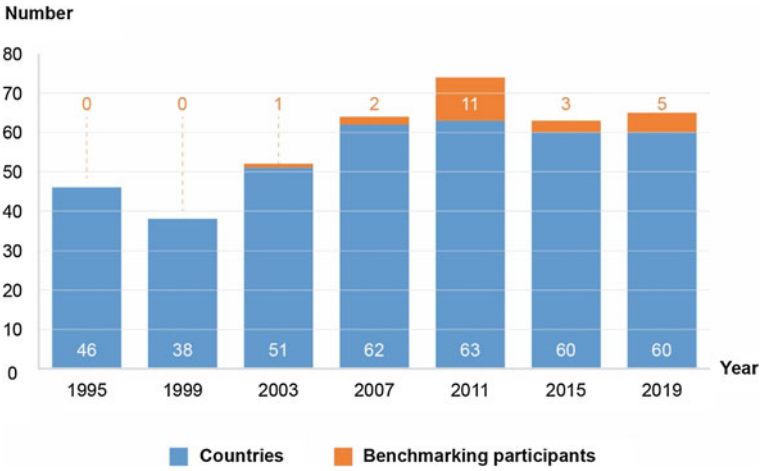


Fig. 2.2 Compositional changes and growth in participation in TIMSS, 1995–2015

2.3.6 Expansion in Assessment Activities

Along with the compositional changes in study participation came an expansion of assessment activities into other subject-matter areas (see Fig. 2.3, which also makes evident the development of the trend design for TIMSS and PIRLS since 1995).

The establishment of regular trend cycles, focused on understanding and measuring changes over time, for TIMSS and PIRLS, and later ICCS and ICILS, introduced a new era of assessment for IEA. But each of these trend assessments has unique characteristics that have increased the complexity of their design and analysis of their outcomes.

TIMSS, for example, assesses two subjects, mathematics and science, within one test administered to the same students; previously, the two subjects were assessed

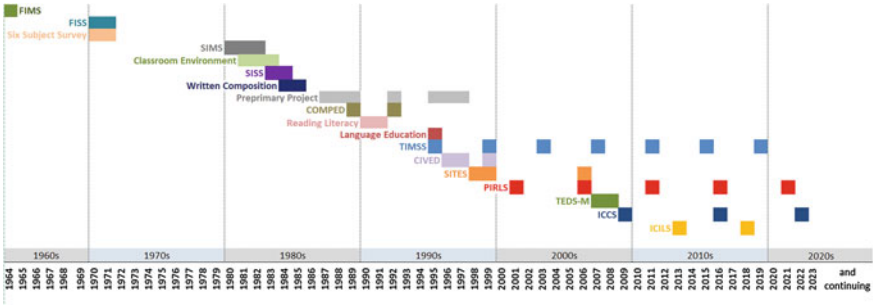


Fig. 2.3 Development of IEA's program of ILSAs from 1995 onwards. *Source* Hastedt (2020) Springer International Publishing, reproduced with permission of Springer Nature Switzerland AG. Disclaimer: this figure is copyright protected and excluded from the open access licence

separately. PIRLS, in an attempt to provide an authentic reading experience for the students being assessed, includes a "reader" booklet as a distinct element. ICCS developed "regional modules," namely assessments adapted for groups of similar countries (for example, European, Asian, and Latin American) designed to reflect distinct regional, cultural, and political interests. Finally, ICILS, over successive iterations, has evolved to a fully computer-based assessment delivery and includes modules that explore students' ability to solve programming problems. The more recent iterations of TIMSS and PIRLS have also provided computer-based delivery options for participants.

2.3.7 Heterogeneity

In addition to addressing the issue of scale, these changes in participation brought in a greater diversity of languages and scripts (ideo vs logo vs phonography, as well as directionality), culture (non-Western, non-European), ethnicity, and educational structures and conventions (e.g., school starting age). They have also allowed for greater variance in student ability within a given target grade. The 2019 TIMSS assessment, for example, tested students in 50 languages and seven orthographies. Furthermore, as countries have become more aware of the need to address issues related to internal diversity, an increasing number of them (29 in TIMSS 2019) are administering the assessment instruments in more than one language (including South Africa, which assessed students in 12 languages in TIMSS 2019) (IEA 2020b).

New sampling strategies have also been needed to accommodate countries with complex population structures based on, for example, ethnicity, and for studies assessing new populations (i.e., populations not previously part of the IEA studies), such as teachers. Modalities associated with increased use of computers and computer-based assessment (CBA), such as those evident in ICILS, have demanded the application of new technologies in complex environments and again challenged existing strategies for ensuring the quality and fairness of the assessment. The more obvious examples of these challenges include the variability in students' familiarity with computers and other digital technologies within and among countries.

2.3.8 Advances in Technology

Technological advances in assessment technologies have introduced another set of complexities to IEA's program of assessments. One of the most important concerns test design. The assessment design for early IEA studies was based on a common test format and a psychometric model that was state-of-the-art at the time (classical test theory and relatively simple regression models). Mean achievement scores on their own, however, are limited in their ability to inform policy. Real insight comes

from understanding the associations between learning outcomes and the policy-related antecedents of learning. Similarly, the demand for greater insight into relative strengths and weaknesses in achievement led not only to achievement being seen in terms of both content and cognitive domains but also to the development of the more complex assessment designs that enable in-depth exploration of the linkages between learning outcomes and their antecedents.

To avoid an unacceptable testing burden and to increase the information yield, IEA, borrowing from the experience of the US' National Assessment of Educational Progress (NAEP), adopted, soon after completion of the 1990 Reading Literacy Study, a so-called balanced incomplete block (BIB) spiraling test design (see Johnson 1992), which achieved greater curriculum coverage and greater fidelity in terms of information yield. The trade-off for these gains, though, was greater administrative and analytical complexity. IEA's associated transition to Rasch and other IRT-based assessment models (Smith and Smith 2004), including the use of multiple imputation technologies in the early 1990s, signaled another major development in assessment design, but the improvement in measurement again brought with it further technical challenges. These changes in assessment design have all been associated with and made possible by the rapid advances in computing power.

2.3.9 Assessment Delivery

The availability, introduction, and application of more technologically advanced modes of assessment delivery, including computer-based assessments (CBA), resulted in yet another complexity challenge. Addressing this development has required robust transition arrangements that balance the need to adapt to computer-based delivery systems with the reality that some participating countries or the jurisdictions within them still need to use paper and pencil versions of the assessments.

Establishing equivalence and managing potential effects due to the mode of test delivery continues to be a critical concern, and it is one addressed in detail later in this book. However, as IEA transitions from pencil and paper to CBA, it is having to adjust to the need to accommodate those countries or jurisdictions that are in an earlier stage of educational development or those that wish to assess with greater fidelity the lower boundaries of the performance distribution. In 2019, IEA offered TIMSS in several formats (IEA 2020b).

TIMSS Numeracy was developed for the less educationally-developed jurisdictions; standard TIMSS was available in paper and pencil versions for both grade 4 and grade 8 mathematics and science in addition to electronic versions for the two grades; and eTIMSS comprised electronic versions at both grades 4 and 8 for those jurisdictions wishing to pilot the computer-based technology.

IEA's careful management of the transition from paper and pencil to CBA for TIMSS 2019 has been challenging (see Table 2.1). The complexity of the

Table 2.1 Number of education systems (including benchmarking participants) participating in each assessment delivery mode offered for TIMSS 2019 (in total 64 education systems participated at grade 4 and 46 at grade 8)

	Assessment mode						
	TIMSS numeracy	TIMSS 2019					
		TIMSS pencil and paper		e-TIMSS		e-TIMSS pilot	
	Grade 4	Grade 4	Grade 8	Grade 4	Grade 8	Grade 4	Grade 8
Number of participating education systems	11	17	19	36	27	21	11

process was heightened by the need to accommodate, through TIMSS Numeracy, the requirements of the less educationally-advanced countries.

As part of the PIRLS reading assessment, IEA introduced pre-PIRLS in 2011 and, in 2016, offered this assessment as PIRLS Literacy, together with a computer-based version of PIRLS called ePIRLS.

2.4 Conclusions

The decades following the release of the earliest IEA studies, such as FIMS and SIMS, heralded a new era of large-scale assessment. A global shift in concerns about educational achievement relative to a country's wellbeing and economic standing stimulated countries' willingness to participate in ILSAs. That willingness was further facilitated by the advent of funding from major donor organizations, as well as countries' acceptance that they also needed to help meet their costs of participation. This growth in participation was not merely a problem of scale, however. Rather, the challenge came from the need for the assessments to address greater heterogeneity across and within the participating countries and all that this implies for the reliability of the assessment instruments and the validity of the data emanating from them. While technological advances in the science of large-scale assessment and computing resources have enabled more complex designs and greater fidelity with respect to analyzing, modeling, and interpreting the results of the studies, these changes have also increased complexity (i.e., significant measurement and operational challenges) across all stages of the development and execution of ILSAs.

The increased recognition worldwide that ILSAs can contribute to educational reform and improvement at the local level stimulated participation not only in the work of IEA but also in that of other testing organizations. As the data from international assessments began to receive more attention from researchers and policymakers, so, too, did the quality and interpretation of ILSA data (see Chatterji 2013; Singer and Braun 2018; Wagemaker 2013). This scrutiny continues to this day.

Validity and reliability arguments and measures must therefore reassure study stakeholders that each ILSA satisfactorily addresses concerns related to comparisons of data at the international level and is appropriate in terms of judgments made about learning outcomes at the local level. In essence, the challenge for everyone associated with ILSAs is to satisfy the twin imperatives of international comparability and local relevance. Also, if educational reform and improvement are indeed the ultimate goal of ILSAs, then educational providers must be charged with an extra duty of care, that of ensuring, to the greatest extent possible, that the use and reporting of data meets the quality tests of reliability and validity. With all this in mind, the chapters that follow outline IEA's responses to and strategies for ensuring reliability, thus providing the basis for constructing arguments in favor of validity.

References

Chatterji, M. (Ed.). (2013). *Validity and test use: An international dialogue on educational assessment, accountability and equity*. Bingley, UK: Emerald Publishing.

Cliffordson, C. (2010). Methodological issues in investigations of the relative effects of schooling and age on school performance: The between-grade regression discontinuity design applied to Swedish TIMSS 1995 data. *Educational Research and Evaluation, 16*(1), 39–52. https://doi.org/10.1080/13803611003694391.

Crooks, T. J., Kane, M. T., & Cohen, A. (1996). Threats to the valid use of assessments. *Assessment in Education: Policy and Practice, 3,* 265–285. https://doi.org/10.1080/0969594960030302.

Hastedt, D. H. (2020). History and current state of international student assessment. In H. Harju-Luukkainen, N. McElvany, & J. Stang (Eds.), *Monitoring of student achievement in the 21st century* (pp. 21–37). Cham, Switzerland: Springer International Publishing. https://www.springer.com/gp/book/9783030389680.

IEA. (2020a). *FIMS. First International Mathematics Study [webpage]*. Amsterdam, the Netherlands: IEA. https://www.iea.nl/studies/iea/earlier#section-170.

IEA. (2020b). *TIMSS 2019: Trends in International Mathematics and Science Study 2019 [webpage]*. Amsterdam, the Netherlands: IEA. https://www.iea.nl/studies/iea/timss/2019.

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement, 29,* 95–110.

Lockheed, M. E., & Wagemaker, H. (2013). International large-scale assessments thermometers, whips or useful policy tools? *Research in Comparative and International Education, 8*(3), 296–306. https://doi.org/10.2304/rcie.2013.8.3.296.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *The American Council on Education/Macmillan series on higher education. Educational measurement* (pp. 13–103). New York, NY: Macmillan Publishing Co, Inc; American Council on Education.

NCES. (2020). *National Center for Education Statistics: About us [webpage]*. https://nces.ed.gov/about/.

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

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (Vol. 19, pp. 405–450). Washington, DC: AERA.

Singer, J. D., & Braun, H. I. (2018). Testing international education assessments. *Science, 360*(6384), 38–40. https://science.sciencemag.org/content/360/6384/38.

Singer, J. D., Braun, H. I., & Chudowsky, N. (Eds.). (2018). *International education assessments: Cautions, conundrums, and common sense.* Washington DC: National Academy of Education. http://naeducation.org/wp-content/uploads/2018/08/International-Education-Assessments-NAEd-report.pdf.

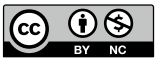
Smith, E. V., Jr., & Smith, R. M. (2004). *Introduction to Rasch measurement: Theory, models and applications.* Maple Grove, MN: JAM Press.

UN. (2015). *We can end poverty. Millennium development goals and beyond 2015 [webpage].* https://www.un.org/millenniumgoals/.

USNCEE. (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. (2008). Choices and trade-offs: Reply to McGaw. *Assessment in Education, 15*(3), 267–268. https://doi.org/10.1080/09695940802417491.

Wagemaker, H. (2013). International large scale assessment (ILSA) programs and the challenges of consequential validity. In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability and equity* (pp. 217–233). Bingley, UK: Emerald Publishing.



Chapter 3
Framework Development in International Large-Scale Assessment Studies

John Ainley and Wolfram Schulz

Abstract At the outset of an international large-scale assessment study, researchers typically develop an assessment framework that defines the outcomes (content, skills, and understandings) that are to be assessed in the study and provide a contextual framework that outlines the scope of contexts believed to be associated with those outcomes. A particular challenge is to establish frameworks for cross-national studies that are valid across a wide range of participating education systems. Some studies have assessment frameworks that reference established curriculum areas. Other studies have assessment frameworks that define outcomes related to learning areas of a cross-curricular nature. A third set of studies have assessment frameworks that put a greater emphasis on what people need to know to successfully participate in society rather than on what is taught in schools. Contextual frameworks, typically embedded in a broader assessment framework, focus on those characteristics that are believed to influence student learning either as teaching and learning processes or as the conditions provided for teaching and learning. Some large-scale international surveys have been entirely concerned with contexts and are based on frameworks that are primarily concerned with contexts. Framework development involves detailed definitions of the outcomes to be investigated in terms of content, cognitive processes, and affective dispositions, as well as the types of items to be used and aspects of assessment design. Frameworks are informed by existing research literature and available information about national and school curricula. The development process is informed by expert advisory groups and country representatives through a series of review cycles. Research questions that guide the study are formulated as part of these frameworks.

Keywords Contextual influences · Framework · Learning outcomes · Measurement · Study design

J. Ainley (✉) · W. Schulz
Australian Council for Educational Research (ACER), Camberwell, Australia
e-mail: john.ainley@acer.org

W. Schulz
e-mail: wolfram.schulz@acer.org

H. Wagemaker (ed.), *Reliability and Validity of International Large-Scale Assessment*, IEA Research for Education 10, https://doi.org/10.1007/978-3-030-53081-5_3

3.1 Introduction

Assessment frameworks define not only what is to be measured but also the how and the why of what is to be measured. The international large-scale assessments (ILSAs) of IEA (and those of other organizations such as the Organisation for Economic Co-operation and Development [OECD]) are all based on such an organizing document. Although assessment frameworks are also a common feature of national assessment surveys, they are especially important for ILSAs because of the need to transcend variations among participating education systems. In many ILSAs, there will be an assessment framework that defines the outcomes (content, skills, and understandings) that are to be assessed in the study, as well as a contextual framework that defines the contexts believed to be associated with those measured outcomes. In some studies these are not distinguished, but both elements are essential to underpin the quality of ILSAs and provide stakeholders, the research community, and a wider audience with a rationale for the study, the definitions and conceptualizations of what is measured and for what reason, and an outline of the study design.

A particular challenge specific to cross-national studies is to establish frameworks that are valid across the range of participating education systems. There is considerable diversity across national contexts in the way curricula are defined and implemented, how systems are structured and governed, and in which form teaching and learning take place within schools. The assessment framework is the main point of reference to understand how international studies define common elements of learning outcomes or contextual factors that should be measured, and how the study is designed to obtain comparable data across countries.

3.2 Assessment Frameworks

Assessment frameworks define the intended learning outcomes from an area that is to be assessed as well as the design of the assessment (Jago 2009). In other words, assessment frameworks reference the structure and content of the learning area, as well as defining the design of the assessment (including the population definitions) and detailing how the assessment is to be implemented. Furthermore, these documents are also of central importance as reference points to assess content and construct validity and, in cases of studies that monitor trends over time, they provide rationales for any innovations or changes of content and how these are integrated into the existing construct(s) (Mendelovits 2017).

An assessment framework details the constructs to be measured in terms of the content, skills, and understandings expected to have been learned by the population at the time of their assessment (e.g., grade 8 students). Pearce et al. (2015, p. 110) argued that an assessment framework provides a "structured conceptual map of the learning outcomes" for a program of study. The National Assessment Governing

Board (NAGB) of the United States Department of Education describes the mathematics assessment framework for the National Assessment of Educational Progress (NAEP) as being "like a blueprint" in that it "lays out the basic design of the assessment by describing the mathematics content that should be tested and the types of assessment questions that should be included" (NAGB 2019, p. 2). The Progress in International Reading Literacy Study (PIRLS; see IEA 2020) conducted by IEA similarly says that its assessment framework is "intended as a blueprint" (Mullis et al. 2009, p. 2).

Assessment frameworks are not the same as curriculum frameworks even though the two are related (Jago 2009; Pearce et al. 2015). Whereas a curriculum framework describes comprehensively what is to be taught in education, an assessment framework defines the construct or constructs being measured by a particular assessment and takes into account what can feasibly be assessed. Consequently, assessment frameworks often define the cognitive skills or processes needed by students to deal with relevant tasks in the respective learning area, and some, such as the International Civic and Citizenship Education Study (ICCS; see IEA 2020), also define affective and behavioral outcomes of relevance in the field.

The recent transition of long-term (cyclical) ILSA studies from paper-based to computer-based assessments has provided opportunities to include new measures of aspects with digital technology that could not be administered on paper. While this is an attractive feature of these new ways of assessment delivery, they may also have implications for the construct(s) measured over time. Together with mode effects, computer-enhanced measurement may have consequences for the monitoring of data across assessment cycles and frameworks have the role of describing the extent to which construct(s) may have broadened and include aspects that were not measured in previous cycles. Furthermore, there may also be other aspects that change between ILSA implementations (such as trends in curricular changes across countries or changes in relevance of particular topics for a learning area under study). The assessment framework serves as a reference point for describing any such changes and providing a rationale for them.

Assessment frameworks also incorporate sets of research questions concerned with the expected variations in achievement among and within countries and covariation with student, school, and system characteristics. Those research questions broadly define the analyses to be conducted and guide the structure of reporting.

3.3 Contextual Frameworks

ILSAs study contexts to aid understanding of variation in achievement measures. The constructs that characterize those contexts are elaborated as variables in contextual frameworks on the basis of relevant research literature. The contextual information necessary to provide the basis for measures of the constructs is also outlined in the framework, which is then used as a guide to for the development of questionnaire material (Lietz 2017; see also Chap. 5). For IEA's assessments, contextual

information is gathered through student, teacher, and school questionnaires, and, in some cases, from parent questionnaires, as well as from information about education systems provided by national research centers and/or experts.

Contextual variables can be classified in various ways. One approach to classification is in terms of what is being measured; for example, whether the variables are factual (e.g., age), attitudinal (e.g., enjoyment of the area), or behavioral (e.g., frequency of learning activities). Another approach recognizes that the learning of individual students is set in the overlapping contexts of school learning and out-of-school learning. A third approach to classifying contextual variables is based on the multilevel structure inherent in the process of student learning (see Scheerens and Bosker 1997; Schulz et al. 2016). It is possible to broadly distinguish the following levels from a multilevel perspective:

- *Context of the individual*: This level refers to the individual ascribed and developed characteristics of individual respondents.
- *Context of home and peer environments*: This level comprises factors related to the home background and the immediate social out-of-school environment of the student (e.g., peer-group activities).
- *Context of schools and classrooms*: This level comprises factors related to the instruction students receive, teacher characteristics, the school culture, leadership, and the general school environment.
- *Context of the wider community*: This level comprises the wider context within which schools and home environments work. Factors can be found at local, regional, and national levels.

Another distinction between types of contextual factors can be made by grouping contextual variables into antecedents or processes (see Fraillon et al. 2019; Schulz et al. 2016):

- Antecedents are those variables from the past that shape how student learning takes place. These factors are level-specific and may be influenced by antecedents at a higher level. For example, training of teachers may be affected by historical factors and/or policies implemented at the national level. In addition, educational participation may be an antecedent factor in studies beyond the compulsory years of schooling.
- Processes are those variables related to student learning and the acquisition of understandings, competencies, attitudes, and dispositions. They are constrained by antecedents and possibly influenced by variables relating to the higher levels of the multi-level structure.

Antecedents and processes are variables that have potential impact on the outcomes at the level of the individual student. Learning outcomes at the student level can also be viewed as aggregates at higher levels (school, country) where they can affect factors related to learning processes.

3.4 Design and Implementation

Being blueprints for ILSA studies, frameworks also describe assessment designs and implementation strategies. An assessment design will include specification of the balance of numbers of items across content and cognitive domains, the types of items (closed or constructed response) and details of a rotated block design for studies where such a design is to be employed. Since the first cycle of the Trends in International Mathematics and Science Study (TIMSS; see IEA 2020) in 1995, most ILSA studies have employed rotated block designs, in which students are randomly allocated booklets based on different blocks of items, so as to provide a more comprehensive coverage of the learning area than could be assessed with a single set of items. An assessment framework for an ILSA study would typically also indicate how scale scores are to be calculated from item response data to sets of items related to the constructs being measured.

An assessment framework may also include a description of available delivery modes (paper-based or computer-based). The International Computer and Information Literacy Study (ICILS; see IEA 2020) has been computer-based since its inception and requires relevant specification of item types (Fraillon et al. 2019). Computer-based methods have since become part of other ILSA studies. Mixed-mode methods, or changes in mode between cycles, require consideration of whether delivery mode influences student responses and how to evaluate any mode effects (see also Chap. 10).

Implementation strategies described in assessment framework documents include population definition, sample design, and administration procedures. For example, frameworks for IEA study populations would provide definitions in terms of the grade to be assessed (which typically includes either grade 4 or grade 8, or both grade 4 and grade 8), while frameworks for the OECD's Programme for International Student Assessment (PISA) define the population in terms of student age (15 years) and secondary school attendance at the time of testing. It is important that the framework provides explicit information about the population and whether exclusions are permitted (e.g., students with special educational needs). Most ILSA studies allow exclusions that meet specific criteria for up to five percent of the population.

Assessment frameworks also tend to provide outlines of the instrumentation and other design features (such as international options) to participating countries (e.g., including additional sets of questions as part of the questionnaires). It is further customary to provide examples of assessment material, which illustrate the way constructs are measured in the respective ILSA. By giving insights into the ways of measuring these constructs, ILSAs provide stakeholders, researchers, and the wider audience with a better understanding of their operationalization and enable them to assess their validity, in particular in cases where new innovative construct(s) are measured (see, e.g., Fraillon et al. 2019).

3.5 Steps in Framework Development

Framework development is based on, and makes use of, understandings of the research literature about the respective learning area and information about educational practice and policy in that field. These sources provide the basis for a proposal for research and help to define the scope of the framework, which guides the assessment and identifies the contextual influences to be measured along with the assessment. They help to identify the content and skills to be assessed in an ILSA study and the basis for the design and format of the assessment (Jago 2009). The process of reviewing literature and educational practice results in the formulation of research questions that become key elements of the assessment framework. At the broadest level this involves a definition of the main construct(s) an ILSA sets out to measure. For example, based on an extensive literature review, PIRLS defines reading literacy as:

> the ability to understand and use those written language forms required by society and/or valued by the individual. Readers can construct meaning from texts in a variety of forms. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment (Mullis and Martin 2015, p. 12).

A broad definition provides the basis for a detailed elaboration of the construct that is embedded in theory. For example, in PIRLS, reading for literary experience and reading to acquire and use information each incorporate four types of comprehension process: (1) retrieving explicitly stated information, (2) making straightforward inferences, (3) interpreting and integrating ideas and information, and (4) evaluating content and textual elements (Mullis and Martin 2015). Research literature, theory, and information about educational practice provide the basis for these elaborations. They also provide a basis for systematically characterizing important aspects of contexts and articulating the research questions for the respective ILSA.

Although the processes of reviewing literature are well established, processes for reviewing educational practice and policy are less clearly set out. ILSA studies frequently make use of reviewing national curricula and curriculum frameworks to augment the insights provided by literature reviews. Reviewing curricula through document analysis is clearer when there is a direct alignment between a discipline or subject and the construct being assessed. This is the case for mathematics and science as assessed in TIMSS (Mullis and Martin 2017). It is also evident when the capability being assessed is a key element of all primary school curricula and part of the language arts learning area as in the assessment of reading in PIRLS (Mullis and Martin 2015). It is less clearly evident when the capability (or assessment construct) being assessed crosses several curriculum areas as is the case in ICCS (Schulz et al. 2016) and ICILS (Fraillon et al. 2019). In studies that set out to measure so-called real-life skills, such as the OECD PISA study or the OECD Programme for the International Assessment of Adult Competencies (PIAAC), the respective frameworks need to provide a theoretical underpinning for the respective conceptualizations of construct(s) that are not referenced to any national curricular frameworks but rather to overarching formulations of expected knowledge and skills based on theoretical

models (OECD 2012, 2019). The increasing heterogeneity among participants in ILSA studies means that the processes to ensure inclusion when defining content and design are crucial to the validity of the assessments.

Structured opportunities for country commentary and expert advice provide important perspectives that contribute to framework development. ILSA studies conducted by the IEA incorporate reviews by national research coordinators (NRCs) so that the framework is applicable to all of the participating countries. NRCs meet in person several times during an ILSA study and communicate outside those meetings on a regular basis. It also essential to seek expert advice during the process of framework development. PIRLS has a reading development group and a questionnaire development group that contribute to framework development (Mullis and Martin 2015). TIMSS has a science and mathematics item review committee and a questionnaire item review committee that contribute to framework development (Mullis and Martin 2017). ICILS and ICCS each have project advisory committees providing expert advice across all aspects of these studies. OECD studies such as TALIS, PISA, and PIAAC have established similar expert groups to provide advice on the development of their respective frameworks.

3.6 Types of Framework

Among the range of ILSA studies there appear to be four main types of framework that reflect the nature of the domains of those studies.

3.6.1 Curriculum-Referenced Frameworks

TIMSS and PIRLS, the longest established of the ILSA studies currently conducted by IEA, are closely related to learning areas of school curricula. For example, the TIMSS curriculum model (Mullis and Martin 2017), which is also present in similar forms in other IEA studies, includes three aspects: the intended curriculum, the implemented curriculum, and the attained curriculum. These represent, respectively, the domain-related aspects that students are expected to learn, what is actually done in classrooms, the characteristics of those facilitating these learning opportunities, how it is taught and offered, what students have learned, what they think about learning these subjects, and how they apply their knowledge.

TIMSS is most clearly aligned with mathematics and science at grades 4 and 8. The assessment frameworks are organized around a content and a cognitive dimension. The content domains for mathematics at grade 4 are number (50%), measurement and geometry (30%), and data (20%), and the cognitive domains are knowing (40%), applying (40%), and reasoning (20%) (Mullis and Martin 2017). At grade 8, the content domains are number (30%), algebra (30%), geometry (20%), and data and probability (20%), and the cognitive domains are knowing (35%), applying (40%),

and reasoning (25%) (Mullis and Martin 2017). In summary there is less emphasis on number at grade 8 than there is at grade 4. At grade 8, there is a little more emphasis on reasoning, and a little less emphasis on knowing than at grade 4.

The content domains for TIMSS science at grade 4 are life science (45%), physical science (35%), and earth science (20%), and the cognitive domains are knowing (40%), applying (40%), and reasoning (20%) (Mullis and Martin 2017). At grade 8, the content domains are biology (35%), chemistry (20%), physics (25%), and earth science (20%) and the cognitive domains are knowing (35%), applying (35%), and reasoning (30%). In brief, at grade 8 there is a little less emphasis on biology (or life science) and a little more emphasis on physics and chemistry (physical science), as well as greater emphasis on reasoning and less on knowing than at grade 4. The context framework for TIMSS is organized around five areas: student attitudes to learning, classroom contexts, school contexts, home contexts, and community or national policies (Mullis and Martin 2017).

PIRLS is only designed as a grade 4 assessment and defines its content as focused on reading for literary experience and reading to acquire and use information, with each incorporating four types of comprehension process: retrieving explicitly stated information, making straightforward inferences, interpreting and integrating ideas and information, and evaluating content and textual elements (Mullis and Martin 2015). The framework specifies that 50% of the assessment is concerned with reading for literary experience and 50% is concerned with reading to acquire and use information (Mullis and Martin 2015). The framework identifies four comprehension processes: retrieving explicitly stated information (20%), making straightforward inferences (30%), interpreting and integrating ideas and information (30%), and evaluating content and textual elements (20%) (Mullis and Martin 2015). The PIRLS 2016 framework observes that, in the literature on reading, there has been a shift in emphasis from fluency and basic comprehension to demonstrating the ability to apply what is read to new contexts and that this has been reflected in the framework.

3.6.2 Frameworks for Measuring Outcomes in Cross-Curricular Learning Areas

While, in spite of cross-national diversity, there is considerable common ground across different national curricula with regard to subject areas such as mathematics or science, there are also learning areas that are cross-curricular and/or embedded in different subjects or subject areas. Two such learning areas that are increasingly regarded as highly relevant for education across a wide range of societies are students' competencies related to civic and citizenship education and their information and communication technology (ICT) related skills. The IEA has recognized the importance of investigating these two areas through the establishment of two continuous studies: ICCS (with its first two completed cycles in 2009 and 2016, and an upcoming

cycle in 2022) and ICILS (with its first two completed cycles in 2013 and 2018, and an upcoming cycle in 2023).

Given the diversity of approaches to these learning areas across national contexts (Ainley et al. 2013; European Commission/EACEA/Eurydice 2017), it was necessary to develop frameworks for each that were appropriate, relevant, and accepted by participating countries, while at the same time recognizing the existence of a wide range of different approaches to teaching and learning of relevant outcomes. To ensure the validity of frameworks that needed to outline common ground for measurement for these studies, it was necessary to implement an iterative process of reviews and revisions with input and feedback from a range of international experts and national centers in each participating country.

ICCS 2009 was established as a baseline study of ongoing comparative research of civic and citizenship education; it built on previous IEA studies of civic and citizenship education conducted in 1971 as part of the six-subject study (Torney et al. 1975) and in 1999 (the Civic Education Study [CIVED]; see IEA 2020; Torney-Purta et al. 2001). To measure cognitive civic learning outcomes, the ICCS assessment frameworks (Schulz et al. 2008, 2016) articulated what civic knowledge and understanding comprised in terms of two cognitive domains that distinguish knowledge related to both concrete and abstract concepts (knowing) from cognitive process required to reach broader conclusions (reasoning and applying). Furthermore, it distinguished four content domains: civic society and systems, civic principles, civic participation, and civic identities. For both content and cognitive domains, the definition focused on issues that could be generalized across societies and excluded nationally specific issues, such as those related to the particular form of government in a country (for details on the measurement of understanding of civics and citizenship, see Schulz et al. 2013).

For a study of civic and citizenship education it is also paramount to appropriately consider affective-behavioral outcomes. ICCS considers attitudes and engagement-related indicators as part of the learning outcomes that need to be measured through data from its student questionnaires. To this end, the assessment framework of this study defines two affective-behavioral domains (attitudes and engagement) that comprise different indicators of relevant learning outcomes. Furthermore, the contextual framework describes a wide range of factors at different levels (the individual, home, and peer context, the school and classroom context, and the context of the wider community ranging from local communities to supra-national contexts) distinguishing antecedents and process-related variables (Schulz et al. 2016).

While civic and citizenship education is a long-established learning area in a wide range of education systems, education for learning about ICT or digital technologies is a more recent development that followed their emergence as important for people's daily lives. Across many education systems the area has been acknowledged as of importance for young people's education, although there has been a diversity of approaches (see, e.g., Bocconi et al. 2016). ICT-related learning is often envisaged as a transversal or cross-curricular skill and ICT subjects are not consistently offered across countries (Ainley et al. 2016).

To capture to the essence of a number of previous conceptualizations, ICILS 2009 defined computer and information literacy (CIL) as "an individual's ability to use computers to investigate, create, and communicate in order to participate effectively at home, at school, in the workplace, and in society" (Fraillon et al. 2013, p. 17). While continuing the measurement of CIL in its second cycle, ICILS 2018 also included an optional assessment of computational thinking (CT), defined as referring to "an individual's ability to recognize aspects of real-world problems which are appropriate for computational formulation and to evaluate and develop algorithmic solutions to those problems so that the solutions could be operationalized with a computer" (Fraillon et al. 2019, p. 27).

Both CIL and CT are described in separate sections of the assessment framework, built on previous conceptualizations, and developed in close cooperation with experts and representatives of participating countries. ICILS also studied the use of digital technologies as a means of continuing earlier IEA research from the 1990s and early 2000s, such as that of the Second International Technology in Education Study (SITES; see IEA 2020).

These two examples of learning areas that are transversal and cross-curricular illustrate that a consistent curricular-referenced approach to the development of instruments to measure outcomes is often difficult to implement. Rather, for both ICCS and ICILS the approach is to develop definitions of constructs for measurement that are regarded as relevant to the diversity of approaches across national curricula. In this respect, the development of assessment frameworks that clearly describe the scope of measurement is of particular importance in terms of the validity of the study results.

3.6.3 Frameworks for Measuring Real-Life Skills

OECD studies of educational outcomes and learning tend to define ranges of previously defined knowledge skills that are viewed as important for citizens instead of referencing these skills to existing curricula. OECD's PISA, which has been conducted every three years since 2000, is designed to assess the extent to which 15-year-old students near the end of their compulsory education "have acquired the knowledge and skills that are essential for full participation in modern societies" (OECD 2019, p. 11). While the study routinely assesses reading, mathematics, and science literacy in each cycle, in particular cycles it also has assessed additional domains such as problem solving, digital reading, or financial literacy. In each cycle, one of the core domains is assessed with more extensive coverage as the major domain, while the two others are measured with less item material.

In its approach to the assessment of educational achievement, PISA emphasizes its policy-orientation that links outcomes to characteristics of students, schools, and education systems, its concept of "literacy" referring to the capacity of 15-year-old students to apply knowledge to different real-world situations, its relevance to life-long learning concepts, its regular assessments and breadth of coverage. While each

domain is defined and described in separate frameworks, the contextual aspects for measurement in each PISA cycle are outlined in a questionnaire framework that focuses on variables that are particularly relevant for the major domain (such as reading literacy in PISA 2018).

The assessment of adults' knowledge and skills in OECD's PIAAC study is administered every ten years and aims at the measurement of key cognitive and workplace skills that are regarded as necessary for participation in society and prosperity at the system level (OECD 2012). As with PISA, the framework defines the competences (without reference to national curricula) and measures the extent to which adults demonstrate them. In its last cycle in 2011/2012, PIAAC set out to measure adult skills in literacy, numeracy and problem solving in technology-rich environments across 25 countries. Again, its assessment framework provides an important reference point for understanding the study's results, as it illustrates the scope of skills that are assessed in terms of what are considered relevant adult skills for participation in modern societies.

While curriculum-referenced studies can be judged in terms of their coverage of what education systems have defined as aims for learning, the OECD approach to assessing achievement sets out overarching (international) learning goals that are defined by looking at what "should" be expected from an "output" perspective. In this respect, the frameworks for PISA and PIAAC were both shaped by the OECD Definition and Selection of Competencies (DeSeCo) project (Rychen and Salganik 2003). Both PISA and PIAAC explicitly define ranges of knowledge and skills that are deemed essential for young people and adults instead of referencing existing curricula. The theoretical bases for their respective frameworks need to be more extensively elaborated than is the case for curricula-referenced frameworks and are of key importance for an understanding of their results.

3.6.4 Frameworks for Measuring Contexts

The OECD Teaching and Learning International Survey (TALIS) and IEA's SITES Module 2 (SITES-M2) are studies of contexts that have not included assessment data.

TALIS is an ongoing large-scale survey of teachers, school leaders, and their learning environments. TALIS was first administered in 2008, and then again in 2013 and 2018. It has been administered in lower secondary schools (ISCED level 2 according to the international ISCED classification, see UNESCO Institute of Statistics 2012) with options to be administered in primary (ISCED level 1) and upper secondary (ISCED level 3) schools and a further option for the survey to be administered in PISA sampled schools. The TALIS 2018 framework built on the cycles in 2008 and 2013. It was developed with advice from a questionnaire expert group (through a series of virtual and personal meetings), national project managers and the OECD secretariat (Ainley and Carstens 2018). The development focus was on effective instructional and institutional conditions that enhance student learning and how these vary within and across countries and over time.

The TALIS 2018 framework addressed enduring themes and priorities related to professional characteristics and pedagogical practices at the institutional and individual levels: teachers' educational backgrounds and initial preparation; their professional development, instructional and professional practices; self-efficacy and job satisfaction; and issues of school leadership, feedback systems, and school climate. It also addressed emerging interests related to innovation and teaching in diverse environments and settings.

SITES-M2 was a qualitative study of innovative pedagogical practices using ICT (Kozma 2003). It aimed to identify and describe pedagogical innovations that were considered valuable by each country and identify factors contributing to the successful use of innovative technology-based pedagogical practices. The framework specified procedures for identifying innovative practices in teaching classrooms in primary, lower-secondary, and upper-secondary schools and methods to be used to collect and analyze data. National research teams in each of the participating countries applied these common case study methods to collect and analyze data on the pedagogical practices of teachers and learners, the role of ICT in these practices, and the contextual factors supporting and influencing them. Data were collected from multiple sources for each case, including questionnaires for school principals and technology coordinators, individual or group interviews, classroom observations, and supporting materials (such as teacher lesson plans).

3.7 Conclusions

Most research studies are based on frameworks that link them to the extant literature in the field, define research questions, and articulate methods. However, frameworks are especially important for ILSA studies because of the need to ensure validity across a great diversity of national contexts and education systems. Frameworks for ILSA studies need to be explicit about the constructs being measured and the ways in which they are measured. An assessment framework should be the main point of reference to understand how common elements of learning are defined and measured, and how comparable data across countries are to be generated and analyzed.

Consistency of definition and measurement is already a challenge, even for achievement in fields such as mathematics and science, but it is more of a challenge in learning areas that are context-dependent, such as reading, and/or of a rather transversal, cross-curricular nature and not consistently articulated in curriculum documents, such as civic and citizenship education and ICT-related skills. The importance of assessment frameworks for providing reference points in terms of construct validity is also highlighted in cases where ILSAs need to document content- and method-related changes in terms of the definition of what and how learning outcomes are measured (such as with the transition to computer-based delivery of assessments or the adaptation of new content areas as a consequence of societal developments that affect the respective learning area).

References

Ainley, J., & Carstens, R. (2018). *Teaching and learning international survey (TALIS) 2018 conceptual framework.* OECD Education Working Papers, No. 187. Paris, France: OECD Publishing. http://dx.doi.org/10.1787/799337c2-en.

Ainley, J., Fraillon, J., Schulz, W., & Gebhardt, E. (2016). Conceptualizing and measuring computer and information literacy in cross-national contexts. *Applied Measurement in Education, 29*(4), 291–309. https://doi.org/10.1080/08957347.2016.1209205.

Ainley, J., Schulz, W., & Friedman, T. (Eds.). (2013). *ICCS 2009 Encyclopedia. Approaches to civic and citizenship education around the world.* Amsterdam, The Netherlands: IEA. https://www.iea.nl/publications/encyclopedias/iccs-2009-encyclopedia.

Bocconi, S., Chioccariello, A., Dettori, G., Ferrari, A., Engelhardt, L. Kampylis, P., et al. (2016). Developing computational thinking: Approaches and orientations in K-12 education. In G. Veletsianos (Ed.), *EdMedia. World Conference on Educational Media and Technology 2016, June 28–30, 2016, Vancouver, British Columbia* (pp. 13–18). Waynesville, NC: Association for the Advancement of Computing in Education (AACE). https://www.learntechlib.org/j/EDMEDIA/v/2016/n/1.

European Commission/EACEA/Eurydice. (2017). *Citizenship education at school in Europe—2017*. Eurydice Report. Luxembourg: Publications Office of the European Union. https://eacea.ec.europa.eu/national-policies/eurydice/content/citizenship-education-school-europe-%E2%80%93–2017_en.

Fraillon, J., Ainley, J., Schulz, W., Duckworth, D., & Friedman, T. (2019). *IEA international computer and information literacy study 2018 assessment framework*. Cham, Switzerland: Springer. https://www.iea.nl/publications/assessment-framework/icils-2018-assessment-framework.

Fraillon, J., Schulz, W., & Ainley, J. (2013). *International computer and information literacy study 2013: Assessment framework.* Amsterdam, the Netherlands: IEA. https://www.iea.nl/publications/assessment-framework/international-computer-and-information-literacy-study-2013.

IEA. (2020). *IEA studies [webpage]*. Amsterdam, the Netherlands: IEA. https://www.iea.nl/index.php/studies/ieastudies.

Jago, C. (2009). *A history of NAEP assessment frameworks*. Paper commissioned for the 20th anniversary of the National Assessment Governing Board 1988–2008. Washington, DC: National Assessment Governing Board, US Department of Education. www.nagb.gov/focus-areas/reports/history-naep-assessment-frameworks.html.

Kozma, R. (Ed.). (2003). *Technology, innovation, and educational change, a global perspective. A report of the Second Information Technology in Education Study Module 2*. Eugene, OR: International Society for Technology in Education (ISTE).

Lietz, P. (2017). Design, development and implementation of contextual questionnaires in large-scale assessments. In P. Lietz, J. Cresswell, K. Rust, & R. Adams (Eds.), *Implementation of large-scale education assessments* (pp. 92–136). Chichester, UK: Wiley.

Mendelovits, J. (2017). Test development. In P. Lietz, J. Cresswell, K. Rust, & R. Adams (Eds.), *Implementation of large-scale education assessments* (pp. 63–91). Chichester, UK: Wiley.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2015). *PIRLS 2016 assessment framework* (2nd ed.). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://www.iea.nl/index.php/publications/assessment-framework/pirls-2016-assessment-framework.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://www.iea.nl/index.php/publications/assessment-framework/timss-2019-assessment-frameworks.

Mullis, I., Martin, M., Kennedy, A., Trong, K., & Sainsbury, M. (2009). *PIRLS 2011 assessment framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://www.iea.nl/index.php/publications/assessment-framework/pirls-2011-assessment-framework.

NAGB. (2019). *Mathematics framework for the 2019 National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board, U.S. Department of Education. www.nagb.gov/content/nagb/assets/documents/publications/frameworks/mathematics/2019-math-framework.pdf.

OECD. (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD survey of adult skills*. Paris, France: OECD Publishing. https://doi.org/10.1787/9789264128859-en.

OECD. (2019). *PISA 2018 assessment and analytical framework*. Paris, France: OECD Publishing. https://doi.org/10.1787/b25efab8-en.

Pearce, J., Edwards, J., Fraillon, J., Coates, H., Canny, B., & Wilkinson, D. (2015). The rationale for and use of assessment frameworks: improving assessment and reporting quality in medical education. *Perspectives on Medical Education, 4,* 110–118. https://link.springer.com/article/10.1007/s40037-015-0182-z.

Rychen, D. S., & Salganik, L. H. (Eds.). (2003). *Key competencies for a successful life and a well-functioning society*. Göttingen, the Netherlands: Hogrefe and Huber.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford, UK: Pergamon.

Schulz, W., Ainley, J., Fraillon, J., Losito, B., & Agrusti, G. (2016). *IEA international civic and citizenship education study 2016 assessment framework*. Cham, Switzerland: Springer. https://www.iea.nl/publications/assessment-framework/iea-international-civic-and-citizenship-education-study-2016.

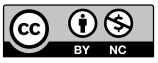
Schulz, W., Fraillon, J., & Ainley, J. (2013). Measuring young people's understanding of civics and citizenship in a cross-national study. *Educational Psychology: An International Journal of Experimental Educational Psychology, 33*(3), 327–349.

Schulz, W., Fraillon, J., Ainley, J., Losito, B., & Kerr, D. (2008). *International civic and citizenship education study 2009: Assessment framework*. Amsterdam, The Netherlands: IEA. https://www.iea.nl/publications/assessment-framework/international-civic-and-citizenship-education-study-2009.

Torney, J., Oppenheim, A. N., & Farnen, R. F. (1975). *Civic education in ten countries: An empirical study*. New York, NY: Wiley.

Torney-Purta, J., Lehmann, R., Oswald, H., & Schulz, W. (2001). *Citizenship and education in twenty-eight countries*. Amsterdam, the Netherlands: IEA. https://www.iea.nl/publications/study-reports/international-reports-iea-studies/citizenship-and-education-twenty-eight.

UNESCO Institute for Statistics. (2012). *International standard classification of education. ISCED 2011*. Montreal, Canada: UNESCO Institute for Statistics. http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf.



Chapter 4
Assessment Content Development

Liz Twist and Julian Fraillon

Abstract The very nature of international large-scale assessments (ILSAs) raises unique challenges for test developers. The surveys are used across many different countries and translated into multiple languages; they must serve as sound measurement instruments across a very wide range of attainment whilst also engaging students from diverse backgrounds and with very different educational experiences. In order for the surveys to meet the stated aims to evaluate, understand, and improve education worldwide, policymakers must have confidence that the conclusions drawn from the results are relevant and meaningful in their particular jurisdictions. The International Association for the Evaluation of Educational Achievement (IEA) has a comprehensive assessment development process; assessment experts and representatives of participating countries review the material at key points during the test development and throughout strict standards for quality assurance are applied. Care is taken to ensure that consideration is given to diverse perspectives whilst maintaining the integrity and rigor of the assessment instruments as specified in the assessment framework. This established and thorough approach has enabled the assessments to evolve whilst continuing to meet the diverse needs of stakeholders.

Keywords Assessment development · Comparative studies · International Civic and Citizenship Education Study (ICCS) · International Computer and Information Literacy Study (ICILS) · International large-scale assessments (ILSAs) · Progress in International Reading Literacy Study (PIRLS) · Trends in International Mathematics and Science Study (TIMSS)

L. Twist (✉)
National Foundation for Educational Research (NFER), The Mere, Upton Park, Slough, Berkshire SL1 2DQ, UK
e-mail: l.twist@nfer.ac.uk

J. Fraillon
Australian Council for Educational Research (ACER), 19 Prospect Hill Rd, Camberwell, VIC 3124, Australia
e-mail: Julian.Fraillon@acer.org

H. Wagemaker (ed.), *Reliability and Validity of International Large-Scale Assessment*, IEA Research for Education 10, https://doi.org/10.1007/978-3-030-53081-5_4

4.1 Introduction

IEA is faced with considerable challenges when developing appropriate questions for international large-scale assessments (ILSAs) that are to be used across a broad range of countries, languages, and cultures. For the assessments to contribute to meeting IEA's stated aim, to evaluate, understand, and improve education worldwide, the assessments must have robust technical functioning and yet retain credibility with stakeholders. Whilst there exist some internationally recognized technical standards relating to the development of educational assessments (see Sect. 4.5), even within countries, there is unlikely to be unanimous agreement about what constitutes an appropriate and high quality assessment instrument. This diversity of views is magnified when considering the number and variety of countries that participate in IEA surveys and the challenges associated with putting into practice the principles and standards relating to educational assessment.

In this chapter, we aim to show how the technical quality and strength of IEA assessments is a result of deliberate strategies to maximize the benefits of the diverse perspectives of IEA's researchers, stakeholders, and expert consultants, and how a collaborative and consultative approach leads to the development of high quality measurement instruments. We discuss the process of assessment item development for IEA surveys, looking at the range of item types and the role of participating countries in item development. We consider the approaches adopted to ensure quality in item development.

4.2 Key Features of ILSAs that Influence Assessment Content Development

ILSAs are, by definition, not aligned to any specific country's curriculum or framework. Further to this, the specificity and content of curricula, standards, or frameworks vary greatly across countries and according to learning areas being assessed. In the IEA context,[1] reading, assessed in the Progress in International Reading Literacy Study (PIRLS), and mathematics and science, assessed in the Trends in International Mathematics and Science Study (TIMSS), are core learning areas with consequently strong and explicit representation in country curricula and, where applicable, local or national assessments within countries. In contrast, there is far greater variation across countries in the explicitness and emphasis given in the curriculum to civics and citizenship education, computer and information literacy (CIL), and computational thinking (CT), skills that are measured in the International Civic and Citizenship Education Study (ICCS) and the International Computer and

[1]For more than 60 years, IEA has been a leading figure in the field of comparative studies of education. Their activities cover all aspects of educational research. Current international IEA studies include TIMSS, PIRLS, ICCS, and ICILS (see IEA 2020).

Information Literacy Study (ICILS). While in all ILSAs it is essential for the assessment content to be drawn from a broad interpretation of the construct defined in the assessment framework (see Chap. 3 for further details of assessment framework development), the different approaches and level of curriculum detail across countries introduce a unique set of challenges to the development of test content in ILSAs.

In order to maximize curricula/domain coverage, a matrix survey design is typically used, in which individual students complete a sub-sample of items, the entirety of which assess the defined domain. This requires the development of an extensive pool of items, each of which is suitable for use in each of the participating countries. This pool needs to include items with a very wide range of difficulty in order to provide all participating countries with sufficient precision in the outcomes of the surveys to meet their objectives. This is particularly challenging when there is a wide range of student achievement across countries.

There is a significant challenge of developing a pool of items that is suitable for use across a range of countries, and agreed by country representatives to represent the learning area as it is understood and assessed within each country. While the test content is developed with reference to a common assessment framework (rather than to any given country's curriculum), expert judgements of the suitability of each item to assess the specified content in each country rightly take into account existing relevant assessments in the same or similar learning areas within countries. In learning areas such as reading, mathematics, and science that are assessed in PIRLS and TIMSS, many countries have well-established pools of existing items that are used in national or local assessments which can provide a frame of reference. ILSA assessment content, while governed by the specifications of the assessment framework, must also be recognizably relevant to the assessment of learning areas as they are understood and represented in national assessment contexts. Evaluating the coherence between ILSA assessment content and national assessment contexts can be more difficult in studies such as ICCS and ICILS. In such studies, while some participating countries may have explicit curricula and standards that, together with contributions from relevant academic literature and expert judgements, can contribute to the content of the assessment framework, many countries may not have existing pools of assessment items that national experts can refer to when evaluating the suitability of the ILSA test items for use in their national contexts. In these cases, expert judgements of the appropriateness of the assessment content may need to be based on more abstract conceptualizations of what is likely to be relevant and suitable rather than in comparison with what is known already to work within countries.

In addition to the objective of measuring and reporting student achievement at a given point in time, a key reason that many countries choose to participate in ILSAs is to monitor achievement over time. ILSAs have varying cycles of data collection. In IEA studies, the PIRLS cycle is five years, TIMSS is four years, ICCS is seven years, and ICILS is five years. This requirement for longevity is discussed further in Chap. 2, but it does place an additional demand on test item development. That is, the item pool needs to include items that are likely to be suitable for use in future cycle(s), as well as in the cycle in which they are developed. Items that are used in more than two cycles in one of the listed ILSAs may therefore need to be appropriate over a

period spanning more than 14 years. In all learning areas this can pose challenges for item development. In IEA studies, this clearly poses challenges for ICILS when working in the domain of rapidly evolving technologies, but similar challenges are emerging as all studies transition to computer-based delivery.

4.3 Validity in International Large-Scale Assessments

The assessment review processes described in this chapter are operational manifestations of the aim to maximize the validity of the assessments. In this context, validating the assessment requires an evaluation of the evidence used to support particular interpretations of the survey results. While this includes an evaluation of the assessment content that is the focus of this chapter, the frame of reference for the review of the validity of the assessment is broader than the contents of the assessment itself (Kane 2013).

The conceptualization of validity proposed by Kane (2013) requires that validity focuses on the use of the assessment outcomes. Also relevant is the work of Oliveri et al. (2018, p. 1), who proposed a conceptual framework to assist participating countries to:

- Systematically consider their educational goals and the degree to which ILSA participation can reasonably help countries monitor progress toward them;
- Use an argument model to analyze claims by ILSA programs against the background of a country's specific context; and
- More clearly understand intended and unintended consequences of ILSA participation.

Others, such as Stobart (2009), have recognized the complexity of producing a validity argument when assessments may be used for multiple and varied purposes. His reservations concerned the validity in the use of one country's national assessments to which many purposes had become attached; the demand is increasingly complex when it concerns the use of an assessment in dozens of countries.

In this chapter, we elaborate on the process of item development in IEA surveys. This is the foundation for the two key sources of validity evidence: expert review, and item and test analysis of the survey data. Other sources of validity evidence may be available within countries; for example the association between performance in the surveys and in other national assessments, but inevitably this is local evidence. While this chapter focuses on how the assessment instrument development process is used to evaluate the validity of the instrument, this notion of validity works in the larger framework (suggested by Kane 2013) in which the evaluation relates to the suitability of the instruments to elicit data that can be used to support defensible interpretations relating to student outcomes in the areas of learning being researched.

4.4 The Assessment Frameworks

As explored in Chap. 3, the assessment frameworks define the construct/s to be assessed and the nature of the assessment to be developed. These documents, publicly available and rooted in the research theory and evidence, are reviewed and revised by expert groups in the early stages of each cycle. They are used to guide the content development but also have the potential to support participation decisions and appropriate interpretation of the outcomes.

In their definitions of the constructs to be assessed, the assessment frameworks inevitably shape the assessment design. In the case of reading, for example, the PIRLS framework defines reading literacy as follows:

> Reading literacy is the ability to understand and use those written language forms required by society and/or valued by the individual. Readers can construct meaning from texts in a variety of forms. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment (Mullis and Martin 2019).

This focus on reading as a meaning-making process requires the assessment to ensure that participating students engage with and respond to the written texts. This response takes a written form. There is no element of the PIRLS assessment that specifically assesses decoding, namely students' ability to convert graphemic (or logographic) forms into sounds. Whilst decoding is implicit in all reading, the starting point for the PIRLS assessment materials is the individual and, in most cases, silent reading of written texts ("passages") and the assessment is of students' ability to comprehend them by answering, in writing, the written questions.

In ICILS, CIL is defined as:

> …an individual's ability to use computers to investigate, create, and communicate in order to participate effectively at home, at school, in the workplace and in society (Fraillon et al. 2019, p. 18).

and CT is defined as:

> …an individual's ability to recognize aspects of real-world problems which are appropriate for computational formulation and to evaluate and develop algorithmic solutions to those problems so that the solutions could be operationalized with a computer (Fraillon et al. 2019, p. 27).

In each of the ICILS constructs, there is a clear emphasis on the use of computers as problem-solving tools. For CIL there is an emphasis on information gathering and communication, whereas for CT the emphasis is on conceptualizing and operationalizing computer-based solutions to problems. Both definitions suggest the use of computer delivered instruments and an emphasis on achievement being measured and demonstrated in real-world contexts. In response to these demands, the ICILS CIL and CT instruments consist of modules comprising sequences of tasks linked by a common real-world narrative theme per module (see Fraillon et al. 2020).

It is particularly important that a clear exposition of the knowledge, skills, and understanding being assessed in the international surveys is provided. While this

begins with the assessment framework, any review of the assessment instruments includes consideration of the degree to which these instruments address the assessment outcomes articulated by the framework. As part of the development process, each assessment item is mapped to the relevant framework and the accuracy and defensibility of these mappings is one aspect of the validity review. While it is not possible to collect other validity evidence, such as the relationship between performance on the survey and performance on another assessment in broadly the same domain (concurrent validity) on an international level, it is possible that some countries could potentially undertake such an exercise.

4.5 Stimulus Material and Item Development: Quality Criteria Associated with Validity

There are well-established criteria that all assessment material should be evaluated against. The *Standards for Educational and Psychological Testing* (American Educational Research Association et al. 2014), for example, use the concepts of validity, reliability, and fairness as organizing perspectives from which to evaluate the quality of assessment material. For the purpose of this chapter, we will address eight groups of evaluation criteria that are routinely implemented in the development of assessment materials in IEA studies:

- representation of the construct
- technical quality
- level of challenge
- absence of bias
- language and accessibility
- cultural and religious contexts
- engagement of test-takers
- scoring reliability.

These quality criteria are applied during all phases of the materials development process.

4.5.1 Representation of the Construct

As described in the previous section, the assessment constructs in ILSAs are defined and explicated in detail in an assessment framework (see Chap. 3 for further details). In the context of ILSA, the assessment constructs do not represent any given national curriculum but are designed to be relevant to and recognizable within national curricula. Once an ILSA assessment construct has been accepted by countries, it is essential that the assessment instrument provides a true representation of the

construct. This evaluation, conducted by assessment developers, country representatives, and other experts takes place throughout the materials development process. When considering this criterion, it is essential that reviewers consider the construct being used in the study without conflating it with what might be used in local (national) contexts. For example, in PIRLS, poetry is not included in the assessment because of the specific challenges associated with translation. In many curricula, the reading and comprehension of poetry is mandated. Despite this omission, what is included in PIRLS is a wide representation of reading comprehension and is accepted as such by stakeholders.

All assessments in IEA studies are also carefully mapped to their constructs. The assessment frameworks typically specify the proportions of items addressing different aspects of the constructs to ensure a full and appropriate reflection of the constructs in the instruments. As part of the review process, both the accuracy of the mapping of items to the constructs and the degree to which the total instrument meets the design specifications in the framework are considered.

4.5.2 Technical Quality

While the technical quality of test materials could be considered a property of the representation of the construct in the items, it warrants independent explication, as it is central to the materials review. Questions that need to be considered in the technical review of the materials can include, but are not limited to:

- Is the material clear, coherent, and unambiguous?
- Is the material self-contained? Or does it assume other prior knowledge, and, if so, is this appropriate?
- Are there any "tricks" in materials that should be removed?
- Is each key (the correct answer to a multiple choice question) indisputably correct?
- Are the distractors (the incorrect options to a multiple choice question) plausible but indisputably incorrect?
- Do the questions relate to essential aspects of the construct or do they focus on trivial side issues?
- Is the proposed item format the most suitable for the content in each case?
- Are there different approaches to arriving at the same answer? If so, do these different approaches represent equivalent or different levels of student ability and should this be reflected in the scoring?
- Is there any local dependence across items within a unit (testlet) or across the instrument? Local dependence occurs when either the content of or the process of answering one question affects the likelihood of success on another item.

4.5.3 Level of Challenge

For the assessments to function well in psychometric terms across the participating countries, they must adequately measure the skills of the highest and lowest attainers. Within all countries there is a range of attainment in the assessed domains; the difference between countries is often in the proportions of students at different points across the range. Of course, there are some notably high achieving countries: in TIMSS 2015 at grade 4, for example, 50% of students in Singapore reached the advanced benchmark. In contrast, in 12 countries, fewer than three percent of students reached the advanced benchmark and, in four of these countries, fewer than 50% of students reached the low benchmark (Mullis et al. 2016). As a measurement exercise, students need to complete some items that they find straightforward and some that are challenging: if a student scores maximum points the measure is not providing full information about their capabilities; similarly a failure to score any points is not informative about what skills have been developed. When selecting content for the main survey data collection, careful consideration is given to the relative proportion and allocation of items of varying difficulty across the assessment. More recently, some studies have included plans for an approach that allows for the balance of item difficulties to vary across countries to better match known profiles of achievement within countries.

4.5.4 Absence of Bias

Whilst this aspect of the quality criteria is part of the psychometric analysis, it is also a consideration during the development process. Bias occurs when factors other than those identified as integral to the assessment impact on the scores obtained by students, meaning that students with the same underlying ability do not achieve equivalent scores. In technical terms, this is construct irrelevant variance. There are a number of potential sources of bias that test developers are mindful of. The benefit of prior experience can be evident in performance on a reading test, for example, where, independent of reading ability, the assessment rewards some test-takers for existing knowledge. A reading assessment would not directly include content from a reading program widely used in a participating country but it would consider for inclusion material published in particular countries as this provides a necessary level of authenticity. The review for bias is one in which the perspectives of country representatives is crucial, as they may identify particular content, themes, or topics that may unfairly advantage or disadvantage test-takers in their national context. The psychometric measures of bias that inform assessment development rely on there being sufficient numbers of test-takers in the sub-groups for comparison. Routinely in ILSA, bias across countries is measured at the item level in what is referred to as “item-by-country interaction” and bias within countries (and cross-nationally) is measured between female and male test-takers.

4.5.5 Language and Accessibility

The assessment content is developed and reviewed in English. In all assessment development there is a need for consideration of the precision in language alongside the amount of reading required. In the case of international surveys, a further consideration is the impact of translation, as discussed in Chap. 6. During the development phase the onus is on country representatives to be alert to any particular concerns about the feasibility of translating particular words and phrases. It is frequently the case that, in discussion, often within multilingual groups, alternative words and phrases are identified that function equally well within the assessment.

Reading ability should not influence test performance when reading is not the construct being assessed. For this reason, accommodations such as readers may be used in the administration of an assessment of science, for example. In studies such as TIMSS, ICCS, and ICILS, where reading is not the domain being assessed, there is a clear and deliberate effort to keep the reading load to a minimum. A rule-of-thumb is that the typical reading load in a non-reading assessment should be equivalent to the level attained by students that are roughly two grades below the grade level of the students being tested. The reading load is primarily influenced by sentence length, sentence structure, and vocabulary use. In some cases, it is feasible to use readability indexes to support the evaluation of the reading load of materials, however, interpreting the output of a readability index must be done with careful consideration of the domain being assessed. For example, in ICCS, terms such as *democracy* or *sustainable development* may represent essential content of the domain, but also inflate the reading load of materials when measured using a readability index. In addition, as the original materials are developed in English, readability index outcomes applied to English language text cannot be assumed to be appropriate when considering how the text will appear under translation to languages other than English.

4.5.6 Cultural and Religious Contexts

Developers of all assessments used within a single country need to be alert to potential cultural or religious issues that may impact on how the assessment is interpreted. It is an even greater focus when the assessments are deployed internationally. The issues are different according to the domains being assessed. For example, certain concepts are accepted as legitimately part of a science assessment and, in fact, required in order to ensure as comprehensive an assessment as possible, but may not be as readily accepted in a reading assessment. Similarly, some texts, such as traditional fables, may focus on explanations of natural phenomena that would have no place in a science assessment and may also challenge some beliefs, yet they are a part of what is accepted as the broad literary canon that may legitimately be included in a reading assessment.

This criterion includes consideration not just of the items but also the images and contexts incorporated into some assessments. When selecting contexts for ICILS content there are, for example, varying rules and laws across countries relating to grade 8 students' access to and engagement with social media platforms. The involvement of representatives of participating countries in the development process ensures that many perspectives are considered at an early stage.

4.5.7 Engagement of Test-Takers

There is ample evidence that more engaged students perform better. This is evident in better learning in the classroom (e.g., Saeed and Zyngier 2012) or in test-taking (e.g., Penk et al. 2014). While test developers make no attempt to entertain students, they do aspire to present engaging material to the young people completing the assessment, namely the sort of content that Ryan and Deci (2000) described as being "intrinsically interesting" for the majority. This may be in the contexts selected for scenarios for tasks, or in the texts selected. In PIRLS, for example, students are asked to indicate how much they enjoyed reading specific texts at the field trial stage. This is one of the sources of evidence that is considered when selecting the final content.

In addition, the assessment should have a degree of coherence for the student, even though each student completes only a part of the overall assessment. Each student will be exposed to items that assess more than solely number operations; for example, in TIMSS and in ICCS, each test booklet includes items that assess content associated with all four content domains in approximately the same proportions as those specified in the assessment framework for the test instrument as a whole.

4.5.8 Scoring Reliability

At the heart of reliable scoring is consistency. The scoring guides must be interpreted in the same way, ensuring that the same responses achieve the same score, regardless of who the scorer is. Scoring reliability is measured across countries in each survey and within countries by looking at consistency between cycles (part of the trend measure). To establish this consistency, those who undertake the scoring require facility in both the language(s) of the test in their country and also in English. In IEA studies, the international scorer training (i.e., the training of the people responsible for scoring and scorer training within each participating country) is conducted in English, and the scoring guides and scoring resources are presented in English. Some countries choose to translate the scoring materials to the language of the national test, and to run their national scoring in their language of testing; in PIRLS 2016, over half participating countries supplemented their scoring materials with example responses produced within their particular country (Johansone 2017).

Good assessment development practice requires the scoring guides to be developed alongside the items; developers need to document the answers that they are expecting to receive credit and to use these to confirm the process being assessed. Scoring guides and the accompanying materials are developed by an expert group using example responses collected during small-scale trialing. These are subject to ongoing review and are reviewed in the light of the field trial data.

4.6 Stimulus and Item Material: An Overview

4.6.1 Stimulus Characteristics, Selection, and Development

Stimulus materials are the essential core of many assessments, and the format, type, length, and content of stimuli can vary depending on the role they play. In the case of PIRLS, the stimuli are the reading passages that contain the text, images, and information that students read in order to respond to the items. In TIMSS, stimulus materials are the combination of text, images, and data that both contextualize items and provide information needed in order to respond to items. In ICCS, stimulus materials perform a similar role to those in TIMSS (except in the context of civic and citizenship education rather than mathematics and science). In ICILS, the stimulus materials provide both the real-world context across the narrative theme of modules and provide information that may be required to support completion of the tasks.

With the increasing use of computer-based assessment, some stimuli are now being developed exclusively for use on computer. These include all materials for ICILS, the reading passages developed for the computer-based version of PIRLS (ePIRLS), the problem solving and inquiry tasks (PSIs) in TIMSS, and the computer-enhanced modules in development for ICCS. In each case, these stimulus materials need to include interactive functionality that extends beyond what can be achieved in paper-based stimuli. While the nature of this functionality varies according to the assessment constructs being measured (as specified in the assessment frameworks) there are common criteria used to evaluate the viability of computer-based stimulus materials. They need to employ features that are accessible to students and that reflect current conventions of interface design. Furthermore the stimuli need to represent plausible (i.e., not contrived) uses of the technology in context, and operate in a broader test narrative for the students in which the use of and the reason for using the technology are apparent without the need for further explanation. Where computer-based stimuli are used, these considerations are, in addition to those that relate to the content and presentation of stimulus materials, necessary in the selection and evaluation of all stimulus materials regardless of their medium.

The development or selection of passages (texts) or other stimulus materials, such as the creation of a context within which a set of items is presented, can be the most challenging aspect of the development cycle when the relatively abstract descriptions of the assessment in the framework are operationalized. Assessment items flow from

good stimulus materials. When sourcing, selecting, and refining stimulus materials, assessment developers keep in mind the types of item that they can see will flow from the materials.

Stimulus selection and development is challenging when an assessment is being developed for use in a single jurisdiction. In the case of ILSAs, the challenge is greater as the material is scrutinized internationally. First and foremost, it must be clear which aspect of the assessable domain the stimulus is targeted at. In TIMSS, for example, the topic, and content domain (as specified by the assessment framework) can be established using the stimulus context, whereas the cognitive domain assessed will more typically be instantiated through the item. In ICCS, while some stimuli are clearly associated with a given content domain, there are also stimuli that are used to introduce civic and citizenship scenarios that elicit items across a range of content and cognitive domains.

In PIRLS, stimulus material is classified as having a literary or informational purpose. The requirement is for texts which are "rich" enough to withstand the sort of scrutiny that comes with reading items. "Rich" texts are those which are both well-written and also engaging for young students. In all assessment contexts, the selection of reading passages includes consideration of a broad suite of criteria. These include, for example, the degree to which the content of the material is: appropriate for the target grade level, inclusive, culturally sensitive, and unlikely to cause distress to readers. While this poses a challenge in local or national contexts, the challenge is significantly increased when selecting texts that are appropriate to use across a broad range of cultures. While these challenges are common when selecting stimulus materials for any ILSA, they are greatest in reading assessments such as PIRLS where there is a balance between maintaining the integrity of an original self-contained text and evaluating its appropriateness for use across countries. In PIRLS, representatives of participating countries are encouraged to submit texts they feel may be suitable; country engagement at this stage helps to ensure that literary and information texts are characteristic of the material read by children of this age around the world. Other texts are sourced by assessment experts within or associated with the international study center. Texts are submitted in English, although this may be a translation from the original. Information texts are likely to be written specifically for this assessment, and generally draw from a range of sources. In PIRLS 2016, literary texts included contemporary narrative and folk tale. Information texts were diverse and included varied purpose, layout, and structure.

In other studies, country representatives are also invited to submit stimulus materials and even ideas for materials. However, in studies other than reading there is also greater flexibility in adapting stimulus materials to ensure they are suitable for use.

An important element of the development process is to obtain the perspective of participating countries on stimulus materials during a review stage. In PIRLS, this takes place before items are written. This is necessary in PIRLS because each passage is the stimulus for a large number of items and the viability of passages must be considered before engaging in the substantial work of creating the items for each passage. Inevitably, there is considerable diversity in the viewpoints expressed. Material that is regarded as well aligned with the needs and expectations

of one country may be seen by others to be: too challenging, uninteresting, or too unfamiliar; too familiar and commonly used in classrooms; or culturally inaccessible or inappropriate. There is a high level of attrition: material "falls" at this stage and is not developed further. This is not simply a case of identifying the most popular or least problematic material; developers need to be sure that the material to be considered further has the potential to form the basis for a robust assessment. The process of "text mapping" is a means of evaluating whether a prospective text is likely to function well at the item writing stage. In text mapping, the characteristics (e.g., length, genre, form, reading load), core and secondary content and themes (explicit and implicit where relevant) of a text are listed and described. In addition there is some explication of the content and focus of items (with reference to the assessment framework) that the text naturally suggests.

In ICILS, where the real-world context of the test modules is essential to their viability, the stimulus materials are first reviewed from this perspective. Country representatives are asked to consider, in addition to the previously described criteria, the degree to which the proposed scenarios are relevant and plausible for target grade students in their national contexts. Where the stimulus materials are shorter and with more opportunity for revision, it is feasible to review them together with their relevant items.

A feature of all IEA ILSAs is the use of expert groups, separate from and in addition to country representatives, who also review all assessment materials. The expert groups typically comprise people with specialist expertise in assessing the learning area. While the size and composition of the expert groups may vary, in most cases the members are experienced researchers with at least some experience of involvement in IEA studies (as members of national research centers, for example). Many too have experience of working on national assessments in their own countries. The expert reviews can be both electronic (i.e., where feedback is sent electronically to the international study center) and delivered in face-to-face meetings. In IEA PIRLS, where the reading texts are integral to the assessment and are large and challenging to develop, the expert group can be involved in the editing and development of stimulus materials. While it varies across studies, it is typical for the expert group to provide input into the development of the stimulus materials during the early (pre-field trial) phases of development.

4.6.2 Item Characteristics and Development

The process of item development begins once the stimulus materials have been selected and revised (although this can be an iterative process in which stimulus materials are further revised as a part of the item development process).

Items fall broadly into two categories: closed response, where the student is making some sort of selection from a given set of answers, or constructed response, where the student is producing their own response. While traditionally responses in ILSAs have been largely restricted to small amounts of text (from a word or

number through to several sentences or a worked solution to a problem), the transition to computer-based testing has brought with it an expanded set of response formats. In ICILS, for example, students create information resources such as online presentations or websites and create computer coding solutions to problems. For all constructed response items, the scoring guides are developed together with the items as the two are inextricably connected. The proportion of item types is specified in the assessment framework for each study (see Sect. 4.6.3).

4.6.3 Item Types

Item type is generally defined by the response the student must give. Among the most recognizable "closed" item types is multiple-choice, when the student must select the correct option (the "key") from a set of alternatives (the "distractors"). It is generally accepted that there should be at least three distractors and these should be plausible but definitively wrong. The key should not stand out in any way from the distractors (such as by being much longer or shorter than the distractors). The development of good quality multiple-choice items is harder than it may first appear, especially in ensuring that all the distractors are plausible. What makes a distractor plausible in the mind of a student is usually that student has a particular misconception, either in comprehension of the stimulus or related to their understanding of the assessment domain that leads them to the incorrect response. The nature of these misconceptions may vary according to the nature and contents of the stimulus material and the nature of the domain. For example, in PIRLS, distractors may represent an incorrect reading of the text, the imposition of assumptions based on students' typical life experience that are not represented in the text, or the retrieval of inappropriate information from a text. In TIMSS, ICCS, and ICILS, distractors may more commonly represent misconceptions relating to the learning area. Some of these may reflect misconceptions that are well-documented in the research literature and others may represent misconceptions or process errors that are plausible based on the content of an item and the knowledge, understanding, and skills required to reach the solution. Considerable effort is undertaken to create plausible distractors. This involves test developers responding to the item from the perspective of the students, including considering the types of misconceptions that a student may have when responding. However, it is also possible to create distractors for which the distinction between the distractor and the correct response is too subtle to be discernible by students. In these cases, even though the distractor is irrefutably incorrect from an expert perspective, the capacity to discern this is beyond the reach of the students, and consequently many high achieving students believe it to be a correct answer. The empirical analysis of multiple-choice questions following the field trial allows for a review of the degree to which the distractors have been more plausible for lower achieving students and less plausible for higher achieving students.

Other closed item types, where the student is not developing their own response but indicating a selection in some way, include sequencing, where a series of statements describing or referring to a sequence are put in order, and other forms of "sorting" or matching where students indicate which pieces of information can be best matched together. Computer-based test delivery in PIRLS, TIMSS, and ICCS allows for the use of a greater range of closed item formats than can be easily developed on paper. In particular, many forms of "drag and drop" items can allow students to respond by sequencing, sorting, counting and manipulating elements. ICILS also includes a suite of closed format computer skills tasks in which students are required to execute actions in simulated software applications. Such tasks are closed, in that the response format is fixed and restricted, but from the perspective of the student the item functions as if students were working in a native "open" software environment (see Fraillon et al. 2019 for a full description of these tasks).

Constructed response items require the student to generate the content of their response. These can vary in length from a single character (such as a number) through to words, equations, and sentences. ICILS includes authoring tasks that require students to "modify and create information products using authentic computer software applications" (Fraillon et al. 2019, p. 49). These tasks can be, for example, the creation of an electronic presentation, or an electronic poster or webpage.

The scoring guide for a constructed response item is, from a content development perspective, an integral component of the item itself. As such, the scoring guide is developed alongside a constructed response item and then refined as evidence is collected. The process that the item is addressing is identified in the scoring guide, along with a statement of the criteria for the award of the point(s). The guide typically includes both a conceptual description of the essential characteristics of responses worthy of different scores as well as examples of student responses that demonstrate these characteristics. There is ongoing refinement, often following piloting, and examples are included of actual student responses. Scoring guides are incorporated into the iterative item review process when there may be further refinement.

In ICILS, the scoring guides for the authoring tasks comprise multiple (typically between 5 and 10) analytic criteria. These criteria address distinct characteristics of the students' products and each has two or three discrete score categories. In most cases, these criteria assess either an aspect of the students' use of the available software features in the environment to support the communicative effect of the product or the quality of the students' use of information within their product. Despite the differences between the analytic criteria used in assessing the ICILS authoring tasks and the scoring guides developed for shorter constructed response items, the approach to the development of both types is the same. It relies on consideration of the construct being assessed as described by the assessment framework and the applicability of the guide to be interpreted and used consistently across scorers. In the case of ILSA, this includes across countries, languages, and cultures.

4.7 Phases in the Assessment Development Process

There is no single common set of activities in the development of ILSA assessments. In each study, what is completed and how is determined by the characteristics, scale, and resources of the study. In spite of this, there is a set of phases that are common to the development of assessment materials in all ILSAs. During each phase, review activities are conducted by national experts, expert groups, and the content developers, applying the quality criteria we described in Sect. 4.5.

4.7.1 Phase 1: Drafting and Sourcing Preliminary Content

This first phase of development is characterized by creative problem solving and breadth of thinking. For studies such as PIRLS or ICILS, where the texts or contexts must be confirmed before detailed item development can begin, this first phase focuses on sourcing and locating texts and conceptualizing and evaluating potential assessment contexts. In studies such as TIMSS and ICCS, in which the stimulus materials can be developed in smaller units (or testlets) with their related items, it is possible to both source and develop stimulus and items in this early phase.

During this phase, contributions from country representatives are actively encouraged and expected. Where face-to-face meetings occur it is common practice to include some form of assessment development workshop with country representatives followed by a period in which submission of texts, stimulus and assessment materials, and ideas are invited. For computer-based assessments, country representatives are typically presented with a demonstration of the testing interface and examples of items that the interface can deliver. They are then invited to propose and submit ideas and storyboards for computer-based items rather than fully developed assessment materials. Any project expert group meetings during this phase will include evaluation and development of content.

4.7.2 Phase 2: Item Development

The item development phase begins when any necessary texts, contexts, and stimuli have been selected. In this phase, the emphasis is on developing the item content and scoring guides that, together with the stimulus material and contexts comprise the draft assessment instrument. In the case of computer-based assessments, this may also include development of new item formats that support assessment that has previously not been possible. For example, in ICILS, the concept for a fully-functional visual coding system as part of the assessment was first proposed and developed in this early phase of item development.

This phase typically includes the opportunity for extensive contribution to and review of the materials by country representatives and the external experts involved in stimulus and text development. The item development phase can include any or all of the following procedures.

- Item development workshops with country representatives can be conducted early in the item development phase. The process of working in small but international teams means that different perspectives are shared and assumptions may be challenged. Whilst the working language is English, at this stage concerns about translation may emerge. Discussion of the issue across countries may ensure a resolution is identified but, in some cases, it will be found that the issue is irreconcilable and the material does not progress further in development. At this point, all items are in a relatively unpolished state, the focus being on identifying the potential for items to be developed rather than on creating the finished product.
- Piloting (and/or cognitive laboratories), in which draft assessment materials are presented to convenience samples of test-takers from the target population (usually students from countries participating in the given study) for the purpose of collecting information on the students' experiences of completing the materials. The nature of piloting can vary across projects. In some cases piloting is conducted using cognitive laboratory procedures during which students complete a subset of materials and then discuss the materials (either individually or in groups) with an administrator. In other cases test booklets are created for larger groups of students to complete in a pilot with a view to undertaking some simple quantitative analyses of item performance or to collect constructed responses from which the scoring guide can be refined and training materials developed. Piloting is of particular value when developing materials using new or changed constructs or to evaluate the test-taker experience of new item formats. This latter use has become particularly relevant in recent years, as many ILSAs transition from paper-based to computer-based formats. Where possible, piloting should be conducted across languages other than the language in which the source materials are developed (US English for all IEA studies) and can provide some early information about issues that may exist in translation of materials and the degree to which materials under translation have maintained their original meaning.
- Desktop review by country representatives and external experts is often conducted as part of the item development process. Where piloting provides information on how test-takers respond to the assessment materials, the desktop review complements this by providing expert feedback on the technical quality of the material (such as the quality of expression, clarity, and coherence and accuracy of the material), the targeting and appropriateness (such as cultural appropriateness) of the material across a broad range of countries, and how well the material represents the content of the assessment framework and the equivalent areas of learning across countries. Typically this review involves providing country representatives and experts with access to the materials (either as electronic files or through a web-based item viewing application) and inviting critical review of the materials (items, stimulus, scoring guides, and contexts). While it is possible to invite an

open review of the materials (in which respondents complete open text responses) it is common practice to structure the review so that respondents provide both some form of evaluative rating of each item and, if appropriate, a comment and recommendations for revision.

- Face-to-face meetings to review materials with country representatives and other experts are essential in the quality assurance process. While these can take place at any time during the item development cycle, they most frequently occur near the end of the process as materials are being finalized for the field trial. At these meetings, all assessment materials are reviewed and discussed, and changes are suggested. Where possible it is common to have external experts review materials in sufficient time before a face-to-face meeting with country representatives to allow for the materials to be refined before a "final" review by country representatives. A feature of IEA studies is the value placed on the input of country representatives to the assessment content. One manifestation of this is that IEA studies routinely include a meeting of national research coordinators as the final face-to-face review of assessment materials before they are approved for use in the field trial.
- Scorer training meetings occur before each of the field trial and main survey. Feedback from the scoring training meetings, in particular for the field trial, lead to refinements of the scoring guides. In most cases, the scoring guides for a study are finalized after the scoring training meeting, taking into account the feedback from the meetings.

4.7.3 Phase 3: The Field Trial and Post Field Trial Review

The field trial fulfils two main purposes:

- to trial the operational capability within participating countries
- to collect evidence of item functioning of newly developed materials.

The field trial is held approximately a year before the main survey. The size of the field trial sample will vary across studies depending on the number of items and test design in each study; however, it is usual to plan for a field trial in which no fewer than 250 students complete each item within each country (or language of testing within each country if feasible). The processes undertaken in preparation for the field trial, during the administration and afterwards in the scoring and data collection phases, mirror what is to be done during the main survey. As well as obtaining item level data, evidence collected from students may include their preferences or whether or not they enjoyed specific parts of the assessment. In PIRLS, for example, students use the universally recognized emoji of a smiley face to indicate their level of enjoyment of the passages. Item functioning is calculated, based on classical test theory and item response theory (IRT) scaling.

Following the field trial and the review of the findings by country representatives and expert groups, the final selection of material to be included in the survey is made. At this stage there are a number of considerations. Test materials must:

- meet the specification described in the assessment framework in terms of numbers of points, tests, item types, and so on;
- maximize the amount of information the test provides about the students' achievement across countries by presenting a large enough range of difficulty in the test items to match the range of student achievement;
- provide a range of item demand across sub-domains or passages, with some more accessible items at the start;
- discriminate adequately (i.e., the performance of high- and low-achieving students should be discernibly different on each item);
- contain new material that is complementary with material from previous surveys that has been brought forward, ensuring adequate and optimum representation of the construct and a balance of content (for example, a balance in male and female protagonists in reading passages);
- using evidence derived from the field trial, show sufficient student engagement and preferences;
- based on evidence from the field trial, show adequate measurement invariance for each item across countries (measured as item-by-country interaction); and
- demonstrate scoring reliability in the field trial.

The field trial also provides the first opportunity in the instrument development process for the scoring guides to be reviewed in the light of a large number of authentic student responses across countries and languages. This review allows the assessment development team to:

- check and when necessary refine the descriptions of student achievement included in the scoring guides in the light of actual student responses;
- refine the scoring guides to accommodate any previously unanticipated valid responses; and
- supplement the scoring guides with example student responses that are indicative of the different substantive categories described in the guides and develop scorer training materials.

In the lead-up to the main survey in each IEA study, national research coordinators meet to assess the data from the field trial and recommendations for the content of the assessment to be used in the main survey. Implementation of the decisions at this meeting can be considered to be the final step in the process of instrument development for that cycle of the assessment.

4.7.4 Post Main Survey Test Curriculum Mapping Analysis

The connection between explicit test and curriculum content within each country and its impact on the suitability of the assessment instrument for reporting national data is of particular importance in the areas of mathematics and science. This is because curricula in these learning areas often are based on sequences of learning content in the development of knowledge, skills, and understanding that are closely associated with specific topic contents. As such, results for an ILSA of mathematics and science achievement may be particularly sensitive to relative differences between curriculum topics in a given country and those in the ILSA instrument used across countries. In IEA TIMSS, this challenge is addressed after the main survey data collection using a test curriculum mapping analysis (TCMA). The TCMA is used to compare a country's performance on items assessing the skills and knowledge that experts in that country's curriculum consider are represented in their intended curriculum with performance on all items included in the assessment. A large discrepancy between the two sets of data (for example, a country having a much higher percentage correct on the items that were represented in their curriculum compared with the percentage correct on all items) would suggest that the selection of items for inclusion in the assessment was affecting the relative performance of countries, something which would be clearly undesirable. The results of the TIMSS 2015 curriculum matching exercise contributed to the validity of the assessment, indicating that there was little difference between each country's relative performance, whether performance across all items was considered or just those considered linked to the country's intended curriculum. Unsurprisingly, most countries perform a little better on items that are considered appropriate to that country.

4.8 Measuring Change Over Time and Releasing Materials for Public Information

While the process of developing ILSA assessment material is centered on a given assessment cycle, it is also conducted with consideration of what has come before and what is planned for the future.

The assessment development plan in an ILSA must take into account any secure material that was used in a previous cycle that will be reused in the current cycle, and also take into account what material from the current cycle may be held secure for future use. It is students' relative performance when responding to this "trend" material across cycles that is the basis for the reporting of changes in student performance over time. As such, it is essential that the content of the trend material fully represents the construct being measured and reported. How this is achieved varies according to the overarching instrument design. For example, in both PIRLS and ICILS, the test items are inextricably linked to their respective texts and modules. For this reason, in these studies, the items linked to a text or module typically span a broad range

of difficulty and cover a large proportion of the constructs. In effect, each PIRLS reading text with its items and each ICILS test module is designed to be as close as is possible to a self-contained representation of the whole assessment instrument. This allows for the selection of trend materials to be made by text or module. However, in PIRLS, where students read for literacy experience and to acquire and use information (Mullis and Martin 2019), it is necessary for both literary and informational texts to be included in the trend materials. In ICILS, a single module may be regarded as a proxy for the full assessment instrument, although it is typical to select more than one module to establish trends. In TIMSS and ICCS, where the items are developed in much smaller testlets, a large number of testlets are selected as trend materials to represent the construct. In all studies, the proportion of trend to new material is high to support robust measurement of changes in student performance across cycles.

An important aspect of communicating with stakeholders is the provision of released material from the assessments. This serves to illustrate how the assessment of the defined domain is operationalized and is indicative of the material seen by students; a useful and practical element, given the diversity of participating countries and the variety of assessment styles. Even though there is not the same measurement imperative for the released materials to represent the construct as there is for trend materials, material selected for release will ideally provide stakeholders with an accurate sense of the nature, level of challenge, and breadth of content covered in the assessment. In many cases, the released material is also made available for other research purposes.

4.9 Conclusions

The process of instrument development in any large-scale assessment is challenging and requires careful planning and expert instrument developers. However, development for ILSA introduces additional challenges. Ultimately the developers' aim is to produce instruments that function effectively in their role as a means of collecting data and assessing performance within many different countries, which present a variety of languages and cultures. While the developers cannot prescribe all possible uses of the assessment outcomes, they can and do ensure the quality of the instruments.

In this chapter we have described the constituent components and processes in the development of ILSA assessment content with a focus on four key IEA studies. While these four studies span six different learning areas that are approached and represented in different ways across countries, the fundamental characteristics and principles of assessment development are common across the studies. With the aim of maximizing the validity of assessments, the development process comprises phases of conceptualization, development, and finalization that are informed by the application of expert judgement and empirical data to interrogate the materials according to a range of quality criteria.

What lies at the core of the pursuit of excellence in this process is the feedback from experts who provide a broad and diverse set of linguistic and cultural perspectives on

the materials. Without these perspectives in the creation of ILSA materials it would not be possible to present materials that can be used confidently and defensibly assert that they can be used to measure student attainment within and across all of the countries that take part in each study.

As the range of countries participating in ILSA continues to increase and the transition to computer-based delivery continues, the ways in which computer-based assessment may improve and expand ILSA will continue to evolve. Computer-based delivery offers the opportunity to include a much broader range of item and stimulus formats than have previously been used on paper, with the opportunity to enhance assessment in existing domains and to broaden the range of domains in which ILSA can be conducted. However, this expanded repertoire of assessment content brings with it additional demands when evaluating the validity of assessments. In addition, the possibility of including process data to better understand and measure achievement is a burgeoning area that requires careful planning and integration into the assessment development process. As ILSA instruments evolve, so too must the evaluation criteria applied in the assessment development process to ensure that the principles underpinning the development of high quality assessment materials continue to be implemented appropriately.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA Publications.

Fraillon, J., Ainley, J., Schulz, W., Duckworth, D., & Friedman, T. (2019). *International computer and information literacy study 2018 assessment framework*. Cham, Switzerland: Springer. https://www.iea.nl/publications/assessment-framework/icils-2018-assessment-framework.

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2020). *Preparing for life in a digital world: International computer and information literacy study 2018 international report*. Cham, Switzerland: Springer. https://www.iea.nl/publications/study-reports/preparing-life-digital-world.

IEA. (2020). Studies [webpage]. Amsterdam, the Netherlands: IEA. https://www.iea.nl/studies.

Johansone, I. (2017). Survey operations procedures in PIRLS 2016. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 6.1–6.26). Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center. https://www.iea.nl/publications/study-reports/international-reports-iea-studies/methods-and-procedures-pirls-2016.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center. https://www.iea.nl/publications/timss-2015-international-results-mathematics.

Mullis, I. V. S., & Martin, M. O. (2019). PIRLS 2021 reading assessment framework. In I. V. S. Mullis & M. O. Martin (Eds.), *PIRLS 2021 assessment frameworks* (pp. 5–25). Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center. http://pirls2021.org/frameworks/home/reading-assessment-framework/overview/.

Oliveri, M. E., Rutkowski, D., & Rutkowski, L. (2018). *Bridging validity and evaluation to match international largescale assessment claims and country aims.* Research Report No. RR-18-27. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12214.

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-scale Assessments in Education, 2,* 5. https://doi.org/10.1186/s40536-014-0005-4.

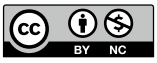
Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25,* 54–67. https://doi.org/10.1006/ceps.1999.1020.

Saeed, S., & Zyngier, D. (2012). How motivation influences student engagement: A qualitative case study. *Journal of Education and Learning, 1*, 2. http://www.ccsenet.org/journal/index.php/jel/article/view/19538.

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research, 51*(2), 161–179. https://doi.org/10.1080/00131880902891305.

Liz Twist is joint head of the Centre for Assessment at the National Foundation for Educational Research (NFER) in England. She has been involved with IEA's Progress in International Reading Literacy Study (PIRLS) since 2000 and has led a range of large scale national assessment development projects.

Julian Fraillon is the Director of the Assessment and Reporting (Mathematics and Science) Research Program at the Australian Council for Educational Research (ACER). Julian is the Study Director of the the IEA International Computer and Information Literacy Study (ICILS) 2018 and was the Study Director of the inaugural ICILS in 2013. He has been the Assessment Coordinator of the IEA International Civic and Citizenship Education Studies (2009, 2016, and 2022) and directs ACER's work on the Australian National Assessment Program studies of Civics and Citizenship and ICT Literacy. He is a member of the IEA Progress in International Reading Literacy Study (PIRLS) Reading Development Group.



Chapter 5
Questionnaire Development in International Large-Scale Assessment Studies

Wolfram Schulz and Ralph Carstens

Abstract Questionnaires are a crucial part of international large-scale studies of educational achievement, such as the International Association for the Evaluation of Educational Achievement (IEA) studies on mathematics, science, reading, computer and information literacy, and civic and citizenship education. Building on IEA's well-established technical standards, this chapter provides an overview of the purpose of this type of instrument, approaches to its development, and the evolving challenges in this area. While large-scale assessment studies have traditionally employed questionnaires to gather contextual information to explain variation in the respective outcome variables of interest, over time there has been a shift toward also collecting information and reports on students' attitudes, dispositions, or behaviors as outcome measures. More recently, development of alternative item formats and approaches have further increased the reliability, validity, and comparability within and across education systems. Approaches to questionnaire purpose and design can vary and instrument can be targeted to different groups and populations. Contextual information has to be collected across highly diverse educational systems creating issues regarding cross-national validity. The challenges of maximizing measurement invariance across highly diverse national contexts and the opportunities provided by a computer-based delivery may lead to future changes and improvements in the approach to questionnaire development.

Keywords Contextual information · Pre-testing · Questionnaire design · Question types and formats · Response process

W. Schulz (✉)
Australian Council for Educational Research (ACER), Camberwell, Australia
e-mail: wolfram.schulz@acer.org

R. Carstens
International Association for the Evaluation of Educational Achievement (IEA), Hamburg, Germany
e-mail: ralph.carstens@iea-hamburg.de

H. Wagemaker (ed.), *Reliability and Validity of International Large-Scale Assessment*,
IEA Research for Education 10, https://doi.org/10.1007/978-3-030-53081-5_5

5.1 Introduction

International studies of educational achievement, such as those conducted by IEA, routinely employ questionnaires to gather contextual information that can be used to explain variation in the outcome variables measured by educational achievement tests. These questionnaires are administered to students, teachers, schools, and/or parents, as well as national research coordinators in the participating countries. In some studies, questionnaires (in particular those administered to students) also play a key role in measuring affective-behavioral student learning outcomes in addition to cognitive test results. This important purpose is particularly relevant for IEA's civic and citizenship education studies (see IEA 2020a). There are also international large-scale assessments (ILSAs) that rely exclusively on (school and teacher) questionnaires to provide self-reported data on educational contexts as the main focus of cross-national surveys (such as IEA's Second Information Technology in Education Study [SITES]; see IEA 2020a).

Here we describe and discuss the different approaches regarding the purpose and design of the questionnaires that support ILSAs and the methods used to target these instruments to the most proximal information source, groups, populations and contexts. Differing question formats reflect the cognitive response process, and rigorous development procedures and quality assurance procedures are used to ensure the validity, reliability, and comparability of data. As a consequence, the IEA has published technical standards for its studies (Martin et al. 1999), and there are other salient guidelines and recommendations, especially the Survey Research Center's (2016) cross-cultural survey guidelines or summary volumes of contemporary issues and approaches (e.g., Johnson et al. 2018; Presser et al. 2004). Given the ongoing transition of paper-based to computer-based delivery in ILSAs, there are also implications related to the different modes of administration.

With regard to these different aspects, we discuss implications of recent developments in questionnaire design and reflect on the future role and characteristics of this type of instrument, including possibilities for improving the validity, reliability, and comparability of questionnaire data from international studies.

5.2 Approaches to Questionnaire Design and Framing

Questionnaire instruments in educational research may serve a range of different purposes. In many studies focused on measuring student achievement, the traditional role of questionnaires has been to provide information that may explain variation in the primary survey measures (e.g., mathematics skills or reading abilities). This type of questionnaire focuses on students' characteristics and home background, teachers' reports on learning environments, and/or principals' report on the educational context in which students learn.

Increasingly, other variables have been recognized as important in their own right, such as students' attitudes toward subject areas, their sense of self-efficacy, or self-regulated learning. While these variables may be related to achievement, their relationship is less clear in terms of expectations of possible causality and there is a growing recognition of their importance as outcome variables of interest in their own right. In certain studies, such as the IEA studies on civic and citizenship education (the Civic Education Study [CIVED] and International Civic and Citizenship Education Study [ICCS]; see IEA 2020a), measures derived from questionnaires (regarding civic attitudes and engagement) are as important in their role as learning outcomes as cognitive measures (civic knowledge and understanding). Obviously, questionnaire-only studies may also include important outcome measures, for example teachers' job satisfaction in the case of the Organisation for Economic Co-operation and Development's (OECD's) Teaching and Learning International Survey (TALIS; OECD 2019a), where all primary variables of interest are derived from questionnaire data.

In particular, in earlier IEA studies that were focused on measuring students' cognitive achievement, the development of questionnaires primarily aimed at gathering data about factors that explained variation in test scores. Apart from collecting basic information about student characteristics, such as gender and age, questionnaires of this kind typically aim to gather data about home and school contexts.

The type of information collected for this purpose is selected based on assumptions about what might provide explanatory factors predicting differences in educational achievement, and, consequently hint at aspects of system and school effectiveness. Typically, assessment frameworks for ILSAs not only describe the content of learning outcome domain(s) that should be measured but also include contextual frameworks, which outline the factors regarded as relevant for explaining variation in the learning outcome variables (see Chap. 3 for a more detailed description of the role of assessment frameworks). A secondary purpose of contextual questionnaires of this kind is also that data help to describe contexts in their own right, which is particularly useful in cases where statistical information (e.g., about school resources) is not available. For examples from the most recent cycles of IEA's Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) (see IEA 2020a), we refer readers to Hooper (2016) and Hooper and Fishbein (2017). For international perspectives with a focus on PISA, we advise readers to consult Kuger et al. (2016).

It is possible to distinguish between factual information about contexts (e.g., the number of books at home as reported by students and/or parents, or the number computers available at school as indicated by school principals or information and communication technology [ICT] coordinators) and perceptions of learning environment factors (e.g., student reports on student-teacher relations at school, or teacher reports on teacher collaboration at school). Factual information is often reported based on single variables (e.g., the average age of teachers) or variables are combined to create a new derived measure based on simple arithmetic calculations (e.g., the student-teacher ratio at school). Perceptions and attitudinal aspects are often measured through rating-scale items that then are either reported as single variables

or used to derive a scaled index (in the form of raw, factor, or item response theory [IRT] scores).

International studies with a traditionally strong focus on the measurement of student achievement, such as IEA's TIMSS and PIRLS, or the OECD's Programme for International Student Assessment (PISA), primarily collect questionnaire measures that are helpful to explain achievement. Increasingly though, ILSAs also include a number of affective-behavioral outcomes that relate factors such as attitudes toward learning, self-efficacy, or self-regulated learning, which do not always show strong correlations with achievement but help describe learning contexts. In the IEA's International Computer and Information Literacy Study (ICILS; see IEA 2020a), the measurement of students' and teachers' use of, familiarity with, and attitudes toward digital technology have always played an important role in its reporting of study results, and have received attention both in the initial reports and later secondary research (see Fraillon et al. 2014, 2020).

In IEA's studies of civic and citizenship education, measuring variables related to students' attitudes and engagement has traditionally been as important as the measurement of cognitive aspects (Schulz et al. 2010, 2018b; Torney et al. 1975; Torney-Purta et al. 2001). Here, affective-behavioral domains have a similar weight and importance for the reporting of study results and therefore the study places considerable emphasis on the development of questionnaire items that measure both contextual information and important civic-related learning outcomes. In this way, studies related to this learning area have achieved a more balanced representation of: (1) knowledge- and skill-related aspects; (2) attitudes and dispositions; and (3) practice, engagement, and behavioral intentions.

There are also studies that do not gather cognitive data reflecting student achievement where all reported results are derived from questionnaire instruments. Such studies include IEA's SITES 2006, OECD's TALIS 2008, 2013, and 2018 (OECD 2019a), and the OECD TALIS Starting Strong Survey 2018 (OECD 2019b), all of which have conducted surveys of school staff (teachers and/or educators) and management (school principals, ICT coordinators, and center leaders) (for more details see, e.g., Law et al. 2008 for SITES 2006, and Ainley and Carstens 2018 for OECD TALIS 2018). The aim of these studies is to gather and report information about school contexts, pedagogical beliefs, and teaching practices that are reported as criterion variables independently of achievement outcomes.

Each large-scale study and survey in education requires a careful articulation of research interests, aims, and data needs in order to obtain more advanced empirical insights with potential implications for educational policy, and the respective frameworks have important implications for the role assigned to questionnaires. For example, TIMSS 2019 used national curricula as the major organizing concept in considering how educational opportunities are provided to students, and emphasized those factors that influence these provisions and how they are used by students and teachers (Mullis and Martin 2017). In contrast, for example, the framework for ICCS 2016 (Schulz et al. 2016) laid more emphasis on the role of out-of-school contexts

and included perspectives beyond education, such as general perceptions of important issues in society, the economy and environment, and students' engagement in the local and wider community (e.g., through the use of social media).

5.3 Targeting of Questionnaires to Different Groups and a Diversity of Contexts

Consequently, there are a variety of contexts that may be of interest when collecting contextual material for a study. However, some respondents to a survey may not be sufficiently knowledgeable to provide data on all of the contexts of interest or (which is the central idea of this publication) may not be able to provide information and views with the same degree of validity and reliability. For example, while it is relatively straightforward to ask students and teachers about their perceptions of what happens during classroom interactions, school principals could provide broad information about what is expected with regard to policies or teaching practices at the school, but yet be unable to report on their actual implementation and/or the variation in perceptions regarding implementation.

The contexts that are of relevance for a particular study should be defined in the contextual framework together with the scope for gathering relevant data from respondents about aspects of interest. Contexts that may be of relevance in a particular international study can be mapped against a number of particular survey instruments that could be used to collect information about them (Table 5.1).

It is often also possible to ask more indirect questions. For example, school principals may be asked about the social background of enrolled students or their general expectations of teaching practices in classroom contexts. Furthermore, in smaller contexts (for example, in early learning studies), it might be appropriate to ask teachers (or early education center staff) about individual children.

Generally, when measuring contextual aspects, there are differences in how much information respondents may have about particular topics. For example, students (at least from a certain age onwards) and teachers can be expected to provide relatively reliable information about their personal characteristics (such as age or gender). However, judgments by school principals about the social context of the local community or the socioeconomic background of students at their school are likely to be less accurate. Furthermore, information gathered from students or teachers about what happens in the classroom also tends to differ considerably given their differing perspectives, even if the point of reference (the classroom) is identical (Fraser 1982).

The sample design may also have implications for the targeting of contextual questionnaires to contexts. In cases of a grade-based design (as is customary in most IEA studies), it is possible to gather specific data about defined classroom (or course) contexts from sampled students at school. In the case of an age-based design, where students are randomly sampled from a specific age group enrolled at selected schools (as in PISA), it is likely that student data reflect a wider range of experiences

Table 5.1 Mapping contexts to questionnaire types

Contexts	Student questionnaire	Parent questionnaire	Teacher questionnaire	School questionnaire	National contexts survey
Individual student	×	×			
Individual teacher			×		
Student peer context	×				
Student home context	×	×			
Classroom context	×		×		
School context	×	×		×	
Community context		×	×	×	
National context					×

Notes × = group is targeted by this questionnaire

than those from grade-based samples, as students from the same age group may be enrolled in different classrooms or course grade levels, or even study programs.

Having a classroom sample (as, e.g., in TIMSS or PIRLS) that relates to a common entity of pedagogical practice also provides an opportunity to ask the corresponding subject teacher specific questions about their teaching practices (see, e.g., Hooper 2016). However, in countries where a broader subject area is taught as part of different individual subjects (such as chemistry, biology, or physics), it remains challenging to relate information from teachers to student learning outcomes because pedagogical input into the content area may come from more than one subject teacher. This becomes even more difficult in cross-curricular learning areas (such as civic and citizenship education), where related content could be spread out across a larger variety of subjects with an even larger number of teachers who might have contributed in different ways and to differing extents to students' learning outcomes (see, e.g., Schulz and Nikolova 2004). Within countries, such as those with a federal state system or different jurisdictions, several approaches to organizing a learning area's curriculum may coexist (see European Commission/EACEA/Eurydice 2017).

One particular challenge related to questionnaire development is the inclusion of aspects that are relevant in many but not all countries. For example, questions about differences across study programs within schools may be of relevance in particular countries (such as Belgium or the Netherlands) but not in those where all students follow the same program in the grade under study (such as Australia, Finland, or Slovenia). To allow countries to pursue particular research as part of an international study, sometimes questionnaire sections are developed as international options that

are only included in those countries where these are regarded as relevant. To ensure the comparability of other (core) questionnaire data, there are limits to how much additional item material can be added. Furthermore, even if administered after the core questions overly long questionnaire may affect the response rates in countries participating in an option. A review of the extent to which optional material can be appropriately added to a study would ideally be part of an international field trial.

There are also cases where item material is only relevant to a sub-group of the target sample, for example for teachers of a particular subject or subject area. Here, questionnaires may include a filter question for teachers regarding their subject or subject area after which only certain teachers are asked to complete a section with questions relevant to their subject area. An example of such an approach can be found in IEA's ICCS, where all teachers teaching at the target grade are surveyed but only those teaching civic-related content are presented with more specific questions related to this particular learning area (see Agrusti et al. 2018).

Aspects of learning areas or subjects under study may also differ across geographic regions with a common historical, cultural, political, and/or educational context. Therefore, certain aspects may be relevant in one particular region but not in another. ICCS addresses this by including regional questionnaires that are developed in close cooperation with national research centers and their experts in these regions. In ICCS 2009, there were regional instruments for countries in Asia, Europe, and Latin America (Kerr et al. 2011), while ICCS 2016 administered regional questionnaires in Europe and Latin America (Agrusti et al. 2018). These regional instruments included civic-related content that was regarded as of particular importance for the respective geographic region or was related to specific aspects that would not have been appropriate for countries outside that region. For example, the European questionnaires in ICCS 2009 and 2016 measured perceptions related to the European Union or cooperation between European countries (Kerr et al. 2010; Losito et al. 2018) while the Latin American instruments focused on issues related to government, peaceful coexistence, and diversity (Schulz et al. 2011, 2018a).

For the successful development of questionnaires in international studies it is important to clearly define the scope of the survey with regard to the targeting of relevant aspects for measurement and appropriate sources of information. As part of the planning of the survey, instruments and the type of questions should be designed so that all aspects can be covered appropriately. It is important to consider which respondents can provide valid and reliable information about the contexts that are seen as relevant in a study. In cross-national studies, it is also crucial to keep in mind that the appropriateness of sources of contextual information may vary across countries. For example, when studying science education in classes, some countries may teach content in one combined subject or in separate subjects, which may require adaptations to the wording or design of subject-related survey instruments or when collecting data from teachers, students, and schools.

5.4 Typology of Questions, Item Formats and Resulting Indicators

As discussed in Sect. 3, a vast array of conceptual considerations, contexts, aims, and reporting needs drive the overall questionnaire design principles and the targeting to particular populations and contexts. The final questions used in the instruments are the primary interface between the aspirations and priorities of researchers working on the development and implementation of comparative surveys in education, and respondents participating in these studies. It needs to be emphasized that questionnaire material in ILSAs is typically delivered through the means of written, self-administered questionnaires. In the case of non-student populations, this is routinely done without the presence of (and possible assistance from) a survey administrator. While many other formats and methods are easy to imagine, including recordings, interview transcripts, or work products, the vast majority of ILSAs rely on the cost-effectiveness of written survey instruments.

Consequently, the questionnaire itself, and perhaps some framing comments made on informational letters, are the only sources of guidance available to respondents regarding the aims of the research, the types of information requested, and instruction to adequately respond to more complex questions. As a consequence, there are possible conflicts between research interests to collect data on complex characteristics, antecedents, inputs, processes, and outcomes and the need to develop and phrase survey questions that can actually be understood by respondents. This tension needs to be resolved in order to maximize the potential yield of valid and reliable information. In international, cross-cultural research, it is also common to encounter issues related to comparability of instruments. More specifically, there is a need to phrase questions in such a way that these can be appropriately translated into different sociocultural contexts and languages. This requirement adds another layer of complexity to the development of questionnaires in ILSAs.

The IEA's technical standards, which were developed at the end of the 1990s to enhance reliability and validity, state that questionnaires should be "clear, simple, concise and manageable" (Martin et al. 1999, p. 43). While the standards do not reflect important developments in survey methodology that have occurred more recently, this premise has not lost its relevance. The standards also request questionnaire development to be specific in terms of the results to be reported; to consider whether the questions will produce credible information, and to review each newly developed question carefully so that it relates to one idea only (i.e., to avoid double-barreled questions); to eschew open-ended questions in the final questionnaires; to ensure that response categories match the question intent, are mutually exclusive, and elicit responses that apply to all of the respondents; to use directions to skip to a later question if they do not apply to respondents; and to arrange questions within an instrument so that their flow is natural and sensible.

In contemporary ILSAs, including all IEA studies, these principles are key to the design of questionnaires, but there are also many other criteria that apply. For example, the use of a specific terminology could be appropriate for questions directed

at one population, such as teachers, but these may not be correctly understood by members of other populations, such as students or parents. Furthermore, seemingly similarly defined and worded terms (such as "students with special education needs") may be consistently understood within one education system but the terminology may differ for other education systems. Using examples in survey questions with the intention of clarifying certain terms (such as providing "Maths Olympics" as an example when asking principals or teachers about the frequency of "school activities to promote mathematical learning") may trigger a particular reference or frame of mind but could also potentially narrow the scope of responses to the particular set of examples.

In general, international study center staff working on questionnaire development, associated expert groups, and national research coordinators will need to carefully consider the cognitive processes involved in responding to surveys in order to match research aspirations with the realities of obtaining valid and reliable data. The response process itself may introduce measurement error or bias. Tourangeau et al. (2000) advised that, when asking respondents about information, researchers need to consider the following aspects: the original encoding/acquisition of an experience; the storage of that experience in long-term memory; comprehension of the survey question's task; retrieval of information from memory; integration/estimation from information retrieved; and mapping and editing a judgment/estimate to the response format. Bias in questionnaire data can be introduced related to each of these aspects, for example through the misunderstanding of a question, lack of memory when asked about relevant information too far back in time, poor estimation of requested information, or deliberate misreporting.

The type of information sought, such as a home context or a teacher's perception, will drive most of the wording of the corresponding question(s) for which there is a range of different approaches from simple to complex formats. The depth (or richness) of the desired characteristic or process that should be measured will be another criterion for the development of questions (i.e., whether researchers would like to check only the occurrence of an aspect, its frequency and/or intensity, information about its actual workings, or possible impacts).

Factual questions on the existence or frequency of current or very recent characteristics or events, or low inference sociodemographic questions (such as age or gender) for that matter, paired with simple closed response formats have a high probability of yielding valid and reliable information. These can be considered as low inference, namely easily observable or verifiable; appropriate formats would be multiple choice questions with nominal or ordinal response options (e.g., the type of school, frequency of a particular school process) or semi-open formats that require respondents to provide numbers (such as counts of enrolled students at school).

There are many aspects of relevance to ILSAs that cannot be measured in such a direct way. Studying behaviors, attitudes, intentions, or expectations requires the assumption of underlying constructs that cannot be observed directly; these can be termed high inference measures. Instead of formulating direct questions about these constructs, they tend to be measured by administering sets of items to respondents using response (rating) scales that typically have an ordinal level of measurement.

To gather indicators of underlying constructs ILSAs tend to use matrix-type formats to measure dimensions of interest (such as the observed use of ICT during lessons, respondents' sense of self-efficacy, or respondents' interest in learning a particular subject). Commonly used formats are frequency scales with fuzzy quantifiers (such as never, rarely, sometimes, often, or always) to measure frequencies of observations or behaviors, or rating scales reflecting the extent of agreement or disagreement (such as strongly agree, agree, disagree, and strongly disagree) regarding sets of statements relating to the same construct and its various aspects and dimensions.

In light of the model developed by Tourangeau et al. (2000), questions that require a high degree of cognitive burden at each stage and/or relate to more distant events or occurrences would be expected to have a higher probability of introducing measurement error. Similarly, questions of highly personal or sensitive nature may elicit deliberate over- or underreporting in order to preserve the desired self-image of a respondent.

Data for which there are no obvious coding schemes or categorizations, may be initially measured using questions with an open format at the pilot or field trial stage in order to identify appropriate factual response options that are comprehensive and mutually exclusive. Classification of open-ended questions using responses from smaller samples can identify aspects that are not clear at the outset of the development process but relevant from the perspective of respondents.

These examples by no means provide a complete picture of the issues related to finding appropriate questionnaire item formats and contents. Harkness et al. (2016) have further illustrated the richness of the debate, and the options and choices available for the development of questionnaires. Lietz (2010, 2017) and Jude and Kuger (2018) summarized persisting and emerging debates surrounding the design of questionnaires, links to theory, and the overall aim of generating data for educational change. Each study needs to find an appropriate balance between its research aims and aspirations and corresponding practical limitations by using expert judgement in the process of writing new questions, adapting existing questions from validated sources, or refining existing questions. For example, asking students about their parents' or guardians' occupations has been one of the most debated questions in international studies of education, since staff at national research centers generally have to interpret students' (often limited) responses and code these to international standards; it is difficult to evaluate the cost-effectiveness of this process.

Critical debates surrounding best practice for questionnaire development (see, e.g., Harkness et al. 2016; Lietz 2010) have also focused on the use of even versus odd category numbers for Likert-scales (i.e., whether it is appropriate to include or exclude a neutral midpoint in the response categories), unipolar versus bipolar response scales, the direction of response scales (i.e., should they always run positive to negative, or from most to least frequently used, or always in the direction of the latent construct), the concurrent use of positive and negative statements (often resulting in some effect), and the use of definitions and examples to guide respondents' answers. Here, IEA studies routinely aim to minimize the cognitive burden for respondents and avoid inconsistent question and response option design within and across cycles.

Since 2010, novel and innovative item and question formats in ILSAs have evolved. Developmental work on new formats has been conducted primarily in the context of OECD's PISA (OECD 2014), but to some extent also in IEA's ICCS and OECD's TALIS. Examples of these innovative research activities include experiments using candidate methods for improving the reliability and cross-cultural validity of self-assessment measures, so-called Bayesian Truth serum, topic familiarity, forced choice, anchoring vignettes, and situational judgment tests (SJTs) (see, e.g., Jude and Kuger 2018). For example, the field trial of TALIS 2013 included a specific measure to capture teachers' negative and positive impression management behavior (seen as an indicative of socially desirable responses), the field trial of IEA ICCS included forced choice formats, and the field trial of TALIS 2018 included SJTs.

However, as Jude and Kuger (2018) found, these formats and methods have only been able to demonstrate limited success in increasing the validity, reliability, and comparability of questionnaire measures, and many of these formats have only relatively poor cost-effectiveness. This may be related to ethical concerns (e.g., when using fictitious concepts), the cognitive complexity of some measures (in particular when using anchoring vignettes or forced-choice formats with students), the hypothetical nature of situation (in particular SJTs), increased reading load (SJTs and anchoring vignettes), and recognition that these alternative formats demonstrated limited potential to measure and correct for differential response styles within and across countries. With respect to the examples above, while novel item formats were trialed in TALIS 2013, ICCS 2016, and TALIS 2018, they were not included in the main data collections. While research related to new item formats continues within and outside the field of ILSAs, their usage is currently quite limited and how effectively these formats can augment or replace established questionnaire design formats in the future remains unclear.

An ILSA's success depends on the representation of different types of actors and experts in the drafting process. There appears to be an increasing level of convergence across different studies, which is further facilitated by the sharing of expertise by technical advisory boards (such as the IEA's Technical Executive Group), experienced study center staff, distinguished research experts and experts from international organizations (such as IEA), and international collaboration and exchange; all helps to advance the quality, validity, and reliability of questionnaire measurement.

5.5 Development Procedures, Process and Quality Management

Any development of questionnaires should ideally be grounded in a conceptual framework that describes the aspects that should be measured with such instruments (see Chap. 3 for further discussion). This framework needs to define the range of factors that are of relevance, either in terms of providing explanation for learning

outcomes or in terms of deriving survey outcome variables. Developing a conceptual underpinning for questionnaire development tends to be particularly challenging when a wide range of diverse national educational contexts are involved, as is typically the case in ILSA.

Large-scale assessments that are designed as cyclical to monitor changes over time face a particular challenge when it comes to questionnaire development for each new cycle. There is demand from stakeholders to retain material so the same measures can be repeated and hence study data can inform on how contexts, perceptions, attitudes, or behaviors have changed since the last cycle(s). However, there is also a conflicting demand to include new material that addresses recent developments or improves previously used measures.

Questionnaires should be completed within an appropriate time frame that avoids respondent fatigue or refusal to participate due to overly long instruments. Experiences from previous cycles and international field trial studies are used to determine an appropriate length and this can depend on different factors, for example, whether a questionnaire is administered after a lengthy cognitive assessment of two hours or a relatively short assessment of less than one hour. Typically, questionnaires for students, teachers, and other respondents are expected to take between 30 min to one hour to complete (including additional optional components).

Given these time restrictions on instrument length, it is a challenge reconciling the need to retain "trend" measures with providing sufficient space for newly-developed items that address evolving areas and/or replace questions that may be viewed as outdated. The process for making decisions on the retention of old material or inclusion of new material can become particularly difficult within the context of international studies, where the diversity of national contexts may lead to differing priorities and views on the appropriateness of retention and renewal of item material.

Once a conceptual framework has been elaborated, the procedure for item development should include as many reviews and piloting activities as permitted by the often quite restricted time frames. Ideally, the item development phase should include:

- Expert reviews at various stages (for international studies these should also include national representatives with expertise in the assessed domain);
- Cognitive laboratories and focus group assessments for qualitative feedback;
- Translatability assessments for studies where material needs to be translated from a source version (typically in English) into other languages (as is usually the case in international studies);
- Initial piloting of new item material with smaller samples of respondents (in international studies this should involve as many participating countries as possible); and
- A general field trial (in international studies this should include all participating countries) that provides an empirical basis for item selection.

Piloting activities (either qualitative or quantitative) have a strong focus on the suitability of the new item material, but are often not conducted in all participating countries and tend to be based on smaller convenience samples. The inclusion of

questionnaire material in international field trials, in turn, aims to review the appropriateness of an instrument that broadly resembles the final main survey instrument. While this may not always include all of the retained item material from previous cycles, it often includes both old and new items in conjunction; this enables questionnaire designers to look into associations between constructs and review how well new items developed for already existing scales measure the same underlying constructs as the old material.

As already is often the case with quantitatively oriented piloting activities, for the field trial it may be appropriate to use more than one form in order to trial a broader range of item material, given the constraints in terms of questionnaire length. It is possible to arrange the distribution of item sets so that there is overlap and all possible combinations of scales and items can be analyzed with the resulting data sets (see, e.g., Agrusti et al. 2018). Another advantage provided by administering questionnaire material in different forms is the ability to trial alternative formats. For example, researchers may be interested in finding out whether it is more appropriate to use a rating scale of agreement or a scale with categories reflecting frequencies of occurrence to measure classroom climate (as perceived by students or teachers).

Analyses of field trial data tend to focus on issues such as:

- Appropriateness of instrument length and content for the surveyed age group (e.g., through a review of missing data);
- Scaling properties of questionnaire items designed to measure latent traits (e.g., self-efficacy or attitudes toward learning) using classic item statistics, factor analysis, and item response modeling;
- Comparisons of results from questionnaire items included in (a) previous cycle(s) with those newly developed for the current survey;
- Analyses of associations between contextual indicators and potential outcome variables; and
- Reviews of measurement invariance across national contexts using item response modeling and/or multi-group confirmatory factor analysis.

A variety of factors may affect the comparability of questionnaire response data in cross-national studies, and the formats typically used to gauge respondents' attitudes or perceptions may not always consistently measure respondents' perceptions and beliefs across the different languages and cultures (see, e.g., Byrne and van de Vijver 2010; Desa et al. 2018; Heine et al. 2002; van de Gaer et al. 2012; Van de Vijver et al. 2019). With this in mind, international studies have started to build in reviews of measurement invariance during the development stage (see, e.g., Schulz 2009; Schulz and Fraillon 2011). At the field trial stage in particular, with data collected across all participating countries, this type of analysis may identify a potential lack of measurement invariance at item or scale level prior to inclusion in the main survey.

Another important challenge when developing questionnaires is to avoid questions that cause respondents to give answers that are biased toward giving a positive image of themselves, their work, or their institution. Tendencies to provide socially desirable responses in a survey may also vary across national contexts and can be regarded as a potential source of response bias in international studies (Johnson and

Van de Vijver 2003; Van de Vijver and He 2014). While researchers have proposed scales that were developed to measure a construct of social desirability (see, e.g., Crowne and Marlowe 1960), research has also shown that it is difficult to use them for detection and/or adjustment of this type bias, given that they also measure content that cannot be easily disentangled from style (see, e.g., McCrae and Costa 1983). Therefore, while it is important to acknowledge tendencies to give socially desirable answers to certain types of questions, which should be considered in the process of developing and reviewing a question's validity, there is no agreed way of empirically investigating this as part of piloting activities or a field trial.

In summary, any questionnaire development should ideally undergo multiple quality assurance procedures embedded throughout the process. A clear reference document (framework) that outlines research questions, scope, design, and content relevant for the development of questionnaire material in international studies is of critical importance. A staged process that includes different stages of review by national staff and experts, qualitative and quantitative vetting at the earlier stages (ideally including translatability assessments), and a field trial that allows a comprehensive review of cross-national appropriateness and psychometric quality of the item material provide the best option for thorough evaluation of the item material.

5.6 Questionnaire Delivery

In recent years, the delivery of ILSA questionnaires to different target populations has transitioned from traditional paper-based instruments to the use of computer-based technology and the internet. Throughout the questionnaire development process, the choice of the delivery mode for questionnaires and its design have important implications for pretesting, and adaptation and translation (Survey Research Center 2016) Computer-based delivery also provides additional opportunities to collect and use auxiliary process data from electronically delivered questionnaires, as well as the ability to design instruments that enable matrix-sampling of items.

Paper-and-pencil administration of questionnaires was the only viable option for ILSAs of education during the 20th century, although research into internet-delivered surveys had been conducted during the 1990s in relation to public-opinion, health, or household-based surveys (see, e.g., Couper 2008; Dillman et al. 1998). In ILSAs, questionnaires designed for self-completion were typically administered to students as part of a paper-based test during the same assessment session managed by a common administrator. Other contextual questionnaires delivered to adult populations, such as school principals, teachers, or parents, were truly self-administered on paper at a time and location chosen by the respondents and later returned to a school coordinator or mailed directly to the study center. Questionnaire completion as part of a student assessment session was loosely timed, while the self-administration to an adult population was untimed (i.e., respondents could take as little or as much time as they needed).

With the rapidly growing penetration and availability of computers and internet connectivity in schools and at home in the late 1990s and early 2000s, the conditions for educational surveys also changed. The IEA pioneered and trialed the first web-based data collection as part of SITES 2006 (Law et al. 2008). Here, the mode of data collection matched the study's research framework (i.e., the study investigated how and to what extent mathematics and science teachers were using ICT within and outside the classroom). The study offered online administration of teacher, principal, and ICT-coordinator questionnaires to all participating countries; however, not all of them chose this the primary mode and some opted for a primarily or exclusively paper-based delivery (Carstens and Pelgrum 2009). While some countries made online administration the primary mode of collection, others made it optional, while other countries decided to administer the survey on paper only. Overall, about 72% of all questionnaires were administered online and, in the 17 (out of 22) countries that used online collection, about 88% of all respondents used the online mode. There was very little variation between the different groups of respondents (e.g., teachers and principals) but choice of delivery mode differed considerably by other characteristics, in particular across age groups.

Furthermore, SITES 2006 investigated issues of measurement invariance across modes using a split-sample design at the field trial stage. As expected, based on prior research findings, the study observed no major differences in response behavior, styles, non-response, or completion time. Regardless of the delivery mode, questionnaires were self-administered without the presence of an administrator, which is viewed as a key factor explaining differences in response behavior (Tourangeau et al. 2000).

The SITES study and other studies, such as the first cycle of OECD's TALIS in 2008 (a survey implemented by IEA), and IEA's ICCS 2009, paved the way for further work in the area and yielded important insights for the design and administration of online questionnaires accompanied by an alternative paper-based delivery. For example, studies had to find efficient and effective ways to manage the instrument production processes, including adaptation, translation, and verification, without duplicating work (and hence duplicating the chance of errors) for international and national study centers planning to administer paper and online questionnaires side by side.

This paradigm shift has also raised important questions regarding the instrument layout. When using dual delivery modes, obtaining comparable data across the two modes is essential. However, this does not necessarily require an identical design and presentation in both modes, which would be a rather challenging and, possibly impossible endeavor. For example, ILSAs using a computer-based delivery typically present one question at a time in online mode, whereas paper instruments might include multiple (albeit short) questions on one single page.

Skipping logic in online mode has the potential of reducing the response burden further by taking respondents directly to the next applicable question rather than relying on the respondent to omit irrelevant questions. Additional validation and review options can be included, such as a hyperlinked table of contents or checks for plausible number ranges and formats. Furthermore, in cases where a dual mode

is available within the same country (as is often the case in online questionnaires for school principals, teacher, or parents), respondents have the option of requesting or accessing paper versions of questionnaires instead of completing them online, a technical standard aimed at preventing respondents from being excluded because of technical requirements.

As time progresses, access to the internet and the number of respondents able to complete questionnaires online is expected to grow. Correspondingly, across different cycles of IEA studies (see the respective technical reports for TIMSS, PIRLS, ICCS, and ICILS; IEA 2020b), the uptake of online delivered questionnaires has generally increased. For example, while in ICCS 2009 only five out of 38 participating countries opted for online delivery, in ICCS 2016, 16 out of 24 countries selected this option. Finally, the technical design of the questionnaire delivery systems used in these studies make no assumptions or requirements about a particular device, internet browser make or type, or available auxiliary software (such as JavaScript), allowing unrestricted access to online questionnaires by reducing or eliminating technical hurdles. However, issues of confidentiality, security, and integrity of online collected data have started to play an increasingly important role in recent years, in response to public concerns and tightened legal standards.

The shift of the primary collection mode for questionnaires, and later assessments, from paper-based to computer-based delivery has introduced two important opportunities of high relevance. First, computer-based/online delivery enables the collection of process and para-data that can facilitate important insights into the quality of question materials and resulting responses through an analysis of response behavior. Second, electronic delivery potentially enables a more targeted delivery of questionnaires, which could lead to improvements to the so far relatively simple rotational approaches used by some ILSAs, for example, at the field trial stage in ICCS, ICILS, and OECD's TALIS (Agrusti et al. 2018; Ainley and Schulz 2020; Carstens 2019), or in the main survey as in PISA 2012 (OECD 2014).

With paper-based questionnaires, there is only very limited information on response behavior to assert the quality of the instruments. In student sessions, report forms completed by test administrators provide information regarding certain aspects, such as timing or anomalies and deviations from uniform conditions during assessment sessions. Therefore, information on the way in which respondents react to the questionnaire material delivered on paper is generally only obtained through pretesting at the pilot and field trial stages, which generates narrative and partly anecdotal information.

Electronic delivery of instruments, however, provides information beyond this, and allows statistical and other analyses of the substantial response data in conjunction with log data. Kroehne and Goldhammer (2018) proposed a framework to conceptualize, represent, and use data from technology-based assessment, explicitly including log data collected for contextual questionnaires. Their model encompasses an access-related category (including, e.g., assessment setting and device information), a response-related category (e.g., input-related events), finally a process category (e.g., navigation). Hu's (2016) cross-cultural survey guidelines provide a similar conceptualization and recommendations for the purpose of reviewing and

explaining non-response at the case and item level, or analyzing aberrant responses and/or "satisficing" (a term combining satisfy and suffice, referring to the idea that people do not put as much effort into responding as they should; see Tourangeau et al. 2000).

A particular benefit from the approach proposed by Kroehne and Goldhammer (2018) relates to the aim of generating indicators from individual events, states, and the in-between transitions that have the potential of informing survey designers about response processes from the perspective of the respondents (e.g., regarding timing, navigation, drop-out and non-response), and the individual questions and items (such as average time needed to respond by assessment language, scrolling, or changes of responses). Data and indicators can then generate insights with respect to the technical behavior of systems, access limitations, and device preferences, which may all assist with optimization at the system level. More importantly, the data and indicators can provide powerful insights into the functioning of the question materials, identifying questions and items requiring a disproportionate long response time or indicating error-prone recollection requirements (Tourangeau et al. 2000). In addition, process data may provide data about respondents' engagement, or disengagement, and the extent to which the collected information validly relates to the questions or is a result of inattentive or otherwise aberrant behavior, which may include deliberate falsification (for an overview of detection methods, see, e.g., Steedle et al. 2019).

The second promising aspect of electronic delivery relates to the ability to deliver questionnaires in a non-linear way to respondents. To date, virtually all ILSAs deliver one version of a questionnaire to respondents at the main data collection stage. However, there is a strong interest in broadening the conceptual depth and breadth of measures in questionnaires to yield additional insights for educational policy. The situation is similar to one of the most important late 20th century advancements in cognitive assessments in education, the so-called "new design" implemented by the National Assessment of Educational Progress (NAEP) to broaden the assessment's domain scope and insights for policy while managing the response burden for individuals (Mislevy et al. 1992). Essentially, the responses (and, by extension, the derived variables) for items not administered to an individual student were treated as a missing data problem, addressed through the use of IRT and latent regression modeling based on Bayesian approaches accounting for imputation variance.

IEA's technical standards (Martin et al. 1999) acknowledged the similar potential for questionnaires early on:

> Consider whether matrix sampling may be appropriate in the development of the questionnaire. Matrix sampling in this case means that not all respondents are asked all questions. Although this method will add some cost and complexity, it can greatly reduce response burden. Matrix sampling should only be considered if the study objectives can be met with adequate precision.

While this matrix sampling is nowadays firmly established in internationally comparative cognitive assessments, it remains to be seen if such an approach can be carried over to the case of questionnaires. Some relevant research related to these aspects has already been undertaken (Adams et al. 2013; Kaplan and Su 2018; von Davier 2014). Electronic delivery coupled with modern statistical approaches,

such as predictive mean matching, is believed to have potential for more elaborate sequencing of materials, matrix sampling approaches, and other aspects. Insights from such research could in due time become the basis for new technical standards for future IEA studies.

5.7 Conclusions

Questionnaires are a well-established component of ILSAs and provide crucial information on context variables and potential outcome variables. From earlier uses as auxiliary instruments to provide data to explain achievement results, questionnaires have grown in importance in recent international studies.

The sophistication of development and implementation procedures has also grown in recent decades. With regard to cross-national educational research, there has been an increasing recognition of the importance of considering the validity and reliability of questionnaire instruments and measures. The requirement of cross-national comparability is a crucial element that is emphasized by the fact that, across ILSAs, increasing attention is paid to questionnaire outcomes as a way of comparing student learning and educational contexts. The recognition of the potential bias resulting from differences in national contexts when using questionnaires in these studies has led to an increased focus on thorough review, piloting, and trialing of material as part of questionnaire development, with further analyses aimed at detecting or ameliorating non-equivalence of translations or resulting measures.

Here, we have described some of the main approaches to the questionnaires applied in international studies of educational achievement, their targeting and tailoring to distinct respondent groups, the variety of measures, indicators, and formats, the procedures typically implemented to ensure thorough development, challenges resulting from cross-national measurement, and the transition from paper-based to electronic delivery. In particular the challenges of maximizing measurement invariance across highly diverse national contexts and the opportunities provided by computer-based delivery are expected to result in interesting developments in the near future, which may lead to further changes and improvements in the approach to questionnaire elaboration and implementation in ILSAs.

References

Adams, R. J., Lietz, P., & Berezner, A. (2013). On the use of rotated context questionnaires in conjunction with multilevel item response models. *Large-scale Assessments in Education, 1*, 5. https://largescaleassessmentsineducation.springeropen.com/articles/10.1186/2196-0739-1-5.

Agrusti, G., Ainley, J., Losito, B., & Schulz, W. (2018). ICCS questionnaire development. In W. Schulz, R. Carstens, B. Losito, & J. Fraillon (Eds.), *ICCS 2016 technical report* (pp. 21–32). Amsterdam, The Netherlands: IEA. https://www.iea.nl/publications/technical-reports/iccs-2016-technical-report.

Ainley, J., & Carstens, R. (2018). *Teaching and Learning International Survey (TALIS) 2018 conceptual framework*. OECD Education Working Papers, No. 187. Paris, France: OECD Publishing. https://doi.org/10.1787/799337c2-en.

Ainley, J., & Schulz, W. (2020). ICILS 2018 questionnaire development. In J. Fraillon, J. Ainley, W. Schulz, T. Friedman, & S. Meyer (Eds.). *ICILS 2018 technical report* (pp. 39–47). Amsterdam, the Netherlands: IEA. https://www.iea.nl/publications/technical-reports/icils-2018-technical-report.

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing, 10*(2), 107–132.

Carstens, R. (2019). Development of the teacher and principal questionnaires. In OECD (Ed.), *TALIS 2018 technical report* (pp. 53–84). Paris, France: OECD Publishing.

Carstens, R., & Pelgrum, J. P. (2009). *Second Information Technology in Education Study. SITES 2006 technical report*. Amsterdam, The Netherlands: IEA. https://www.iea.nl/publications/technical-reports/second-information-technology-education-study-technical-report.

Couper, M. (2008). *Designing effective web surveys*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511499371.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4), 349–354.

Desa, D., van de Vijver, F., Carstens, R., & Schulz, W. (2018). Measurement invariance in international large-scale assessments: Integrating theory and method. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methodology* (pp. 881–910). Chichester, UK: Wiley.

Dillman, D. A., Tortora, R. D., & Bowker, D. (1998). *Principles for constructing web surveys: An initial statement*. Technical report No. 98-50. Pullman, WA: Washington State University Social and Economic Sciences Research Center.

European Commission/EACEA/Eurydice. (2017). *Citizenship education at school in Europe–2017*. Eurydice Report. Luxembourg: Publications Office of the European Union. https://eacea.ec.europa.eu/national-policies/eurydice/content/citizenship-education-school-europe-%E2%80%93-2017_en.

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2020). *Preparing for life in a digital world: IEA International Computer and Information Literacy Study 2018 international report*. Cham, Switzerland: Springer. https://www.iea.nl/publications/study-reports/preparing-life-digital-world.

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). *Preparing for life in a digital age. The IEA International Computer and Literacy Information Study international report.* Cham, Switzerland: Springer. https://www.iea.nl/publications/study-reports/international-reports-iea-studies/preparing-life-digital-age.

Fraser, B. J. (1982). Differences between student and teacher perceptions of actual and preferred classroom learning environment. *Educational Evaluation and Policy Analysis, 4*(4), 511–519.

Harkness, J., Bilgen, I., Córdova Cazar, A., Hu, M., Huang, L., Lee, S., Liu, M., Miller, D., Stange, M., Villar, A., & Yan, T. (2016). *Questionnaire design. Guidelines for best practice in cross-cultural surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. http://www.ccsg.isr.umich.edu/.

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference group effect. *Journal of Personality and Social Psychology, 82*(6), 903–918.

Hooper, M. (2016). Developing the TIMSS 2015 context questionnaires. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 2.1–2.8). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. http://timss.bc.edu/publications/timss/2015-methods/chapter-2.html.

Hooper, M., & Fishbein, B. (2017). Developing the PIRLS 2016 context questionnaires. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 2.1–2.8). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-2.html.

Hu, M. (2016). *Paradata and other auxiliary data. Guidelines for best practice in cross-cultural surveys.* Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. http://www.ccsg.isr.umich.edu/.

IEA. (2020a). IEA studies [webpage]. Amsterdam, The Netherlands: IEA. https://www.iea.nl/studies/ieastudies.

IEA. (2020b). Publications [webpage]. Amsterdam, The Netherlands: IEA. https://www.iea.nl/publications.

Johnson, T. P., & van de Vijver, F. J. R. (2003). Social desirability in cross-cultural research. In J. A. Harkness, F. J. R. van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 195–204). Hoboken, New Jersey: Wiley.

Johnson, T. P., Pennell, B. E., Stoop, I., & Dorer, B. (Eds.). (2018). *Advances in comparative survey methodology*. Chichester, UK: John Wiley & Sons Ltd.

Jude, N., & Kuger, S. (2018). *Questionnaire development and design for international large-scale assessments (ILSAs): Current practice, challenges, and recommendations.* Workshop series on methods and policy uses of international large-scale assessments (ILSAs). Washington, DC: National Academy of Education. http://naeducation.org/wp-content/uploads/2018/04/Jude-and-Kuger-2018-FINAL.pdf.

Kaplan, D., & Su, D. (2018). On imputation for planned missing data in context questionnaires using plausible values: a comparison of three designs. *Large-scale Assessments in Education, 6*, 6. https://largescaleassessmentsineducation.springeropen.com/articles/10.1186/s40536-018-0059-9.

Kerr, D., Schulz, W., & Fraillon, J. (2011). The development of regional instruments. In W. Schulz, J. Ainley, & J. Fraillon (Eds.), *ICCS 2009 technical report* (pp. 45–49). Amsterdam, The Netherlands: IEA. https://www.iea.nl/publications/technical-reports/iccs-2009-technical-report.

Kerr, D., Sturman, L., Schulz, W., & Bethan, B. (2010). *ICCS 2009 European report. Civic knowledge, attitudes and engagement among lower secondary school students in twenty-four European countries.* Amsterdam, The Netherlands: IEA. https://www.iea.nl/publications/study-reports/international-reports-iea-studies/iccs-2009-european-report.

Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika, 45,* 527. https://doi.org/10.1007/s41237-018-0063-y.

Kuger, S., Klieme, E., Jude, N., & Kaplan, D. (Eds.). (2016). *Assessing contexts of learning: An international perspective*. Cham, Switzerland: Springer.

Law, N., Pelgrum, W. J., & Plomp, T. (2008). *Pedagogy and ICT use in schools around the world. Findings from the IEA SITES 2006 study*. Hong Kong: CERC-Springer.

Lietz, P. (2010). Research into questionnaire design: A summary of the literature. *International Journal of Market Research, 52*(2), 249–272. https://doi.org/10.2501/S147078530920120X.

Lietz, P. (2017). Design, development and implementation of contextual questionnaires in large-scale assessments. In P. Lietz, J. Cresswell, K. Rust, & R. Adams (Eds.), *Implementation of large scale education assessments* (pp. 92–136). Chichester, UK: Wiley.

Losito, B., Agrusti, G., Damiani, V., & Schulz, W. (2018). *Young people's perceptions of Europe in a time of change: IEA International Civic and Citizenship Education Study 2016 European report*. Amsterdam, The Netherlands: IEA. https://www.iea.nl/publications/study-reports/international-reports-iea-studies/young-peoples-perceptions-europe-time.

Martin, M. O. Rust, K., & Adams, R. J. (Eds.) (1999) *Technical standards for IEA studies.* Amsterdam, The Netherlands: IEA. https://www.iea.nl/publications/iea-reference/technical-standards-iea-studies.

McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology, 51,* 882–888.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133–161. https://doi.org/10.2307/1434599.

Mullis, I. V. S., & Martin, M. O. (Eds.) (2017). *TIMSS 2019 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://www.iea.nl/publications/assessment-framework/timss-2019-assessment-frameworks.

OECD. (2014). *PISA 2012 technical report*. Paris, France: OECD Publishing. https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf.

OECD. (2019a). TALIS: The OECD Teaching and Learning International Survey. Paris, France: OECD. http://www.oecd.org/education/talis/.

OECD. (2019b). OECD Starting Strong Teaching and Learning International Survey. Paris, France: OECD. http://www.oecd.org/education/school/oecd-starting-strong-teaching-and-learning-international-survey.htm.

Presser, P., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., et al. (Eds.). (2004). *Methods for testing and evaluating survey questionnaires*. Hoboke, NJ: Wiley.

Schulz, W. (2009). Questionnaire construct validation in the International Civic and Citizenship Education Study. In M. von Davier & D. Hastedt (Eds.), *Issues and methodologies in large-scale assessments*, IERI Monograph Series (Vol. 2, pp. 113–135). Hamburg, Germany: IERI. http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_05.pdf.

Schulz, W., & Fraillon, J. (2011). The analysis of measurement equivalence in international studies using the Rasch model. *Educational Research and Evaluation, 17*(6), 447–464.

Schulz, W., & Nikolova, R. (2004). Translation procedures, field operations and quality assurance. In W. Schulz & H. Sibberns (Eds.), *IEA Civic Education Study technical report (*pp. 27–40). Amsterdam, The Netherlands: IEA. https://www.iea.nl/publications/technical-reports/iea-civic-education-study-technical-report.

Schulz, W., Ainley, J., Cox, C., & Friedman, T. (2018a). *Young people's views of government, peaceful coexistence, and diversity in five Latin American countries. The International Civic and Citizenship Education Study 2016 Latin American Report*. Amsterdam, The Netherlands: IEA. https://www.iea.nl/publications/study-reports/international-reports-iea-studies/young-peoples-views-government.

Schulz, W., Ainley, J., Fraillon, J., Kerr, D. & Losito, B. (2010). *ICCS 2009 international report. Civic knowledge, attitudes and engagement among lower secondary school students in thirty-eight countries*. Amsterdam, The Netherlands: IEA. https://www.iea.nl/publications/study-reports/international-reports-iea-studies/iccs-2009-international-report.

Schulz, W., Ainley, J., Friedman, T. & Lietz, P. (2011). *ICCS 2009 Latin American report: Civic knowledge and attitudes among lower secondary students in six Latin American countries*. Amsterdam, The Netherlands: IEA. https://www.iea.nl/publications/study-reports/international-reports-iea-studies/iccs-2009-latin-american-report.

Schulz, W., Ainley, J., Fraillon, J., Losito, B., & Agrusti, G. (2016). *IEA International Civic and Citizenship Education Study 2016 assessment framework*. Cham, Switzerland: Springer. https://www.iea.nl/publications/assessment-framework/iea-international-civic-and-citizenship-education-study-2016.

Schulz, W., Ainley, J., Fraillon, J., Losito, B., Agrusti, G., & Friedman, T. (2018b). *Becoming citizens in a changing world: The International Civic and Citizenship Education Study 2016 international report*. Cham, Switzerland: Springer. https://www.iea.nl/publications/study-reports/international-reports-iea-studies/becoming-citizens-changing-world.

Steedle, J. T., Hong, M., & Cheng, Y. (2019). The effects of inattentive responding on construct validity evidence when measuring social–emotional learning competencies. *Educational Measurement: Issues and Practice*, *38*, 101–111. https://doi.org/10.1111/emip.12256.

Survey Research Center. (2016). *Guidelines for best practice in cross-cultural surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. http://www.ccsg.isr.umich.edu/.

Torney-Purta, J., Lehmann, R., Oswald, H. & Schulz, W. (2001). *Citizenship and education in twenty-eight countries: Civic knowledge and engagement at age fourteen.* Amsterdam, The Netherlands: IEA. https://www.iea.nl/publications/study-reports/international-reports-iea-studies/citizenship-and-education-twenty-eight.

Torney, J., Oppenheim, A. N., & Farnen, R. F. (1975). *Civic education in ten countries: An empirical study*. New York, NY, USA: Wiley.

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511819322.004.

von Davier, M. (2014). Imputing proficiency data under planned missingness in population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 175–201). Boca Raton, FL: CRC Press.

Van de Gaer, E., Grisay, A., Schulz, W., & Gebhardt, E. (2012). The reference group effect: An explanation of the paradoxical relationship between academic achievement and self-confidence across countries. *Journal of Cross-Cultural Psychology, 43,* 1205–1228.

Van de Vijver, F. J. R., & He, J. (2014). *Report on social desirability, midpoint and extreme responding in TALIS 2013*. OECD Education Working Papers, No. 107. Paris, France: OECD Publishing. https://www.oecd-ilibrary.org/education/report-on-social-desirability-midpoint-and-extreme-responding-in-talis-2013_5jxswcfwt76h-en.

Van de Vijver, F. J. R., Jude, N., & Kuger, N. (2019). Challenges in international large-scale educational surveys. In B. Denman, L. E. Suter, & E. Smith (Eds.), *Sage handbook of international comparative research* (pp. 83–102). London, UK: SAGE.

Wolfram Schulz is a Principal Research Fellow (formerly Research Director International Surveys) at the Australian Council for Educational Research (ACER) where he has worked on a large number of national and international large-scale assessment studies. He is International Study Director of the IEA International Civic and Citizenship Education Study (ICCS) and Assessment Coordinator for the IEA International Computer and Information Literacy Study (ICILS). He is also a member of the IEA Technical Executive Group (TEG).

Ralph Carstens is a Senior Research Advisor at the International Association for the Evaluation of Educational Achievement (IEA) in Hamburg. With a professional background in primary and lower secondary teaching, Ralph acquired deep knowledge and experience relating to the conceptual development, instrument design, survey methodology, data work, analysis, reporting, and communication of international large-scale surveys in education. Ralph recently co-directed the IEA International Civic and Citizenship Education Study (ICCS) 2016 study and chaired the Questionnaire Expert Group for the OECD Teaching and Learning International Survey (TALIS). He coordinates the initiation and conceptual development of new studies, projects, and partnerships with a focus on non-IEA studies. He further provides advice on technical and methodological solutions and innovations to IEA management.

Chapter 6
Translation: The Preparation of National Language Versions of Assessment Instruments

Paulína Koršňáková, Steve Dept, and David Ebbs

Abstract To ensure the quality of its large-scale comparative studies, the International Association for the Evaluation of Educational Achievement (IEA) has created technical standards and guidelines, and developed a wealth of experience related to instrument production. The content experts who served as test editors and the professional linguists working toward achieving equivalence of assessment materials during the verification processes informed this chapter: it provides an overview of the rationale and implementation of linguistic quality control procedures for the instruments used in IEA studies, covering the general steps, key aspects, vocabulary, and discourse related to the translation of the study instruments. There are many challenges related to fine-tuning translations and preparing the national instruments used in IEA's large-scale comparative studies in education. The use of technology and development of tailored online translation systems can contribute here, both by guiding and streamlining the workflow and by guarding consistency. The numerous consecutive steps required to produce any national version of a large-scale international comparative assessment that strikes a good balance between faithfulness to the source and fluency in the target language are increasingly regarded as a collaborative effort between international and national experts and stakeholders rather than a system of checks and balances.

Keywords Adaptation · Centralized verification · Decentralized translation · Equivalence · Forward translation · Instrument production · Online translation system (OTS) · Reference version · Source version · Statistical review · Target version

P. Koršňáková (✉) · D. Ebbs
International Association for the Evaluation of Educational Achievement (IEA), Amsterdam, The Netherlands
e-mail: p.korsnakova@iea.nl

D. Ebbs
e-mail: d.ebbs@iea.nl

S. Dept
cApStAn Linguistic Quality Control, Brussels, Belgium
e-mail: steve.dept@capstan.be

H. Wagemaker (ed.), *Reliability and Validity of International Large-Scale Assessment*, IEA Research for Education 10, https://doi.org/10.1007/978-3-030-53081-5_6

6.1 Introduction

Producing comparable national versions of the international source instruments[1] is a key methodological issue in international comparative studies of learning outcomes such as those conducted by IEA. Once the validity and reliability of the international source version is established, procedures need be put in place to ensure linguistic equivalence of national versions with a view to collecting comparable data. As outlined in Chap. 2, the assessment landscape has changed considerably, not only in terms of the absolute numbers of participants (national and sub-national) but also in terms of their linguistic, ethnic, and cultural heterogeneity. In this chapter, our focus is on understanding the complexity of the issues as they relate to translation and the methodological response to this changing landscape.

English is a relatively concise Indo-European language with a simple grammatical and syntactic structure. A straightforward translation of a question into a language with a more complex structure may result in an increased reading load. When measuring education outcomes in, for example, mathematics or science, a fair translation should not increase reliance on reading proficiency. To achieve this level of fairness, a subtle alchemy of equivalence to the source version, fluency in the target language, and adaptations devised to maintain equivalence needs to be applied and monitored. Language components such as sentence length, quantifiers, direct speech, use of interrogatives, reference chains, active or passive voice, use of tenses, use of articles, idiomatic expressions, use of abbreviations, foreign words, or upper case may all need to be treated differently depending on the target language. For example, Slavonic languages do not use articles, in Chinese the context provides additional information, there is no upper case in Thai or in Arabic, there are more possible forms of address in Korean: all these elements have to be balanced so that comparability across language versions can be maximized.

6.2 Translation Related Developments in IEA Studies

The first notes on translation in IEA studies can be found in the results of an international research project undertaken between 1959 and 1961 on educational achievements of thirteen-year-olds in twelve countries (Foshay 1962). Although translation was of great concern for the participants, it was not the main focus of the study, so they agreed to leave to each participant the translation of the items into their own language. The most interesting feature of the recorded procedures followed in translating (mainly already existing tests originally developed in England, France, Germany, Israel, and the United States) into eight languages was the role of

[1] Instruments here refers to the test and questionnaire items, including the test and questionnaire booklet covers, introductory texts, and any scripts read by test administrators. Other important assessment materials, such as manuals and scoring guides, are reviewed as part of the international quality control measures.

a test editor, who reviewed any criticism and suggestions gathered through a pre-test (involving a small number of children in each country) and approved the test prior to its duplication and circulation, including approval of any alterations in the substance of items (such the change in units of measure to conform with the custom of the country). As expected, some difficulties in translation were found and reported, but these were "small in number and so scattered as to be insignificant" (Foshay 1962, p. 19), and there was no evidence that these would have influenced the national scores.

In the late 1960s and in 1970s, however, researchers began to realize that the operation of translating an assessment (or, as a matter of fact, any data collection instrument) into different languages, for use in different cultural contexts involved a cluster of challenges, each of which had implications on fairness and validity. Articles in the literature emerged claiming that translation and adaptation changes test difficulty to the extent that comparisons across language groups may have limited validity (Poortinga 1975, 1995). As a consequence, linguistic quality control methods were introduced in the 1970s, mostly to check the linguistic quality and appropriateness of tests translated from English and, more importantly, their semantic equivalence versus the English source version.[2]

IEA has implemented different translation procedures over time. Ferrer (2011) highlighted three careful steps: (1) comparative curricular analysis, and creation of the conceptual frameworks and specification tables; (2) cooperative procedures like collaborative item development, and multiple discussion and review stages; and (3) rigorous procedures for (back) translation of the items into the different languages of the participating countries. The last step was in line with the initial focus on validation of the translated versions of the assessment through a back translation[3] procedure (Brislin 1970, 1976). It should be noted, however, that Brislin pointed out the limitations of back translation.

With increasing interest and scrutiny in cross-national studies, IEA's Third International Mathematics and Science Study in 1995, embarking on a cyclical trend assessment that was to become known as the Trends in International Mathematics and Science Study (TIMSS), made ensuring the validity of translations a priority and commissioned in-depth research into translating achievement tests (Hambleton 1992). With Hambleton's report as a guide, TIMSS 1995 established the basis for all IEA's translation procedures. To verify the translations, TIMSS 1995 relied on multiple forward translations, translation review by bilingual judges (translation verifiers), and, due to concerns about the limitations of a back translation approach, a final statistical review. Reasons for not using back translation include the resources needed for back translation and the concern that flaws in the translation can be missed if the back translator creates a high quality English back translation from a

[2] In this chapter, the term "source version" systematically refers to the language version from which the materials are translated into other languages, resulting in "target versions." The source language is the language of the source version; the target language is the language of a specific target version.

[3] Back translation is a three-step procedure. The test is translated from English into the target language; a different translator translates that version back into English, and finally an English-speaking person compares the original test with the back-translation (see Hambleton 1992).

poor quality initial translation, thus resulting in a non-equivalent translated national version (Hambleton 1992). Thus, back translation was not used in TIMSS 1995 (Maxwell 1996) nor in any later IEA study.

6.3 Standards and Generalized Stages of Instrument Production

In IEA's technical standards (Martin et al. 1999), the task of translating and verifying translations is mentioned under standards for developing data collection instruments. In addition, when discussing standards for developing a quality assurance program, Martin et al. (1999) acknowledged that IEA studies depend on accurate translation of materials (e.g., tests, questionnaires, and manuals) from the source version (usually English) into the target languages of participating countries. Consequently, IEA makes every attempt to verify the accuracy of these translations and ensure that the survey instruments in target languages conform to the international standard, and that "no bias has been introduced into survey instruments by the translation process" (Martin et al. 1999, p. 27).

Verification of the translations of instruments into the different languages of the participating countries is seen as a measure ensuring that the translated instruments will provide comparable data across countries and cultures, or more concretely "that the meaning and difficulty level of the items have not changed from the international version" (Martin et al. 1999, p. 32). The standard is set as follows: "When translating test items or modifying them for cultural adaptation, the following must remain the same as the international version: the meaning of the question; the reading level of the text; the difficulty of the item; and the likelihood of another possible correct answer for the test item" (Martin et al. 1999, p. 43). In addition, verification is supposed to keep the cultural differences to a minimum, and retain the meaning and content of the questionnaire items through translation.

The survey operations documentation provides guidelines and other materials for the participating countries that describe the translation and cultural adaptation procedures, the process for translation verification, and serve as a means to record any deviation in vocabulary, meaning, or item layout. During translation and adaptation, participants can submit their adaptations to the international study center (ISC) for approval. Upon completion of the translation and prior to finalizing and using the instruments, countries submit their translations to the ISC for verification of the translations and adaptations of the tests and questionnaires. Professional translators also assess the overall layout of the instruments: "[t]he professional translator should compare each translated item with the international version and document the differences from the international version" (Martin et al. 1999, p. 44).

After the completion of the verification, the national center receives the feedback and addresses the deviations (e.g., incorrect ordering of response options in a

multiple-choice item, mislabeling of a graph that is essential to a solution, or an incorrect translation of a test question that renders it no longer answerable or indicates the answer to the question).

While this chapter provides a simplified overview of the process of securing high quality translations (Fig. 6.1), determining and conveying what this process actually aims to achieve is a far more complex endeavor.

In the standards, field testing is (among its other benefits) seen as a way to produce the item statistics that can detect and reveal errors in the translation and/or adaptation processes that were not corrected during the verification process and check on any flaws in the test items. Through the use of item statistics from the field test, items may be either discarded or revised and corrected, minimizing the possibility of translation errors in final versions of the test instruments.

While some problems arising from the translations of the English versions of attitudinal and value statements in the questionnaire scales were already noted in relation to the First International Mathematics Study (FIMS; Ainley et al. 2011), few further details were provided. Van de Vijver et al. (2017) conducted some further analyses using IEA data that indicate that part of the problems attributed to translations could be related to response styles and cultural differences rather than translation errors.

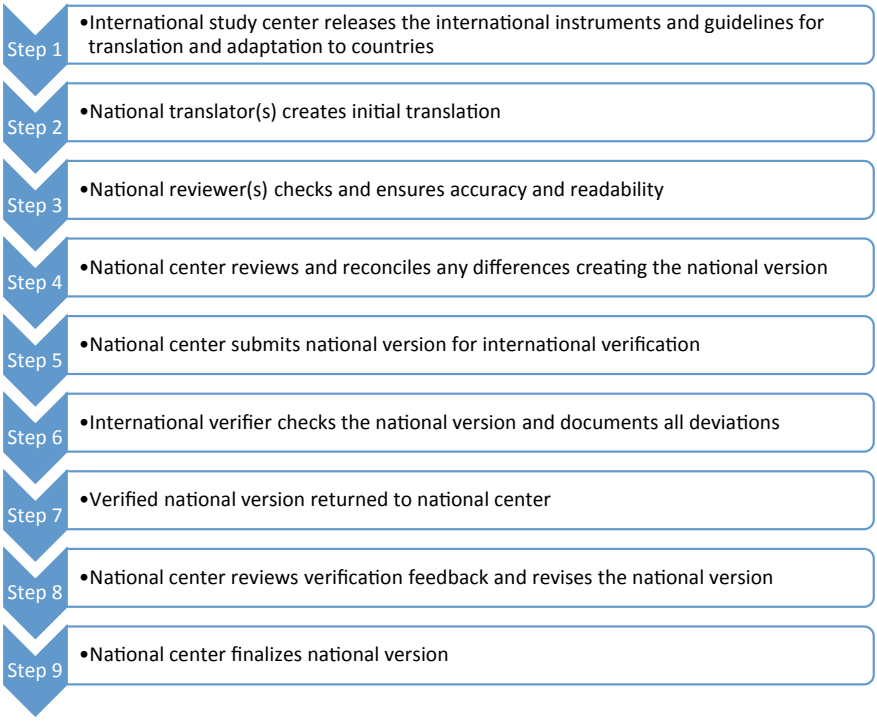


Fig. 6.1 Generalized and simplified stages of instrument production

6.4 Source Version and Reference Version

Nowadays, in international large-scale assessments (ILSAs), it is widely accepted that there is an international source version that creates a base for development of national instruments. In IEA studies, this is crafted in a collaborative effort by non-native and native speakers of the source language, which is English. By the time that this source version is ready for translation, it has been through a number of revisions, piloted, and/or gone through one or several rounds of cognitive pre-testing. By the time the source version is released for translation, it is regarded as a mature draft of the data collection instrument.

In this context, translation can be viewed as an attempt to mirror the source version in the target languages, under the assumption that the highest degree of faithfulness to the source version will be conducive to the highest degree of functional equivalence. If the quantity and quality of the information present in the source version is scrupulously echoed in the target version, the translated instrument should function the same way as the original. This, however, is not a given: Hambleton and Patsula (1999) and Hambleton (2002) described some common myths in what they refer to as test adaptation.

6.4.1 Terms Used: Translation Versus Adaptation

It should be noted that authoritative authors propose different definitions of "test translation" and "test adaptation." They share the view that the term "translation" is too narrow to capture the scope of the challenges that need to be addressed when producing multiple versions of assessment instruments while considering the objectives of functional equivalence and cross-linguistic comparability. Joldersma (2004) deemed the term "translation" too restrictive to describe the process of culturally adjusting a test rather than just translating it literally. Hambleton et al. (2005) suggested that the term "test adaptation" is preferable. Harkness (2003, 2007) and Iliescu (2017) regarded test translation as a subset of test adaptation. Iliescu (2017) explained that test translation is linguistically driven (content over intent), while test adaptation is validity-driven (intent over content), and this is certainly a workable distinction. In this chapter, however, we shall use the term translation in a broad sense and adaptation in a narrower sense, because we view the latter as an integral part of translation process.

The aim of a test translation should be to minimize the effect of language, culture, or local context on test difficulty. A straightforward translation process would not ensure this fairness. To prevent a given target population or culture being placed at an advantage or a disadvantage, it is necessary to deviate from the source version to some extent. If, for example, a general text contains references to July and August as summer months, the translator will need to consider whether, for countries in the southern hemisphere, it would be preferable to keep July and August but refer

to them as winter months; to change July and August to January and February; or to translate literally and explain in a note that this text refers to the northern hemisphere. In this light, we use the following working definition for adaptation, used by the Organisation for Economic Co-operation and Development (OECD) in its Programme for International Student Assessment (PISA):

> An adaptation is an intentional deviation from the source version(s) made for cultural reasons or to conform to local usage (OECD 2016, p. 3)

An adaptation is needed when there is a risk that respondents would be (dis)advantaged if a straightforward translation were used. While general guidelines for test translation may prescribe that each translated item should examine the same skills and invoke the same cognitive processes as the source version, while being culturally appropriate within the target country, this is a tall order. One of the myths described in Hambleton (2002) is that "translators are capable of finding flaws in a test adaptation."

No honest linguist or psychometrician can claim that the combination of a robust translation design and an expert translation verification will ensure that items examine the same skills or elicit the same cognitive processes in the source version and in the target versions. However, IEA procedures have been established to maximize comparability and ensure that the most egregious errors or misunderstandings are avoided.

In comparative assessments, cross-linguistic, cross-national, and cross-cultural equivalence is not only an objective but also a fundamental requirement without which the whole notion of quantitative cross-cultural comparison is invalidated (Dept et al. 2017). Even in translations produced by the most experienced professionals, verified by local subject matter experts, by teachers, and by trained reviewers, readers may still observe language-driven meaning shifts and/or culture-driven perception shifts. While dictionaries may provide a direct equivalent for the word "coffee" in most languages, the cultural context will lead to a different semantic loading: it is hardly possible to sip on an Italian coffee for half an hour and it would not be safe to gulp down a mug of American coffee in four seconds (Eco 2003).[4]

The concept of "mother tongue" translates as "father tongue" in some languages, and as "language of the ancestors" or "language of the fatherland" in others, with all the different connotations that this implies (Banks 2006).

Therefore, maximizing cross-language and cross-cultural comparability is a subtle balancing exercise, whereby (1) different players work together to strive for a balance between faithfulness to the source version and fluency in the target version; and (2) at the test and questionnaire design stage, it is desirable to identify concepts, terms, or contextual elements that will need to be adapted for which an intentional deviation is required or desirable to maintain equivalence, while a literal translation might jeopardize this equivalence.

[4]The example taken from Umberto Eco's *Experiences in translation* illustrates the challenge of semantic portability in general and is not related to IEA assessments.

6.4.2 *Collaborative Efforts*

In parallel with the growth in awareness as outlined in the previous section, IEA studies like TIMSS (see Fig. 6.2) and the Progress in International Reading Literacy Study (PIRLS; see Fig. 6.3) also accommodate a growing number of participating countries, and, even more importantly, a growing number of the national sets for verification, exceeding the number of participating countries. The latter shows that implementation of IEA assessments at the national level became more inclusive of different language minorities. It is important to point out that for the number of languages listed, all versions of a language (e.g., English) are combined and counted as one language, making the number of languages lower than the number of the countries involved. The different versions of a language (e.g., British English, American English, and Australian English) are accounted for in the number of verified national sets.

As in every translation procedure, the quality of the professionals who are involved in the process is one of the key determinants of a high quality result (Iliescu 2017). Considering that the majority of ILSAs (and IEA studies are no exception) have adopted a decentralized translation approach whereby national research centers are responsible for the translation of assessment instruments into their language(s) of instruction, it is important to agree on as many procedural aspects as possible to reduce disparities. Different national study centers (NSCs) may have different approaches to translation: some may outsource the translation to language service providers, while others will produce the translation in-house with more involvement of subject matter experts than professional linguists. With a view to keeping the

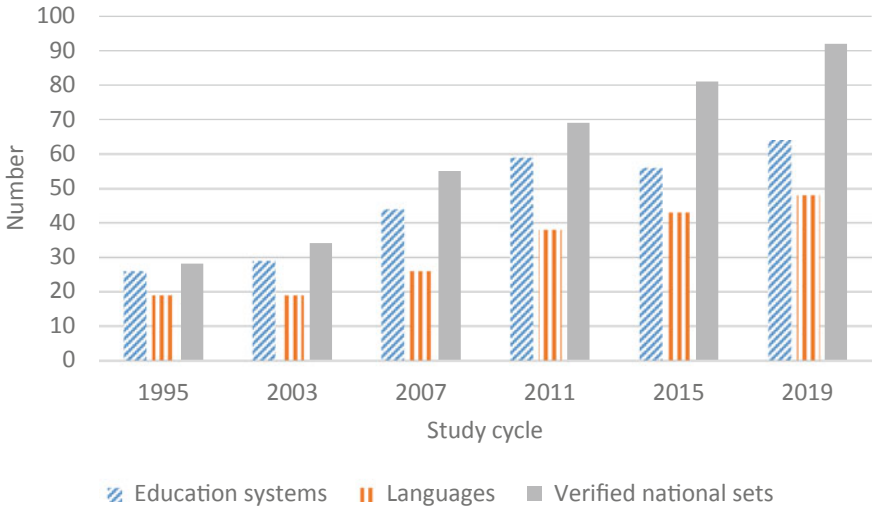


Fig. 6.2 Number of education systems, languages, and verified national sets involved in each cycle of the TIMSS grade 4 from 1995 to 2019

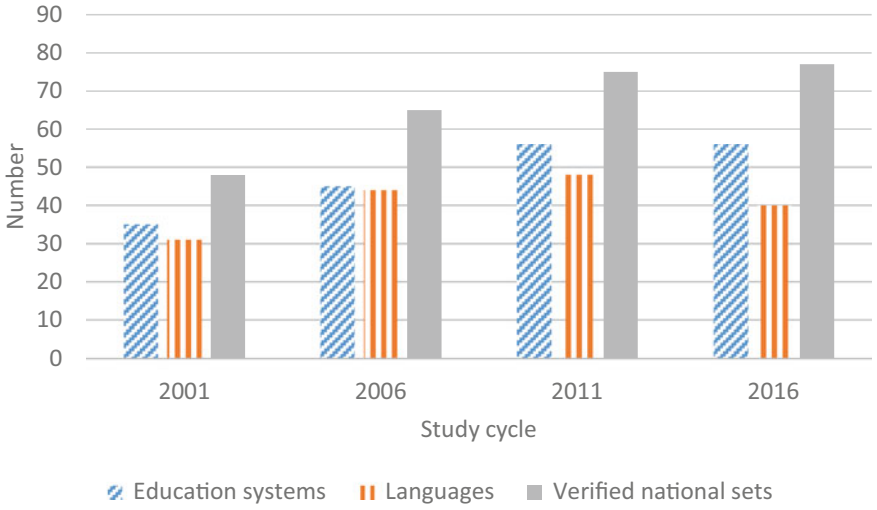


Fig. 6.3 Number of education systems, languages, and verified national sets involved in each cycle of the PIRLS from 2001 to 2016

"translator effect" in check, IEA prepares comprehensive translation and adaptation guidelines for the different instruments and tools and offers extensive technical support to NSCs during the translation and adaptation process.

While working on translations, some countries engage in collaborative efforts, like producing jointly translated instruments or sharing translations. These efforts improve quality (more faithful and fluent target versions of instruments), because such translations undergo additional reviews and the collaboration involves discussions that can reveal differences in understanding and facilitate clarification.

As the studies developed, IEA has engaged in additional efforts. In TIMSS 2007, Arabic became the largest linguistic community; IEA therefore prepared an Arabic reference version of the international instruments for Middle East and North African countries (which was based on the international source version and prepared after its release), providing Arabic-speaking countries with an initial translation of the instruments that could be easily adapted or used as a starting point for creating their national instruments. IEA oversaw and managed the collaborative process of creating the Arabic reference version in cooperation with cApStAn (an independent linguistic quality control agency in Brussels, Belgium) and staff at the TIMSS & PIRLS International Study Center at Boston College in the United States.

The process of creating the Arabic reference version began with the creation of an initial translation produced by a skilled team of translators from different Arabic-speaking countries. Following the IEA translation and adaptation guidelines, each translator produced a separate translation that a reviewer checked and compared against the other translations. The reviewer selected the best translation from the translators for use in the Arabic reference version. Upon completion of the translation, a panel of experts with experience and knowledge of school subjects at the

target grades reviewed the translation. In addition to reviewing the translation, the experts checked the consistency and correctness of terminology and commented on possible translation and adaptation issues. Based on the feedback from the experts, the translation underwent further revisions. Then the revised translation was sent to the TIMSS & PIRLS International Study Center for production of the instruments to be released to the countries (For an example of the overall translation and translation verification procedures in TIMSS 2015, please refer to Ebbs and Korsnakova 2016).

6.5 Translation and Adaptation

As indicated earlier in this chapter, we use the term translation in its broadest sense and with full awareness of the limitations of literal translations. Nevertheless, using different scripts, implementing spelling reforms, working within the grammatical constraints of the target language, and trying to achieve a subtle balance between faithfulness to the source and fluency in the target language can all reasonably be regarded as the remit of a trained professional translator. Conversely, determining which intentional deviations are acceptable, desirable, required, or ruled out should be regarded as the remit of the test authors. The latter may, of course, seek advice from cultural brokers and subject matter experts in the target culture or language. In this context, the working definition we have adopted here for adaptation (see Sect. 6.4.1) is relevant in the context of IEA assessments.

In its international studies, IEA prepares guidelines for the adaptation of test instruments and questionnaires, in which the focus is on clear prescriptions about adaptations that are required, desirable, acceptable, and/or ruled out.

The following general information is based on the guidelines for IEA studies. It aligns with the most recently published chapters on translation and verification for the completed studies (see, e.g., Malak et al. 2011; Noveanu et al. 2018; Yu and Ebbs 2012).

The two distinct steps within the production of national instruments (translation and review) at the participating countries level are designed to build up the comparability as well as linguistic quality of the national instruments. When translating (and adapting) the study instruments, NSCs are advised to pay attention to the following:

- finding words/terms and phrases in the target language that are equivalent to those in the international version;
- ensuring that the essential meaning of the text and reading level do not change;
- ensuring that the difficulty level of the items does not change;
- ensuring correspondence between text in the stem/passage and the items;
- ensuring that national adaptations are made appropriately; and
- ensuring changes in layout due to translation are minimized.

When NSCs review their translations, they are advised to use the following guidelines to evaluate the quality of their national translations:

- the translated texts should have the same register (language level and degree of formality) as the source texts; if using the same register could be perceived as inappropriate in the target culture, then the register needs to be adapted and this needs to be documented;
- the translated texts should have correct grammar and usage (e.g., subject/verb agreement, prepositions, or verb tenses);
- the translated texts should not clarify or remove text from the source text and should not add more information;
- the translated text should use equivalent social, political, and historical terminology appropriate in the target language;
- the translated texts should have equivalent qualifiers and modifiers appropriate for the target language;
- idiomatic expressions should be translated appropriately, not necessarily word for word; and
- spelling, punctuation, and capitalization in the target texts should be appropriate for the target language and the country's national context.

For assessment materials, some words or phrases might need to be adapted in order to ensure that the students are not faced with unfamiliar concepts, terms, or expressions. Common examples include reference to the working week, units of measurement, and expression of time. For questionnaires, some words and phrases require adaptation to the country specific context. To aid NSCs in identifying the words and phrases requiring adaptation, the text is placed in carets (angle brackets) in the international source version. Examples of such required (sometimes referred to as forced, obligatory, or compulsory) adaptations are < language of test >, < target grade >, and < country >.

Examples of acceptable adaptations include: fictional names of people and places that can be changed to other fictional names; measurement units that can be changed from imperial to metric or vice versa with correct conversions/numerical calculations (e.g., 3000 feet to 900 m); time notation (e.g., 2:00 p.m. to 14:00); the names of political institutions (e.g., parliament to congress) or representatives that may need to be adopted to the local context; and the names of school grades (fourth grade to year 5), programs, or education levels.

The above examples of adaptation illustrate the requirement to conform to local context and usage as necessary. However, it is useful to note that, in international assessment instruments, there should also be a requirement to deviate from the international source version each time a straightforward, correct translation is likely to put the respondent at an advantage or at a disadvantage. If a well-crafted, linguistically equivalent translation elicits different response strategies, test designers should consider adapting the translation to approach functional equivalence. For contextual questionnaires, where the notion of advantage or disadvantage does not apply, deviations from the international source version need to be considered when a straightforward translation is likely to introduce a perception shift and could affect response patterns.

In this sense, a reference to a student's boyfriend or girlfriend may become a more potent distractor in a predominantly Muslim country, for example. If this can be adapted to the student's cousin or niece without changing the information needed to respond to the question, this may be a desirable adaptation.

Likewise, if a country proposes to use a music instrument as a < country-specific wealth indicator > , there needs to be a clear assessment establishing how owning a music instrument is a socioeconomic status marker or if the term "wealth" could have been perceived as "cultural wealth" rather than "economic wealth," leading the translation to assume a different underlying construct.

6.6 Decentralized Translations and Adaptations

IEA studies have adopted the decentralized approach where the NSCs are responsible for translating and adapting the instruments into their language(s) of instruction. To aid the NSCs, the ISC always releases documents and manuals intended to guide NSCs through the processes and procedures of instrument preparation. Activities covered in the documents include:

- translating and/or adapting the study instruments;
- documenting national adaptations made to the study instruments;
- international verifications (translation, adaptation and layout); and
- finalizing the national instruments for administration.

For the process of translating and adapting the study instruments, the advice given to NSCs of early IEA studies was to have multiple translators create translation that would be consolidated into a single version by the NSC. Over the years, the advice has evolved to the use of at least one translator and one reviewer per language. The recommended criteria for the translator(s) and reviewer(s) are (see Ebbs and Friedman 2015; Ebbs and Wry 2017; Malak et al. 2011; Noveanu et al. 2018; Yu and Ebbs 2012):

- [an] excellent knowledge of English;
- [an] excellent knowledge of the target language;
- experience of the country's cultural context;
- [a familiarity with survey instruments], preferably at the level of the target grade; and, if possible,
- experience in working with students in the target grade.

The translator creates the initial national version by translating and adapting the international study instrument according to the translation and adaptation guidelines provided by the ISC. If an NSC uses more than one translator to create multiple translations of a single language version, it is the NSC's responsibility to review the translations, reconcile the differences, and produce a single version of the

instruments in that language. Upon completion of the initial national version, the reviewer proofreads and checks that the translation is of high quality, accurate, and at an appropriate level for the target population. If an NSC uses more than one reviewer, the NSC is responsible for reviewing all feedback from the reviewers and ensuring the consistent implementation of any suggestions or changes. If an NSC prepares translations in more than one language, they are advised to use professionals that are familiar with the multiple languages to ensure consistency across the national language versions. Before submitting their national language version(s) for international translation verification, the NSCs are advised to perform a final review of the language version(s) in an effort to reduce and prevent errors in the instruments that will be verified.

6.7 Centralized Verification

The international verification consists of three steps: adaptation verification, translation verification, and layout verification. The order in which these verification steps are conducted has changed over the years.

These quality control steps are centralized. For example, during the translation verification (TV) stage the ISC may: (1) entrust this step to an external linguistic quality assurance (LQA) provider; (2) perform an internal review of the feedback provided by this LQA provider; (3) send the feedback to the NSCs, who have the opportunity to review, accept, reject, or edit the LQA interventions; and (4) perform a formal check, including a layout check of the final version after TV and review of TV are completed.

Prior to TIMSS 1995, the national versions of study instruments underwent national verification procedures but did not undergo international verification procedures. The main reason was related to limited resources for conducting the studies. This resulted in a dependence on the data analysis for identifying and removing non-comparable items from the database based on discrimination and item functioning. With increased funding and requirements for verifying and ensuring the comparability and quality of the data collected, international verification procedures were put in place to support the quality and comparability of the national instruments during the field test stage and prior the main data collection.

In TIMSS 1995, the international translation verifiers conducted layout, adaptation, and translation verification at the same time (Fig. 6.4). The initial international verification procedure required the international verifiers, professional translators, to check the layout of the national instruments, followed by comparing the national translation against the international source version. If the national versions differed in any way, the translation verifiers documented the deviations. Upon completing the verification, the verifiers reported their findings to the international coordinating center, ISC, and NSCs. The NSCs reviewed the verification report and implemented

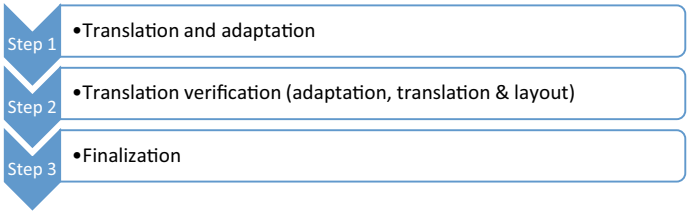


Fig. 6.4 Translation related steps in TIMSS 1995

the comments and suggestions that improved their national instruments. In addition, the verification reports were consulted during the data analysis when anomalies were found for possible translation related explanations.

With each new study and cycle, IEA's focus on ensuring the quality and comparability of the data evolved, leading to changes in the procedures. Starting with PIRLS 2006 (Malak and Trong 2007) and the Second Information Technology in Education Study (SITES) 2006 (Malak-Minkiewicz and Pelgrum 2009), the responsibility for layout verification shifted from the translation verifiers to the ISCs. This change in procedure (see Fig. 6.5) occurred to allow the translation verifiers to focus more on ensuring the quality and comparability of national translations.

Starting with TIMSS 2007 (Johansone and Malak 2008), the ISC also assumed the responsibility for adaptation verification. This change allowed the translation verifiers to further concentrate on the linguistic aspects.

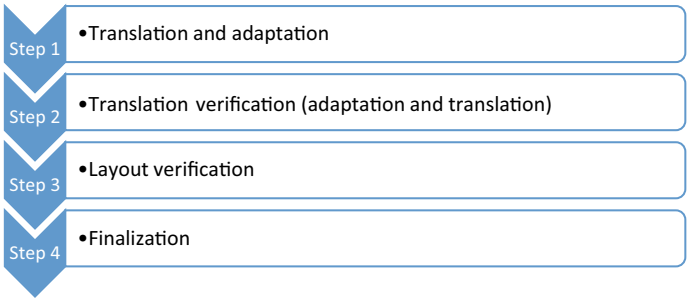


Fig. 6.5 Translation related steps in PIRLS 2006 and SITES 2006

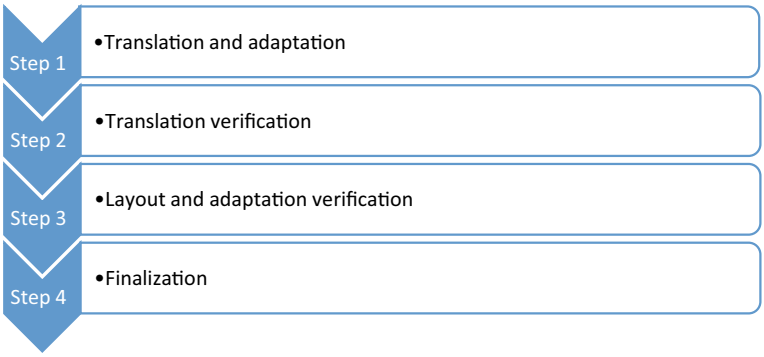


Fig. 6.6 Translation steps in TIMSS 2007

The separation of adaptation verification from translation verification led to the creation of two different pathways for the verification procedures. In the first case (e.g., as followed in TIMSS 2007; see Fig. 6.6), translation verification is conducted first (for information how this situation is handled by verifiers, please see the code 4 description in Sect. 6.8), followed by layout and adaptation verification. This option requires fewer resources and allows for a shorter timeline for completing the verification steps. One concern with this path involves situations when a national adaptation is not approved during adaptation verification and requires further changes. Upon approval of the revised adaptation, the sole responsibility for ensuring the quality of the revised translation resides with the NSC.

The second case (e.g., as followed in ICCS 2009; see Fig. 6.7) starts with adaptation verification, then translation verification, followed by layout verification. This path requires more resources and time than the first path, but ensures that all adaptations are approved and the translations revised prior to translation verification. During translation verification, the translation verifiers review the approved adaptations ensuring the correctness of the documentation and implementation of the adaptation in the national instruments.

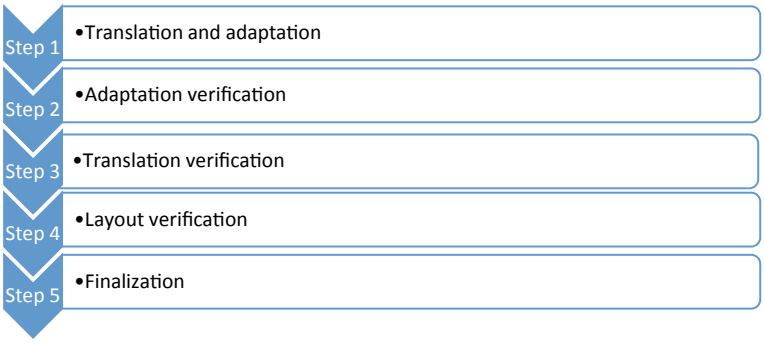


Fig. 6.7 Translation steps in ICCS 2009

6.8 Translation Verifiers

Unlike the test editors in early IEA studies, the external reviewers are linguists, not domain/content nor measurement experts, and so it is important that they judge on the linguistic aspect rather than making decisions on whether the changes that occur are appropriate.

The current IEA severity codes used by translation verifiers are:

Code 1 Major change or error: These changes could affect the results. Examples include incorrect ordering of choices in a multiple-choice item; omission of an item; or an incorrect translation that indicates the answer to the question.
Code 2 Minor change or error: These changes do not affect the results. Examples include spelling and grammar errors that do not affect comprehension; or extra spacing.
Code 3 Suggestions for alternative: The translation may be adequate, but the verifier suggests a different wording.
Code 4 Acceptable changes: Used to identify and document that national conventions have been properly documented and implemented.

If in doubt, verifiers are instructed to use Code 1? as an annotation so that the error or issue is referred to the ISC for further consultation.

If the translation verifier finds a change or error in the text while verifying the national version of the instruments, they are instructed to correct the error or add a suggestion and document the reason for the intervention. Included in the documentation of the intervention, the verifiers are to assign a code to indicate the perceived severity of the change or error corrected. A concern with the use of the severity code relates to their subjective nature. An example of this relates to the possibility that one verifier could list a grammar issue as a Code 1 error and another verifier could consider the same grammar issue to be a Code 2.

When reviewing the verifier feedback, the severity code does not inform the reader about the type of intervention performed, but does indicate the possible influence the error could have had on the item. For more information about the intervention, additional comments are added following the severity code.

IEA prepares clear and concise translation verification guidelines for the verifiers. Differences in verification style need to be kept to a minimum to avoid compounding the translator and verifier effects. It is the LQA provider's responsibility to: (1) adopt a coherent, prescriptive stance on procedures and their implementation; (2) supplement the IEA's guidelines with a face-to-face or a web-based training session for verifiers; (3) provide continuous technical and procedural support to verifiers; and (4) review the feedback provided by the verifiers, clear residual issues, and check the consistency of verifier interventions within and across instruments before returning the feedback to IEA for a second review.

The face-to-face or web-based training sessions for verifiers typically consist of:

- a general presentation on the aims of IEA studies, and on IEA's general and project-specific standards as regards translation and adaptation;
- a presentation of the materials to be verified;
- a presentation of the verifiers' tasks;
- hands-on exercises based on a selection of items from the study under verification (here a variety of errors and/or controversial adaptations may be introduced and the verifiers asked to identify the problem and assign the appropriate code); and
- information about the technical characteristics of the tools, formats, or environment in which the verification needs to be performed.

After the training, when the national version is dispatched, the package sent to verifiers includes the project-specific verification guidelines made available by IEA, a link to a recorded webinar, and a link to a resource page with step-by-step instructions on how to verify cognitive tests and questionnaires.

To further reduce disparities between commenting styles, an additional measure was implemented during the TIMSS 2019 main study translation verification. The measure required translation verifiers to select an IEA severity code and category from drop-down menus followed by using a corresponding standardized comment for each problem spotted (see Table 6.1).

Table 6.1 Examples of standardized comments in each linguistic category

Category	Example of predefined standardized comment
Formatting	Text not bolded/italicized/underlined/capitalized as in source. Changed by verifier
Grammar	Prepositional error. Corrected by verifier Syntax error. Corrected by verifier Incorrect tense. Corrected by verifier Alignment/agreement error. Corrected by verifier
Inconsistency	Translation inconsistent within item. "…" Harmonized by verifier Translation inconsistent between items. "…" Harmonized by verifier
Missing translation	Omission of text. "…" Added by verifier Untranslated text. Translated by verifier
Punctuation, symbols, and spelling	Punctuation error. Corrected by verifier Incorrect use of symbol in target. Corrected by verifier Spelling error/typo. Corrected by verifier

6.9 Layout Verification

Since differences to the layout can also affect the international comparability of the data, the ISC conducts a verification of the national instrument layout. During layout verification, the national instruments are compared to the international instruments and any discrepancies are documented. The layout verifiers check items such as the pagination, page breaks, text formats, location of graphics, order of items, and response options. All differences found are documented and need to be corrected before the national instruments are sent for printing. The goal of layout verification is to ensure minimal deviations in the comparability of the layout of national instruments. Since different languages require a different amount of space and page sizes differ in some countries, the international version of the instruments is designed with extra space in the margins to accommodate the differences in text length and page sizes. These differences are taken into consideration during layout verification. In digital assessments, the layout verification is followed by the player review and then similar duplication (of USB display instead of printing the assessment booklets) and distribution.

6.10 Development Linked to Computer-Based Assessment

As digital technologies have advanced since the millennium, the demand to use these technologies for large-scale educational assessment has increased (Walker 2017).

While IEA has been investigating the role of information and communication technology (ICT) in teaching and its use by teachers and students for a long period (see Law et al. 2008), the International Computer and Information Literacy Study (ICILS) 2013 and PIRLS 2016 were the first IEA studies to administer tests and questionnaires to students on computers. IEA contracted the development of ICILS 2013 to SoNET Systems,[5] covering the costs related to the development of computer-based assessment (CBA), as well as its use for the data collection (head counts), but the ePIRLS 2016 instruments were developed in-house by IEA. From the experience gained while using the SoNET Systems platform for ICILS 2013, IEA saw the possibilities of using an online platform for instrument production and, combined with the knowledge of the process and procedures of instrument production for international studies, IEA began development of the IEA eAssessment system. For ePIRLS 2016, development began with the translation system. The goal was to create a system that was easy to use and incorporated all the basic functions needed for translating and adapting instruments through the stages of verification. From the experiences encountered during ePIRLS 2016, further improvements to the IEA eAssessment system were implemented for TIMSS 2019.

While descriptions of the immense potential and numerous advantages of technology-based assessments abound, the transition from pencil-and-paper tests

[5] SoNET was acquired by RM Results in 2019.

to computer-delivered assessments has also given rise to considerable challenges. In the field of test translation and adaptation, this transition may have left some old school translators behind. At the same time, it has sometimes been experienced as a step backwards by professional language service providers. This is partly due to insufficient awareness of the complexity of translation/adaptation processes by system architects, who have frequently chosen to include some translation functionalities in the platform, but without giving consideration to exploring how the power of state-of-the-art translation technology could be harnessed. When discussing translation in computer-based delivery of cognitive assessment and questionnaires, Walker (2017, p. 249) stated that "at a minimum, a method should be available to replace text elements in the source/development language with translated/target equivalents."

This minimum approach fails to recognize that translation technology evolved considerably in the 1990s, and that computer-assisted translation tools (CAT tools) went mainstream by the end of the 20th century. Translation memories, term bases, spelling and grammar checkers, and style guides are functionalities that most professional translators use on a daily basis, so that, by the time international large-scale assessments transitioned to a computer environment, it had long become standard practice to use CAT tools to translate and verify pencil-and-paper tests.

So, when an e-assessment merely accommodates multilingual content but does not offer access to advanced translation technology, language service providers have to work without the tools of their trade. This often implies that achieving consistency becomes more work-intensive when producing or verifying translations in a computer-based testing platform than in word processing applications that allow the use of CAT tools.

Regardless of the authoring format, best practice in the localization industry is to separate text from layout. Translation editors do not handle layout and style (e.g., color, spacing, border thickness, or bullet formats); they handle only text. Untranslatable elements are kept separate from the translatable text, so that they cannot be altered during translation. These elements are merged with the translation at the end of the process.

Therefore, it is desirable to use adequate technology to extract translatable text from the source version. Layout markers are represented as locked tags that are easy to insert without being mishandled. Ideally, all elements that should not be translated or modified are either protected or hidden. Under these conditions, translators and verifiers can make full use of computer-aided translation tools to produce and edit text in their language. Once the text is translated and reviewed, the technology used for export can be used to seamlessly import the translation (the target version) back into the environment from which it was extracted, so that each text segment falls in the correct location.

At the same time, it is important that linguists and national reviewers can preview both the source version and their own work at any time, without file manipulation. It is necessary to preview the items in context because the translator needs to understand the task that respondents will be asked to perform. That is, the translator needs to read the entire stimulus and try to answer each question before translating it, and go

through the exercise again with the translated unit, to make sure that it functions the same way in the target language.

Content that will be used over different publication channels (webpages, mobile apps, print, etc.) should ideally be produced independently from the delivery modes, following a single-source/multi-channel publishing approach.

The Online Translation System (OTS) in the IEA eAssessment platform is a work in progress. In its initial form, it was a repository for multiple language versions of survey instruments rather than a tool to perform quality assurance routines and equivalence checks. For ePIRLS 2016, the OTS met the minimum function of allowing editing and replacement of the text with a translated version, but did not allow for the use of CATs. In addition, the OTS had basic functions for documenting adaptations and comments to segments, quick review, and resolution of translation verifier feedback, accessing previews of the international or national version of the instruments, and exporting PDF version of the national instruments.

After ePIRLS 2016, IEA made further improvements to the OTS in preparation for TIMSS 2019. A few of these improvements included the addition of a basic export/import function that would allow for the use of CATs, improved options for documenting comments and adaptations and sharing of translations. Even though the system has improved, there is room for more improvements and the interaction between platform engineers and translation technologists aims to close the gap and gradually build up functionality that will make it possible to harness the power of state-of-the-art translation technology.

6.11 Reviewing Results of Translation and Verification Processes

In our experience, quantifying the quality of a translation is difficult. The absence of comments may imply that the translation is near perfect, but it may also mean that the translation is so poor that it would make no sense to edit it. The item statistics have their own limitations, since a differential item functioning (DIF) can result from factors that are not related to translation, for example the curriculum and/or vocabulary used in textbooks, as well as their changes overtime. In addition, although item statistics may look acceptable, they do not indicate how the conveyed meaning was understood by respondents.

This can only be achieved by means of well-designed procedures that include multiple review loops aiding the focus, clarity, and fluency of the instruments. While test materials are skillfully crafted, they are not works of art, and there are usually multiple solutions to obtaining a satisfying target version. It is important that the experts involved in any stage appreciate the work already done, and build on it rather than change it according their particular language "taste." In most languages, grammar rules allow several possibilities, particularly as languages develop.

The international versions of IEA study instruments are finalized after multiple iterations that involve country and international expert views. Assuming that the experts from participating countries have voiced their concerns, the resulting materials contain concepts and words that are translatable to the target languages and cultures. Nevertheless, it remains a demanding process to create a target version that would be true to the international source, while fulfilling other important criteria, such as the appropriateness for the target grade and the country context (including curricula represented in the textbooks and taught in schools). While the translation verifiers comment on the submitted target language versions from the linguistic perspective, and the international quality control observers comment on the instruments and manuals used in a particular target language from a user perspective (including some qualitative input from the respondents as observed during the testing sessions or interviewed school coordinators and test administrators). In IEA studies, national research coordinators (NRCs) hold the responsibility of finalizing the instruments they will use when implementing an IEA study. Therefore, the major review loop starts and ends with the dedicated NRCs.

To begin with, NRCs review and comment on the international version. Once this source is finalized and released, NRCs engage and orchestrate the production of target versions by involving linguists and educators. Upon completion, NRCs submit the target versions for international verification. The translation verifiers review and comment on the submitted target versions. Afterwards, the verification feedback is returned to the NRCs for their review and finalization of the target version. Then NRCs document their experience in survey activity questionnaires that serve as a base for further development and improvement of procedures.

In IEA's trend studies, there is another challenge: trend instruments. These instruments are administered across study cycles to measure change. Ideally, trend instruments should not change unless there is an objective reason, and language experts are discouraged to make any cosmetic or preferential changes. In TIMSS 2003, trend items made up approximately 22% of the grade 4 items and 40% of the grade 8 items. In PIRLS 2006, 40% of the passages were trend. During the TIMSS and PIRLS 2011 cycle, the amount of trend items was increased to 60% and this has persisted to the TIMSS 2019 and PIRLS 2021 cycles. To ensure that the trend instruments are not changed from one cycle to the next, during international translation verification the translation verifiers compare the text of the trend instruments against the version used in the previous cycle. The translation verifiers document all differences found between the versions for review by the NRCs and ISC. A thorough documentation of any change made during the national instrument production is key to preventing unnecessary changes.

At the same time, if, for example, the currency has changed in the target country, or if a spelling reform has been implemented, it may be necessary to update the trend item/s. At this point, based on our research into trend item modification in TIMSS 2015 and the consequent DIF implication, we see a beneficial effect of the modest actions taken by NRCs in order to cope with the changes in the educational context (Ebb et al. 2017).

With the shift to the digital environment, we expect CBA development to provide a new base of information allowing more to be learned about the actions taken by NRCs in modifying the trend items (or any wording used in previous studies cycles), as well as the impact on response rate and achievement in relation to the particular items.

6.12 Procedure Chain and Timeline

IEA studies (indeed, all ILSAs) only have limited time available for their development and implementation (Fig. 6.8). The period available starts from the moment when the source version becomes available (released to participating countries) and must be finished at the moment when the target populations of students and related respondents (such as the parents of sampled students, teachers, and school principals) respond to the instruments.

The available time is shared by all key personnel engaged in the preparation of the national language(s) version(s) of assessment instruments. This includes NRCs, the translators and reviewers that work at a national level, IEA and ISC professionals, and expert translators performing the verification tasks. The reality resembles an assembly line, where any delay or misconduct in a single step affects the following step(s). This means that any time lost by a delayed submission of a language version can be covered by more support and swift accomplishment of the following step(s).

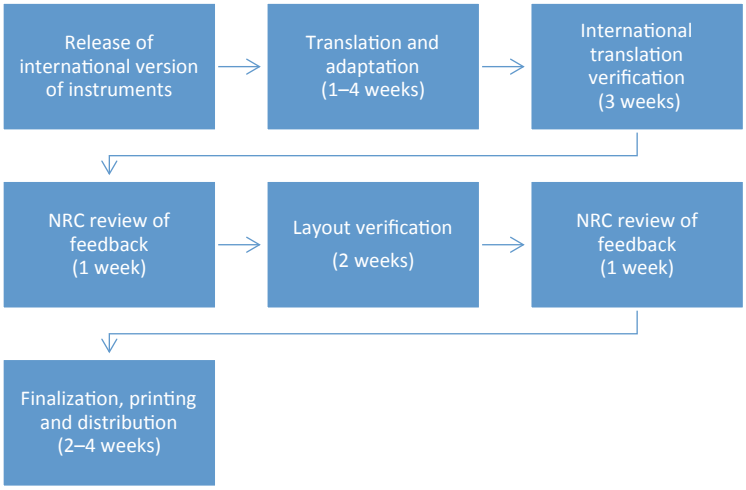


Fig. 6.8 Approximate instrument preparation timeline (based on use of printed booklets for TIMSS)

Fig. 6.9 TIMSS 2015 field test timeline

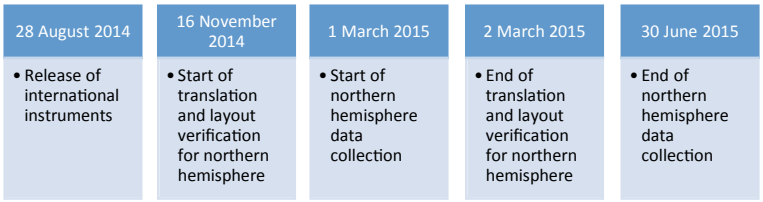


Fig. 6.10 TIMSS 2015 data collection timeline

For example, in the case of multiple national language versions of assessment instruments, these can be prepared in parallel at a country level step of translation and adaptation, and then reviewed for consistency (within and across languages) during the international translation verification. (Some examples of "real timelines" are shown in Figs. 6.9 and 6.10.)

6.13 Conclusions

Historically, participation in assessments that evaluate knowledge presupposes language proficiency, regardless of the evaluation domain. The Imperial examinations in Ancient China, a series of high-stakes tests accessible to all, were administered only in Mandarin. Acquiring advanced knowledge of the dominant language was a prerequisite toward success. A similar pattern could (and can) be observed in colonial and post-colonial states, where only the mastery of, for example, English, French, or Portuguese could open access to academic careers or high positions in the civil service: good results in admission exams taken in these colonial languages are/were regarded as a proof of competence.

In the case of IEA studies, it is of the utmost importance to diminish the impact of limitations in language proficiency that prevent participating students from demonstrating their content domain-related knowledge and their thinking skills, and hinder their engagement in responding to background and context questions expressing their own experience, attitudes, and opinions.

The review loop (see Figs. 6.1, 6.7, and 6.8), has been helpful in continuous development and improvement of the processes involved in producing the target

language instrument versions, enabling focus on the key aspects of the perceived and expected quality of the target versions of the international instruments.

The role of translation verifiers has evolved; verifiers can now concentrate on the linguistic aspects of the translations (with less need to look at layout and no need to judge the adequacy of adaptations), and, consequently, their comments have become more precise. It is also noteworthy that the numerous consecutive steps in elaborating a target version that strikes a good balance between faithfulness to the source and fluency in the target language are increasingly regarded as a collaborative effort rather than a system of checks and balances. The different players are less inclined to express their input in terms of errors, issues, or corrections, and more inclined to describe them in terms of semantic equivalence with the international source. While there is still room for improvement, it is clear that sophisticated translation, adaptation, review, and verification procedures have progressively generated a collective focus on maximizing comparability, as well as increasing reliability and validity, which has proven effective. The use of technology and development of tailored online translation systems can contribute here, by guiding and streamlining the workflow and by guarding consistency and eliciting inputs from the engaged professionals.

While engagement of particular stakeholders and professionals involved in the production of national instruments (translators, proofreaders, researchers, teachers, and administrators) is necessary, it is also important to facilitate their collaboration. In addition, cross-border collaboration has proved to be highly beneficial, since more eyes can see and eliminate more obvious mistakes, and more experience and viewpoints across different contexts can contribute to revealing incorrect assumptions and errors that would not be spotted otherwise.

Dealing with a matter as sensitive, abstract, and fluid as language benefits from a set of simple rules to deal with complexity and actions driven by common sense that all experts engaged can adopt and adhere to.

References

Ainley, J., Keeves, J., Lokan, J., Lietz, P., Masters, G., & Thomson, S. (2011). The contribution of IEA research studies to Australian education. In C. Papanastasiou, T. Plomp & E. C. Papanastasiou (Eds.), *IEA 1958–2008: 50 years of experiences and memories* (pp. 317–372). Amsterdam, the Netherlands: IEA. Retrieved from https://www.iea.nl/publications/iea-reference/iea-1958–2008.

Banks, K. (2006). A comprehensive framework for evaluating hypotheses about cultural bias in educational testing. *Applied Measurement in Education, 19*(2), 115–132.

Brislin, R. W. (1970). Back translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*(3), 185–216.

Brislin, R. W. (1976). Comparative research methodology: Cross-cultural studies. *International Journal of Psychology, 11*(3), 215–229.

Dept, S., Ferrari, A., & Halleux, B. (2017). Translation and cultural appropriateness of survey material in large-scale assessments. In P. Lietz, J. Cresswell, K. Rust, & R. Adams (Eds.), *Implementation of large-scale education assessments* (pp. 168–191). Chichester, UK: Wiley.

Ebbs, D., & Friedman, T. (2015). Translation and verification of ICILS 2013 instruments. In J. Fraillon, W. Schulz, T. Friedman, J. Ainley & E. Gebhardt (Eds.), *ICILS 2013 technical report*

(pp. 55–66). Amsterdam, the Netherlands: IEA. Retrieved from https://www.iea.nl/publications/technical-reports/icils-2013-technical-report.

Ebbs, D., & Korsnakova, P. (2016). Translation and translation verification for TIMSS 2015. In M. O. Martin, I. V. S. Mullis & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 7.1–7.16). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from https://www.iea.nl/publications/technical-reports/methods-and-procedures-timss-2015.

Ebbs, D., & Wry, E. (2017). Translation and layout verification for PIRLS 2016. In M. O. Martin, I. V. S. Mullis & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 7.1–7.15). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from https://www.iea.nl/publications/study-reports/international-reports-iea-studies/methods-and-procedures-pirls-2016.

Ebbs, D., Desa, D., & Korsnakova, P. (2017). DIF study on the effects of item modifications based on the TIMSS trend items. Poster presentation. In *Program 7th IEA International Research Conference 2017, 28–30 June 2017, Faculty of Education, Charles University, Prague, Czech Republic* (p. 73). Amsterdam, the Netherlands: IEA. Retrieved from https://www.iea.nl/sites/default/files/2019-04/IRC%202017%20Program.pdf.

Eco, U. (2003). *Dire quasi la stessa cosa. Esperienze di traduzione.* Milan, Italy: Biompani. English translation: Eco, U. (2008). *Experiences in translation*, Toronto, Canada: University of Toronto Press.

Ferrer, A. T. (2011). Experiencing the world as an educational laboratory. In C. Papanastasiou, T. Plomp, & E. C. Papanastasiou (Eds.), *IEA 1958–2008: 50 years of experiences and memories.* Nicosia, Cyprus: Cultural Center of the Kykkos Monastery. https://www.iea.nl/publications/iea-reference/iea-1958-2008.

Foshay, A. W. (1962). The background and the procedures of the Twelve-Country Study. In A. W. Foshay, R. L. Thorndike, F. Hotyat, D. A. Pidgeon, & D. A. Walker (Eds.), *Educational achievements in thirteen-year-olds in twelve countries* (pp. 7–19). Hamburg, Germany: UNESCO Institute of Education, UNESCO. https://www.iea.nl/publications/publications/educational-achievements-thirteen-year-olds-twelve-countries.

Hambleton, R. K. (1992). *Translating achievement tests for use in cross-national studies* (Doc. Ref.: ICC454/NRC127). Report prepared for IEA, New York, NY, and NCES, Washington, DC, Laboratory of Psychometric and Evaluative Research Report No. 241, University of Massachusetts, School of Education, Amherst, MA. https://files.eric.ed.gov/fulltext/ED358128.pdf.

Hambleton, R. K. (2002). Adapting achievement tests into multiple languages for international assessments. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross national surveys of educational achievement* (pp. 58–79). Washington, DC: National Academy Press.

Hambleton, R. K., & Patsula, L. N. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1*(1), 1–13. Retrieved from http://www.jattjournal.com/index.php/atp/article/view/48345.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 33–56). Hoboken, NJ: Wiley.

Harkness, J. A. (2007). Improving the comparability of translations. In R. Jowell, C. Roberts, R. Fitzgerald, & G. Eva (Eds.), *Measuring attitudes cross-nationally: Lessons from the European Social Survey* (pp. 79–93). Los Angeles, CA: Sage.

Iliescu, D. (2017). *Adapting tests in linguistic and cultural situations.* New York, NY: Cambridge University Press.

Johansone, I., & Malak, B. (2008). Translation and national adaptations of the TIMSS 2007 assessment and questionnaires. In J. F. Olson, M. O. Martin, & I. V. S. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 63–75). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from https://www.iea.nl/publications/technical-reports/timss-2007-technical-report.

Joldersma, K. J. (2004). *Cross-linguistic instrument comparability*. Unpublished manuscript, Michigan State University, East Lansing, MI. Retrieved from https://education.msu.edu/cepse/mqm/documents/KJ.pdf.

Law, N., Pelgrum, W., & Plomp, T. (Eds.). (2008). *Pedagogy and ICT use in schools around the world. Findings from the IEA SITES 2006 Study.* Hong Kong: CERC-Springer. Retrieved from https://www.iea.nl/publications/study-reports/international-reports-iea-studies/pedagogy-and-ict-use-schools-around.

Malak, B., & Trong, K. L. (2007) Translation and translation verification of the PIRLS reading assessment and questionnaires. In M. O. Martin, I. V. S. Mullis, & A. M. Kennedy (Eds.), *PIRLS 2006 technical report* (pp. 49–60). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from https://www.iea.nl/publications/technical-reports/pirls-2006-technical-report.

Malak, B., Yu, A., Schulz, W., & Friedman, T. (2011). Translation and national adaptations of ICCS 2009 instruments. In W. Schulz, J. Ainley, & J. Fraillon (Eds.), *ICCS 2009 technical report* (pp. 51–58). Amsterdam, the Netherlands: IEA. Retrieved from https://www.iea.nl/publications/technical-reports/iccs-2009-technical-report.

Malak-Minkiewicz, B., & Pelgrum, W. J. (2009). Translation, national adaptation, and verification. In R. Carstens & W. J. Pelgrum (Eds.), Second Information Technology in Education Study: SITES 2006 technical report (pp. 41–46). Amsterdam, the Netherlands: IEA. Retrieved from https://www.iea.nl/publications/technical-reports/second-information-technology-education-study-technical-report.

Martin, M. O., Rust, K., & Adams, R. J. (Eds.). (1999). *Technical standards for IEA studies.* Amsterdam, the Netherlands: IEA. Retrieved from https://www.iea.nl/publications/iea-reference/technical-standards-iea-studies.

Maxwell, B. (1996). Translation and cultural adaptation of the survey instruments. In M. O. Martin & D. L. Kelly (Eds.), *Third International Mathematics and Science Study (TIMSS) technical report. Volume I: Design and development* (pp. 8.1–8.10). Chestnut Hill, MA: Boston College. Retrieved from https://www.iea.nl/publications/publications/third-international-mathematics-and-science-study-technical-report-volume.

Noveanu, G. N., Friedman, T., & Köhler, H. (2018). Translation and national adaptations of ICCS instruments. In W. Schulz, R. Carstens, B. Losito, & J. Fraillon (Eds.), *ICCS 2016 technical report* (pp. 33–41). Amsterdam, the Netherlands: IEA. Retrieved from https://www.iea.nl/publications/technical-reports/iccs-2016-technical-report.

OECD. (2016). *PISA 2018 translation and adaptation guidelines.* National Project Managers' Meeting, Bangkok, Thailand, March 2013. First Meeting of the PISA 2018 National Project Managers, 14–18 March 2016, Prague, Czech Republic. Paris, France: OECD. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2018-TRANSLATION-AND-ADAPTATION-GUIDELINES.pdf.

Poortinga, Y. H. (1975). Limitations on international comparison of psychological data. *Nederlandse Tijdschrift voor de Psychologie, 30,* 23–29.

Poortinga, Y. H. (1995). Cultural bias in assessment: Historical and thematic issues. *European Journal of Psychological Assessment, 11*(3), 140–146.

Van de Vijver, F., He, J., & Kulikova, A. (2017). Response styles in cross-cultural perspective: TIMSS and PIRLS. In *Keynote presented at the 7th IEA International Research Conference 2017, Prague, Czech Republic [webpage].* Amsterdam, The Netherland: IEA. Retrieved from https://www.iea.nl/index.php/publications/keynotes/response-styles-cross-cultural-perspective.

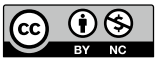
Walker, M. (2017). Computer-based delivery of cognitive assessment and questionnaires. In P. Lietz, J. Cresswell, K. Rust, & R. Adams (Eds.), *Implementation of large-scale education assessments* (pp. 231–252). Chichester, UK: Wiley.

Yu, A., & Ebbs, D. (2012). Translation and translation verification. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011* (pp. 1–13). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from https://www.iea.nl/publications/technical-reports/methods-and-procedures-timss-and-pirls-2011.

Paulína Koršňáková has a background in natural science and mathematics education and holds a PhD in psychology. Prior to joining IEA, Dr. Koršňáková coordinated implementation of multiple IEA and OECD studies in Slovakia. The national adaptation and translation processes were part of the tasks she oversaw and carried on. Since 2010, Dr. Koršňáková worked on IEA ICILS 2013 and OECD TALIS 2013 studies at the international level, and she has overseen or directly managed translation verification processes for multiple other studies at the international level.

Steve Dept is one of the founders of cApStAn Linguistic Quality Control. He received his education in English, Dutch, French and German. He studied Classical Philology and comparative linguistics but is essentially an autodidact and a field practitioner. In 1998, he was commissioned to organize the translation review of PISA 2000 instruments in 24 languages. Within cApStAn, Steve has coordinated linguistic quality control (LQC) of PISA instruments across 8 PISA survey cycles. Other multilingual surveys or assessments for which LQC operations were supervised by Steve include TIMSS, PIRLS, PIAAC, UNESCO/LAMP, ICILS, and TALIS. Over the years, Steve's attention has shifted towards embedding linguistic quality assurance into the test development process with a view to reducing post hoc corrective action. His translatability assessment methodology has become an integral part of the adaptation design of major international large-scale assessments and multilingual surveys.

David Ebbs is a Senior Research Officer at IEA, overseeing the international translation verification process. He has used his considerable experience with IEA processes and procedures to help in the development of the IEA eAssessment system. He has a Master of Education in Educational Leadership and a Bachelor degree in Biology.



Chapter 7
Sampling, Weighting, and Variance Estimation

Sabine Meinck

Abstract The International Association for the Evaluation of Educational Achievement (IEA) undertakes international large-scale assessments (ILSAs) to provide reliable measures of student achievement in the context of learning. These ILSAs can be used to assess the quality and equity of education systems and enable countries to make informed decisions for improving education. Surveying all individuals belonging to the target population (e.g., students, classes, or teachers) would be a costly enterprise. The ILSA design instead focuses on measuring high-quality representative samples, ensuring the results are close to the true values in the populations and can be usefully compared across cultures and countries. Sampling is thus a key element in establishing the validity and reliability of the results obtained from cross-national large-scale assessments. This chapter reviews IEA's sampling strategies, from defining the target populations, to compiling sampling frames, applying complex sampling methods, and accounting for the methodology when analyzing data. These concepts are explained for a non-technical audience in an illustrative way. Common methodologies used when selecting random samples for IEA studies include stratification, multiple-stage and cluster sampling, and unequal selection probabilities. The samples yielded by implementing any one of these methods call for specific methods of data analysis that take account of the sampling design. Sampling weights have to be applied to arrive at accurate conclusions on population features when using sample data; specific methods are needed to compute standard errors and confidence intervals correctly. This comprehensive guide to IEA's methodology illustrates the vital importance of sampling in providing high-quality, reliable data that are generalizable to the targeted populations in an unbiased manner.

Keywords Cluster sampling · Complex samples · Multiple-stage sampling · Nonresponse · Standard errors · Target population · Variance estimation · Weighting

S. Meinck (✉)
International Association for the Evaluation of Educational Achievement (IEA), Hamburg, Germany
e-mail: sabine.meinck@iea-hamburg.de

H. Wagemaker (ed.), *Reliability and Validity of International Large-Scale Assessment*, IEA Research for Education 10, https://doi.org/10.1007/978-3-030-53081-5_7

7.1 Introduction

The studies of IEA, like any other international large-scale assessments (ILSAs) of education, aim to provide a precise (and valid) picture of the state of education systems at a defined point in time for a particular target group and domain (e.g., mathematics achievement of grade 4 students). In doing so, IEA needs to satisfy the twin imperatives of international comparability (fair, valid, and reliable) and national relevance.

Assessing all individuals belonging to the target population would generally be too costly, which is why ILSAs are applied to selected representative samples instead. Based on these samples, researchers approximate the population features of interest. A sample helps in reducing the workload, respondent burden, and costs while providing estimates that are close enough to values from a complete census to meet the intended purposes. This means that the sample must be selected in a way that ensures that it will represent the targeted population in a precise and undistorted manner. The expressions "design-unbiasedness" and "sampling precision" are often used to summarize those characteristics of a statistical sampling design (Dumais and Gough 2012a). Therefore, sampling is key to ensuring the validity and reliability of ILSAs. Sampling strategies developed for IEA studies ensure that: (1) cost and quality requirements are balanced; (2) population estimates are close to what a complete census would have given and precise enough for the intended purposes; (3) they can be applied in a variety of educational systems; and (4) they allow for valid cross-national comparison of the results.

All samples selected for IEA studies are random samples. Common features applied are stratification, multiple-stage sampling, cluster sampling, and sampling with unequal selection probabilities. Samples yielded by implementing any one of these methods are called "complex samples." In most ILSAs, these methods are part of the international sampling design, applied in most, but not all countries, and more than one method can be used. Typically, the international sampling design is optimized with respect to the specific circumstances of a particular country, while complying with the international objectives of design-unbiasedness and sampling precision. Importantly, samples are selected in a way that replicates the structure of the educational system, allowing data to be linked across schools, classes, students, and teachers.

Like any other survey based on random sampling, IEA studies estimate two parameters for each characteristic of interest: its point value (e.g., the years teachers spent on average in the profession, the proportion of students who have access to a school library, the average science score of grade 8 students, the difference in achievement between boys and girls, regression and correlation coefficients) and its precision (e.g., a margin of error for the estimated science score of grade 8 students). Both measures are affected by the complex sampling design, and addressed by specific weighting and variance estimation procedures. After reviewing how target populations are defined in IEA studies, this chapter illustrates the IEA's sampling strategies and weighting procedures, and introduces approaches for estimating population

characteristics and their standard errors. The chapter focuses on how these strategies are related to the reliability and validity of the survey results, and also covers related quality control measures.

7.2 Defining Target Populations

When looking at the study results, readers will intuitively assume they describe features of the whole target population. For example, readers of the Progress in International Reading Literacy Study (PIRLS; see IEA 2020) reports will suppose a particular average pertains to all grade 4 students in a given country. Furthermore, to compare features across different nations or educational systems, equivalent items have to be compared. IEA studies make great efforts to ensure valid comparisons are possible.

At first glance, it may seem a simple task to define the target population of a study. Looking at some exemplary IEA target population definitions, however, it becomes obvious that the details are significant. For instance, the population definitions of students often rely on the internationally accepted International Standard Classification of Education (ISCED) scheme to describe levels of schooling across countries (UIS [UNESCO Institute for Statistics] 2012) and determine the correct target grade in each country. For example, IEA's PIRLS assesses:

> All students enrolled in the grade that represents four years of schooling counting from the first year of ISCED Level 1, providing the mean age at the time of testing is at least 9.5 years. (Martin et al. 2017, p. 3.3)

Reference to the ISCED scheme overcomes the challenge of differently structured educational systems, with different names for grades and different school entry rules (for example, to distinguish pre-primary from primary classes). PIRLS specifies a minimum average age for the cohort of interest, acknowledging that school entry occurs at different ages in different countries and the assessment may pose excessive cognitive demands on very young children. Some educational systems (for example, Northern Ireland and England) have a very early school entry age and therefore test students at grade 5 instead of grade 4 so that their national definition of the target population deviates from the international definition. Compromises are unavoidable when accommodating cross-national comparisons, but also pose a threat to the validity of such comparisons. Not all targeted children have received the same number of years of education, and/or the average age of children within the target population may differ by country, which may affect the maturity of test takers. Nonetheless, the national cohort that is closest to the internationally defined target population represents the optimum choice for cross-country comparisons.

Defining other than student populations is often even more difficult. For example, determining a cross-nationally comparable population of teachers poses significant challenges. This entails detailed wording and operationalized definitions to enable

national teams to identify the individuals targeted by the survey correctly. The International Civic and Citizenship Education Study (ICCS; see IEA 2020) and the International Computer and Information Literacy Study (ICILS; see IEA 2020) both target teachers of grade 8 students. Each study defines precisely which target teachers are to be studied and provides detailed criteria for establishing whether the target groups are teaching grade 8 students (Meinck 2015b; Weber 2018). The definitions are substantiated by practical examples. IEA's sampling experts are available throughout recruitment periods to help national teams interpret the definitions correctly and advise on particular situations.

Another challenge regarding comparability is that not all eligible individuals[1] can be covered by the studies in all countries. Reasons to remove parts of the internationally defined target population usually relate to undue collection costs and/or a lack of fit between the assessment and the abilities of the test takers. Relatively high collection costs occur, for example, for remote or small schools, which is why countries are usually allowed to exclude such schools from the assessment. If clearly distinguishable parts of target populations are removed, such as, for example, minority language groups, this is usually referred to as "reduced coverage." In addition, students with disabilities or those lacking the requisite language skills can be released from the test. As a general guideline, students with physical disabilities, such that they cannot perform in the study test situation,[2] or students with intellectual disabilities, unable to follow even the general instructions of the test, can be excluded (Martin et al. 2016, 2017; Meinck 2015b; Weber 2018). It is advisable that students who are non-native language speakers (i.e., students who are unable to read or speak the language(s) of the test and would be unable to overcome the language barrier in the test situation) are also excluded. Typically, this affects students who have received less than one year of instruction in the language(s) of the test. In all IEA studies, the reasons and scope of exclusions and under-coverage are meticulously examined and documented. To minimize the potential risk of bias, exclusions from the core target population (mostly students) must not exceed five percent of the target population. Countries surpassing this threshold are annotated in every single table in the international report, clearly signaling potential doubts on the comparability of their results to the readers. If the national target population definitions are altered between cycles, trends over time are not reported, or are reported based on homogeneous population parts. An issue of concern is the increasing number of students being excluded because of disabilities in recent study cycles in a number of countries, as can be seen from comparing respective technical documentation over time. For example, the average total exclusion rate in PIRLS increased from 3.8% in 2001 to 5.8% in 2016, including all countries in the analysis who participated in both cycles (Martin et al. 2003, 2017). This may be related to reforms initiated after the United Nations (2019) enacted the Convention on the Rights of Persons with Disabilities, which entered into force in

[1] Eligible individuals are all those belonging to the internationally defined target population.

[2] Students that cannot be accommodated by the test situation or where accommodation cannot be provided as part of normal practice can be excluded.

2008, instigating new procedures on diagnosing and treating students with disabilities. IEA continually reviews this and related developments to ensure comparability is not jeopardized.

Given such challenges, readers of IEA study results are strongly advised to consult the encyclopedias and technical documentation that describe the school systems of each participating country, the chosen target grades and average ages, and information on excluded or not covered parts of the populations. This contextual information needs to be carefully considered when interpreting the results.

To ensure countries identify the suitable target populations in their country, and specify their exclusions correctly, national research coordinators (NRCs), who are responsible for the implementation of a study in a particular country, complete a set of forms. Among other tasks, NRCs identify the target grade, the country's name for the grade, the average age of students in that grade at the time of data collection, and the type and scope of exclusions or reduced coverage. This information is validated by the IEA sampling experts, using information from earlier cycles or other studies, or/and education statistics from reliable online sources, such as the World Bank (World Bank 2020), UNESCO Institute for Statistics (UIS 2020), the Organisation for Economic Co-operation and Development (OECD 2020), Eurydice (European Commission 2020), and national statistical agencies.

7.3 Preparing Valid Sampling Frames for Each Sampling Stage

A sampling frame is a list of units from which the sample is selected. Samples can only be representative of the units enlisted on the frame. Therefore, sampling frames in IEA studies must be comprehensive lists of all eligible units belonging to the target population. As described in Sect. 7.4.1, many IEA studies involve multiple sampling stages, which implies that valid sampling frames have to be compiled for each selection stage. It is critical that all NRCs are able to identify the correct units and list them on the sampling frame. Therefore, detailed definitions are not only needed for the target populations but also for every step of the selection process.

Most often, the first selection step entails a school sample, in which case all schools offering education to the targeted students must be listed on the sampling frame. This list must include a national identification number, allowing staff at the national center to identify the sampled schools correctly. If sampling with probabilities proportional to size (PPS; see Sect. 7.4.3) is to be employed, a measure of school size, such as the number of students or classes, has to be included. Further, variables grouping homogeneous schools together (a process called stratification; see Sect. 7.4.2) need to be added if this design feature is deemed useful for increasing sampling precision or addressing research questions of particular policy relevance. Variables such as addresses and contact information are not needed for sampling, but could be helpful for validating the frame or simplifying the correct identification of schools later on in the process. A school sampling frame is rarely fully up-to-date, as school statistics are usually available only with some delay. However, significant efforts

must be made to make them as accurate as possible. Missing out schools on the frame will lead to an undercoverage bias in the population (i.e., the sample is not representative of, for example, newly opened schools). Carrying ineligible schools in the frame, however (e.g., those that recently closed or no longer teach students in the target grade), can lead to decreasing sample sizes, as they have to be dropped from the sample without replacement when being sampled. Sampling experts check all school sampling frames rigorously for consistency and plausibility. Checking routines include the search for duplicates, and comparison of frame information with information on sampling forms, estimated population sizes from earlier cycles or other studies, and officially available data, such as enrollment figures and birth statistics from reliable online sources (as mentioned in Sect. 7.2).

Deciding upon the unit to be listed on the school frame can be sometimes more difficult than expected. Options can be to list physical buildings, administrative units, or even tracks within units. In some countries, schools as administrative units are not tied to a building, but rather comprise a set of buildings or even locations (sometimes called main and satellite schools). In such a case, it is methodologically sound to list school buildings separately or as complete administrative units. Importantly, all persons involved in the study need to be informed about this decision, as this affects not only the compilation of valid sampling frames for the second sampling stage but also the validity of questionnaire responses. The school coordinator must list the respective target grade classes to prepare for proper class sampling, whether this is just one school building or the complete administrative set. In addition, the principal of a sampled school who is asked to provide information about the school needs to answer the school questionnaire with respect to the sampled unit. Incomplete or imprecise information can lead to noisy data or even severe bias. For example, if school coordinators incorrectly listed only classes based in the school's main building, even though the whole administrative unit was sampled, then classes from satellite buildings (often located in rural areas) would be omitted. The latter would then have a zero chance of being selected and thus would not be represented by any sample selected from the frame. Alternatively, if the satellite of a school, located on a remote island, is sampled and the principal of this school (i.e., the person completing the school questionnaire) has their office at the main school building, their response may instead reflect the whole administrative unit if they are not informed about the location of the sampled unit. This may lead to potentially inaccurate or misleading answers regarding many of the sampled school's features, such as socioeconomic intake, resources, or location. A last pertinent example relates to schools organized in shifts, a situation that is often present in developing countries. In such schools, a school building is used for two, or even three different shifts (morning, afternoon and evening shifts), each with completely distinct student and (often) teacher populations. Again, a decision needs to be made on whether shifts or actual school buildings are listed on the sampling frame, and all people involved in the study (i.e., school coordinators, principals, and test administrators) must be accordingly informed to avoid biased results and thereby threats to the validity of the collected data.

Similar constraints can occur when compiling sampling frames for subsequent selection stages. When sampling classes, a key requirement is that every eligible

student in a school must belong to exactly one, and only one class. In other words, classes must contain mutually exclusive and exhaustive groups of students. If this requirement is violated, students may have zero or multiple selection chances, which would lead to biased results. Further, it is key that all eligible classes (i.e., those containing eligible students) must be listed. Classes exclusively dedicated to students that should be excluded from the test must be marked as such, and will not be sampled. Finally, a complete list of students is needed for each sampled class. With this list, a sample of students within the class can be selected as a last sampling step, or all students in the class invited to participate in the survey.

Listing and sampling classes and students in IEA studies is performed by national staff, using a dedicated software made available by IEA to support participating countries, IEA's Windows Within School Sampling Software (WinW3S). Sampling experts provide comprehensive documentation on listing procedures and train national staff on correct implementation of class and student listing and sampling. IEA's sampling experts are available for help whenever questions arise, a measure that has proved to be highly effective at ensuring the high quality of class and student sampling frames and procedures.

It should be noted that IEA ensures full confidentiality of the information received with the sampling frames. School sampling frames are kept fully confidential; sensitive information about classes, students, and teachers, such as names or addresses, are automatically removed from the WinW3S database when it is submitted by the national center to IEA. Schools, classes, and individuals cannot be identified in the publicly available databases after the survey.

7.4 Sampling Strategies and Sampling Precision

IEA studies rely on sampling strategies recognized as state-of-the art by the scientific community, which are now implemented in most contemporary ILSAs (Rutkowski et al. 2014). The chosen sampling strategies are well described in the technical documentation accompanying all ILSAs and comprehensive introductions to the topic are readily available (for example, Dumais and Gough 2012a; Lohr 1999; Rust 2014; Statistics Canada 2003). Hence, rather than reviewing the sampling approaches used in IEA studies, this chapter focuses on certain aspects directly related to the reliability and validity of the results.

IEA studies implement mostly so-called complex samples, that is, samples that employ at least one of the following features: multiple stage sampling, cluster sampling, stratification, and sampling with unequal selection probabilities. All of these features have a direct or indirect effect on sampling precision (i.e., how close the estimated values are to the true values in the population) or, in other words, how reliable and valid the statements are, based on survey data. The concept of "sampling precision" was neatly illustrated by Meinck (2015b), using the famous photograph of

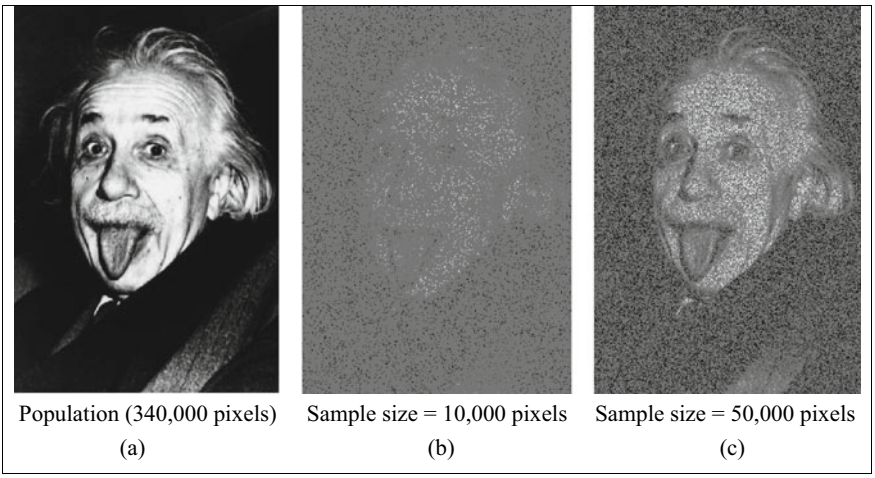


Fig. 7.1 Illustration of sampling precision: simple random sampling. *Source* Meinck (2015b) Copyright © 2015 International Association for the Evaluation of Educational Achievement (IEA)

Einstein taken by Arthur Sasse in 1951. Selecting pixels from a picture and reassembling the picture based only on the sampled pixels shows that the picture gets more precise as the sample size (here the number of pixels) increases (Fig. 7.1).

In educational studies, with students or schools as sampling units, it is not as easy to visualize the effect of sampling precision. Furthermore, most ILSA do not employ simple random sampling (SRS) methods. IEA studies have established precision requirements that allow highly precise estimates for large populations and large population subsets. ICCS and ICILS employ sampling strategies yielding samples equivalent to a simple random sample of 400 sampling units, a precision level deemed sufficient in most ILSAs. Samples with this precision yield percentage point estimates with a confidence interval of ± 5 percentage points (Meinck 2015b; Weber 2018). The Trends in International Mathematics and Science Study (TIMSS; see IEA 2020) and PIRLS sample size requirements are equivalent to a simple random sample of 800 sampling units to permit precise trend measurement (Martin et al. 2016, 2017). However, why and how do sampling strategies affect sampling precision?

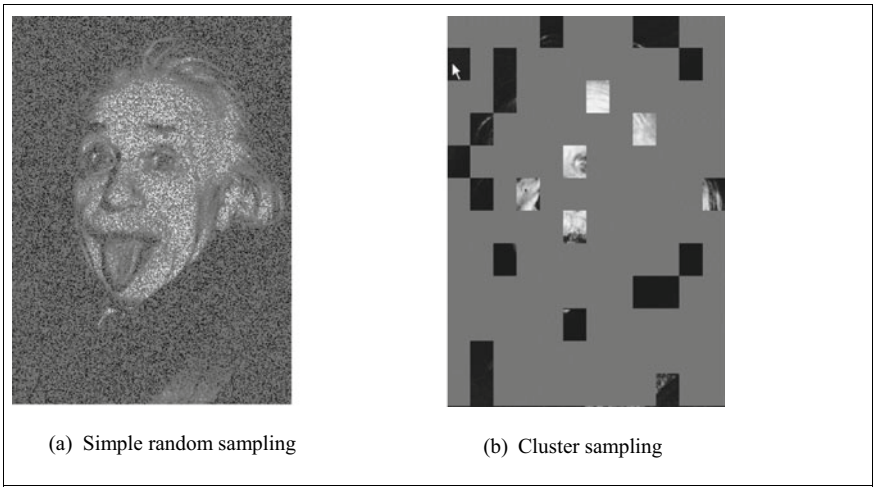
7.4.1 Multiple Stage Sampling and Cluster Sampling

In most IEA studies, samples are selected in multiple stages. Schools are usually selected first; they comprise the so-called "primary sampling units" (PSUs). In a second step, one or multiple classes of the target grade are sampled in TIMSS, PIRLS, and ICCS, while students from across all classrooms of the target grade are selected in ICILS. The approach of selecting multiple units within a previously selected group is called cluster sampling. The method has several advantages: collection costs are

low and the method enables researchers to connect different target populations (e.g., students with their peers, teachers, and schools), and thereby examine the contexts of learning, yielding additional analytical power and possibilities. Comprehensive lists of students are not available in many countries, while lists of schools exist; hence, the process of compiling sampling frames is simplified. However, a significant downside to this approach is that sampling precision is reduced. The effect occurs if elements within a cluster are more similar than elements from different clusters. For example, students attending the same school share the same learning environment, or, similarly, students within a class are instructed by the same teacher, and these students will therefore be more likely to exhibit similar learning outcomes. Studies combining school and class sampling thus suffer from a double cluster effect. A simple comparison of SRS with cluster sampling can be made by using the different sampling methods to select same number of pixels from a picture. This illustrates neatly that the sample derived by SRS provides a far better representation of the original picture (Fig. 7.1a), even though the same number of pixels is displayed (Fig. 7.2).

The cluster effect can differ for any variable of interest, and vary among countries and over time. Its statistical measure is the intraclass correlation coefficient (ICC; Kish 1965). As a very prominent example, the socioeconomic composition of students in schools has a relatively large effect on the ICC of achievement in many countries. This is, variance between schools in student achievement is strongly associated with the "average" socioeconomic status of the students within schools (e.g., Sirin 2005).

The effect of ICC in ILSA is often so large that the required sample sizes actually need to be approximately ten times greater than required for SRS to achieve the same



(a) Simple random sampling (b) Cluster sampling

Fig. 7.2 Sampling precision with equal sample sizes of 50,000 pixels. The results of **a** simple random sampling, and **b** cluster sampling *Source* Meinck (2015b) Copyright © 2015 International Association for the Evaluation of Educational Achievement (IEA)

level of sampling precision. This is why IEA studies that aim for a sample equivalent to an SRS of 400 (ICCS and ICILS) or 800 (TIMSS and PIRLS) units, in fact require about 3000 to 4000 units. While a minimum sample size is specified, countries are able to adjust this upwards to meet the requirements for generating estimates for subpopulations that may be of particular policy interest.

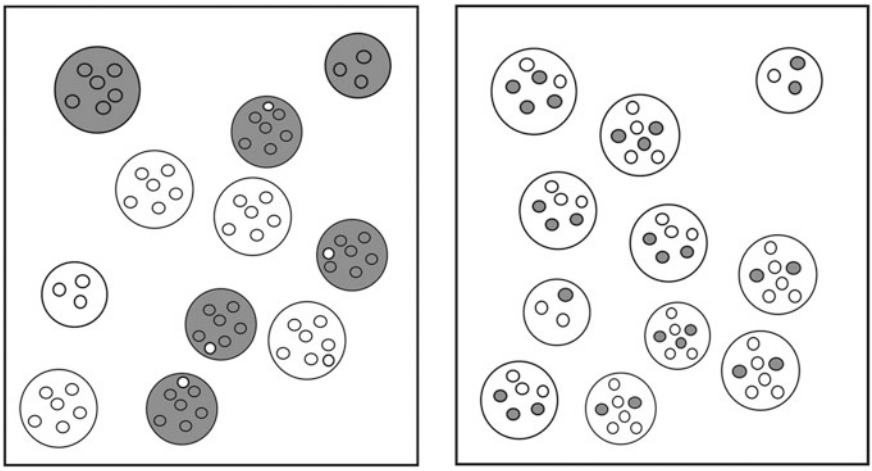
7.4.2 Stratification

Stratification is another feature regularly used in ILSA, including IEA studies. Unlike cluster sampling, stratification can increase sampling precision (Cochran 1977). This effect occurs if units within a stratum have similar survey outcome variables. For example, suppose students' abilities in operating electronic devices are the variable of interest in a survey. If students in rural areas lack access to such devices while students in urban schools use them in lessons on a regular basis, students in rural areas are likely to have systematically lower abilities to operate them compared to students in urban areas. In such a case, stratification by urbanization can increase sampling precision, given a constant sample size. This is because it is enough to select a few students from a larger pool of students that are all alike, making sure that a few are selected from every distinct group of students. Applying this rationale to the Einstein picture (Fig. 7.1a), six strata can be distinguished: the dark background, the mouth, both eyes, the nose, and the hair. Selecting some pixels from each of those parts has a high probability of providing a rather precise representation of the original picture.

Stratification can be employed at any sampling stage. If schools separate their students into high and low ability classes, it would be reasonable to stratify classes by ability before selecting a class sample. Equally, when selecting teachers within schools, it may be useful to stratify by subject, gender, or age group, depending on the variables of interest and expected group differences. IEA studies employ stratification at all stages whenever reasonable, and distinguish between different approaches of stratification (please refer to, e.g., Martin et al. 2016 for details on these methods).

Stratification can also be useful if countries are interested in comparing specific subgroups. In such cases, disproportional sample allocations to strata are possible, for example increasing the sample size for small subgroups within the population, thereby increasing the precision (and reliability) of results for small subgroups (Dumais and Gough 2012a). Computing sampling weights (see Sect. 5) will ensure that unbiased features of the subgroups are estimated, but allows for unbiased estimation of the whole population at the same time.

Stratification and cluster sampling may seem similar to a layperson, as both methods imply assembling units of the target population into homogeneous groups. However, they have opposing effects on sampling precision. The essential difference between these methods is that cluster sampling selects only some clusters from a large pool of clusters, while stratification selects some units within each of the strata (Fig. 7.3).



(a) Cluster sampling: Selecting some groups but all or nearly all units within groups

(b) Stratification: Selecting few units from each cluster

Fig. 7.3 A schematic illustration of the distinction between **a** cluster sampling and **b** stratification. *Note* Large bubbles represent homogeneous groups; shaded bubbles are sampled. Small white bubbles within selected clusters illustrate the fact that clusters are not necessarily sampled in full. Sampling precision will be higher with stratification, even though the number of sampled units is the same in both scenarios

7.4.3 Sampling with Probabilities Proportional to Size

Sampling with probabilities proportional to size (PPS), is also commonly used when selecting samples for IEA studies. This method is again described in detail in many textbooks (e.g., Lohr 1999). With PPS, selection probabilities depend on a "measure of size" of the sampling unit; large schools (e.g., those with many students) have high selection probabilities, and, vice versa, classes or students within large schools have low selection probabilities balancing out the two selection stages. Applying PPS aims for similar estimation weights in the core target population (a so-called "self-weighting design"), a measure that can increase sampling precision for this population but will likely decrease sampling precision for secondary target populations, such as schools, as their estimation weights are more variable (Solon et al. 2015). The effect occurs only if the chosen measure of size is related to the outcome variable. Moreover, PPS is a simple way to ensure selecting an approximately equal number of students in each sampled school while getting roughly similar final sampling weights.

7.4.4 Estimating Sampling Precision

Sampling precision is documented by reporting estimates of standard error[3] or confidence intervals along with any presented population feature. To account for the complex sampling design, standard errors must be estimated by respective statistical methods. In most IEA studies, jackknife repeated replication (JRR; Wolter 2007) is used to achieve unbiased estimates of sampling error, a method not available in statistical standard packages, such as for example, the base module of SAS (SAS Institute Inc. 2013) or SPSS (IBM Corporation 2016). Standard statistical packages use a default method applicable only to simple random samples and thus underestimate standard errors in complex sampling designs. To address this problem, IEA data sets are prepared for proper estimation of standard errors, and IEA offers with its IDB Analyzer a tool that applies the correct method automatically (see Chap. 13 for more information). Applying this method is of utmost importance regarding the reliability and validity of the study results: failure will likely lead to considerably underestimated standard errors, implying overly high precision levels in the results (i.e., confidence intervals that are too small), and can lead to falsely detecting significant group differences.

7.5 Weighting and Nonresponse Adjustment

Sampling weights are a reflection of the sampling design; they enable those analyzing the data to draw valid conclusions about population features from sample data. There are two reasons why it is necessary to compute weights and use them for data analysis, namely varying selection probabilities and nonresponse. Readers should refer to the technical documentation of IEA studies for detailed information on the exact computation algorithms for sampling weights and nonresponse adjustments and how to apply them for analysis; general introductions on weighting in ILSA can be found for example in Dumais and Gough (2012b), or Meinck (2015a). Rather than reiterating these procedures, this section focuses on the effects of weights for unbiased estimation in IEA studies, or, in other words, how weights help to retrieve valid and reliable statements about population features.

IEA studies account for the selection probabilities, varying due to multiple selection stages and the application of PPS, by computing design weights. In thinking about the meaning of the design weight values, the value of the design weight of a sampled unit refers to the number of other frame units represented by the sampled unit; a methodologically more accurate interpretation of the weight value is that the design weights must compensate for the fact that some units are part of a greater

[3] For estimates of achievement, standard errors are composites of measurement and sampling error in IEA studies. The standard error is equivalent to the sampling error for all other variables (see, e.g., Martin et al. 2017).

number of samples than others (Meinck 2015a). For example, if school A has twice the chances of being sampled (due to stratification, oversampling, and PPS effects) compared to school B, the latter school weight must be doubled to compensate for school A being selected twice as often as school B over all possible samples of schools with a given design. By doing so, over all possible samples, the weighted contribution of schools A and B will be identical to a census of all schools.

There is generally no doubt in the research community that weights are needed to achieve unbiased estimates of population features. The following simple example illustrates this effect (Table 7.1). Suppose that a researcher wishes to compare the percentages of female and male teachers teaching in private schools with the respective percentages in public schools. To achieve estimates with the same precision, the sample size for both private and public schools has to be identical, even though private schools compose only 10% of schools in the country.[4] Investigators seek to establish the average percentage of female teachers in each type of school, and overall. In this example, private schools are nine times more likely to be selected and, if the researcher does not apply a weighting factor, the percentage of female teachers in the total population would be drastically overestimated, inflating the average percentage of female teachers in all sampled schools (70%). However, if the appropriate design weights are applied, the estimate becomes design-unbiased and the average percentage of female teachers drops to 62%, a figure that is much closer to the expected average value for the larger part of the population, namely the public schools.

Design weights are computed separately for each sampling stage, and multiplied for sampling units of subsequent stages. If studies aim for multiple target populations, weights are computed separately for each of them.

The validity and reliability of ILSA studies and any other similar surveys can be negatively affected by failure to complete the survey; this is known as nonresponse.

Table 7.1 Example illustrating the effect of disproportional sample allocation on design weights and estimated population features

Factors		Private schools	Public schools	Total
Number of	Schools in population	1000	9000	10,000
	Schools in sample	100	100	200
Design weight		10	90	
Proportion of female teachers (%)	In population and sample	80	60	
	Unweighted sample (biased, descriptive statistic)			70
	Weighted sample (unbiased, inference statistic)			62

[4]The example assumes SRS of teachers in each stratum.

Nonresponse bias can be substantial when three conditions hold: (1) the response rate to the survey is relatively low; (2) there are significant differences between the characteristics of respondents and nonrespondents; and (3) nonresponse is highly correlated with survey outcomes. As opposed to item-level nonresponse (i.e., failure of respondents to answer individual single items), unit-level nonresponse creates higher risk to the validity of the ILSA because usually very little is known about the nonresponding units. Therefore, conditions (2) and (3) are difficult to assess. This applies especially to cross-sectional surveys and hence to ILSA surveys, and is the reason why sophisticated nonresponse models known from, for example, longitudinal studies, cannot be applied in ILSAs.[5] Instead, IEA specifies very strict requirements for participation rates. The requirements for participation rates in IEA studies can be obtained from the technical reports; generally they translate into minimums of 85% participation of sampled schools, 95% of sampled classes, and 85% of sampled students within participating schools/classes. Further, sampling weights are adjusted for nonresponse within supposedly homogenous adjustment cells (e.g., sampling strata, classes), assuming a non-informative response model (i.e., that nonresponse occurs at random) within these cells. For example, if all schools in urban areas participated in a study but not all schools in rural areas, the weight of the rural schools has to be increased to compensate for their loss in the sample. Thus, participating rural schools represent nonresponding rural schools, while urban schools only represent the urban schools they were sampled from.

Nonresponse can occur at every stage of sampling, hence, nonresponse adjustments are computed separately for each stage. All nonresponse adjustment and design weight factors are multiplied to achieve the "final" or "estimation" weight for every unit in the sample. Again, if studies aim for multiple target populations, estimation weights are computed separately for each of them and are stored with the respective data sets. The individual technical reports for each IEA study provide considerable detail on the procedures used to calculate the weights and nonresponse adjustments in each instance.

Computing sampling weights is an elaborative process requiring accuracy and attention to detail. To ensure the quality of the final weights, multiple checks for consistency and validity are conducted. Using exported WinW3S databases, information on classes and school sizes is compared with information on the sampling frame, and significant deviations are probed and clarified. The estimated sizes of total populations and the various subpopulations are compared to the sampling frames, official statistics, and previous survey data. Large deviations may indicate systematic errors in sampling and listing procedures, and are consequently carefully investigated. Variation of the weights is checked, and outliers are identified and handled depending on the outcome of these investigations.

[5]Longitudinal studies can base their nonresponse models on responses of the participants from earlier cycles. It should be noted that nonresponse bias analysis is feasible for countries that can provide the relevant data.

7.6 Sampling Adjudication

A final important quality control measure implemented in all IEA studies is sampling adjudication, mostly conducted as a face-to-face meeting at the end of a study cycle attended by the study directors, the sampling experts, and the sampling referee. The sampling referee assumes a key role regarding sampling quality, acting as an objective external expert who is consulted over the whole cycle of a study on particular cases and challenging situations regarding sampling matters. During the adjudication phase, all debatable issues are brought to the referee's attention, and are resolved in consultation with the adjudication committee, usually aiming for a consensus, but, if in doubt, following the referee's opinion. Violations of the sampling procedures for single schools are usually handled by dropping them from the sample; exceeding limits for exclusion and coverage rates, or moderate shortfalls in participation rate requirements[6] are annotated in all IEA reports; more severe transgressions may lead to separate reporting of the affected participating educational systems as a way to emphasize to readers that there are issues with data validity and due care is required when interpreting this data. In cases of extreme concern, the affected education system may be entirely excluded from the reported data.

Variables affected by high item-level nonresponse are also annotated in IEA reports. Users of complementary questionnaire data (e.g., those provided by parents or principals) are encouraged to pay close attention to completion rates, and to analyze the frequencies of missing values. When undertaking multivariate analyses, listwise deletion (i.e., removing all responses of a person from the analysis) may lead to exponential dropouts of cases, potentially resulting in biased analyses. In ILSA studies, this may often be the case when including information on socioeconomic features collected from parents in the analysis. Students with low achievement have higher missing rates for such features, possibly because parents do not want to reveal their (potentially low) levels of education or occupation, and thus the analysis results may be not representative for affected students.

7.7 Conclusions

The issues related to sampling are complex. Accounting for the sampling methods used is critical for understanding and interpreting study outcomes and correct analysis of the data. Users of IEA data and readers of the international reports should always refer to the extensive technical documentation that accompanies each study release and pay particular attention to the table annotations, which provide critical information designed to ensure valid data interpretation.

[6] That is, meeting participation rate requirements only after including replacement schools.

References

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: Wiley.

Dumais, J., & Gough, H. (2012a). School sampling methodology. In V. Greaney & T. Kellaghan (Eds.), *Implementing a national assessment of educational achievement* (Vol. 3). Washington, DC: The World Bank.

Dumais, J., & Gough, H. (2012b). Weighting, estimation and sampling error. In V. Greaney & T. Kellaghan (Eds.), *Implementing a national assessment of educational achievement* (Vol. 3). Washington, DC: The World Bank.

European Commission. (2020). *Eurydice* [webpage]. Brussels, Belgium: European Commission. https://eacea.ec.europa.eu/national-policies/eurydice/home_en.

IBM Corporation. (2016). *IBM SPSS Statistics for Windows* (Version 24.0) [Computer software]. Amonk, NY: IBM Corp. https://www.ibm.com/analytics/spss-statistics-software.

IEA. (2020). *IEA studies* [webpage]. Amsterdam, the Netherlands: IEA. https://www.iea.nl/studies/ieastudies.

Kish, L. (1965). *Survey sampling*. New York, NY: Wiley.

Lohr, S. (1999). *Sampling: Design and analysis*. New York, NY: Duxbury Press.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. (Eds.). (2003). *PIRLS 2001 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://www.iea.nl/publications/technical-reports/pirls-2001-technical-report.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). *Methods and procedures in TIMSS 2015*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://www.iea.nl/publications/technical-reports/methods-and-procedures-timss-2015.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2017). *Methods and procedures in PIRLS 2016*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://www.iea.nl/publications/study-reports/international-reports-iea-studies/methods-and-procedures-pirls-2016.

Meinck, S. (2015a). Computing sampling weights in large-scale assessments in education. *Survey insights: Methods from the field*, w*eighting: Practical issues and 'how to' approach*. http://surveyinsights.org/?p=5353.

Meinck, S. (2015b). Sampling design and implementation. In J. Fraillon, W. Schulz, T. Friedman, J. Ainley, E. Gebhardt (Eds.), *ICILS 2013 technical report* (pp. 67–87). Amsterdam, the Netherlands: IEA. https://www.iea.nl/publications/technical-reports/icils-2013-technical-report.

OECD. (2020). *OECD data*. Education [webpage]. Paris, France: OECD. https://data.oecd.org/education.htm.

Rutkowski, L., Von Davier, M., & Rutkowski, D. (Eds.). (2014). *Handbook of international large-scale assessment*. New York, NY: Chapman and Hall/CRC. https://doi.org/10.1201/b16061.

Rust, K. (2014). Sampling, weighting and variance estimation in international large-scale assessments. In L. Rutkowski, M. Von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment*. New York, NY: Chapman and Hall/CRC. https://doi.org/10.1201/b16061.

SAS Institute Inc. (2013). *SAS university edition 9.4* [Computer software]. Cary, NC: SAS Institute. https://www.sas.com/en_us/software/university-edition.html.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453.

Solon, G., Haider, S. J., & Wooldridge, J. M. (2015). What are we weighting for? *Journal of Human Resources, 50*(2), 301–316.

Statistics Canada. (2003). *Survey methods and practices*. Ottawa, Canada: Statistics Canada. http://www.statcan.gc.ca/pub/12-587-x/12-587-x2003001-eng.pdf.

UIS. (2012). *International standard classification of education. ISCED 2011*. Montreal, Canada: UNESCO Institute for Statistics. http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf.

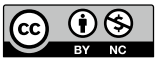
UIS. (2020). *Education* [webpage]. Montreal, Canada: UNESCO Institute for Statistics. http://data.uis.unesco.org/Index.aspx?DataSetCode=edulit_ds.

United Nations. (2019). *Convention on the rights of persons with disabilities* (CRPD). New York, NY: United Nations. https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html.
Weber, S. (2018). Sample design and implementation. In W. Schulz, B. Losito, R. Carstens, & J. Fraillon (Eds.), *ICCS 2016 technical report* (pp. 43–51). Amsterdam, the Netherlands: IEA. https://www.iea.nl/publications/technical-reports/iccs-2016-technical-report.
Wolter, K. (2007). *Introduction to variance estimation* (2nd ed.). New York, NY: Springer Verlag.
World Bank. (2020). *Education statistics* (EdStats) [webpage]. Washington, DC: The World Bank. https://datatopics.worldbank.org/education/.

Sabine Meinck works for the IEA in Hamburg, Germany, being head of both its Research and Analysis Unit and Sampling Unit. Since 2006, she has been involved with the sampling, weighting, variance estimation, and analysis of nearly all contemporary large-scale assessments in education. Her experience as a member of the project management teams for IEA's TIMSS and PIRLS, with the consortia of the IELS and TALIS Starting Strong Surveys, and on the joint management committees of IEA's ICILS, ICCS, ECES, and TEDS-M, have enabled her to develop a diverse knowledge and expertise; she also serves on the board of the IERI Institute.

Dr. Meinck coordinates, guides and supports all research activities within the IEA. Her main research interest lies with the science of cross-national large-scale assessments, and the methodological challenges of complex survey data.

In support of the IEA's enduring commitment to knowledge dissemination, Dr. Meinck has conducted multiple workshops for international audiences designed to share her experiences, and teach best practices and methodologies in educational research. Topics taught range from basic to advanced statistical methods, survey design, and publication and dissemination strategies for diverse audiences. Further, she teaches a Masters Course at the University of Hamburg on "Quantitative methods in educational research." Dr. Meinck is associate editor of the Springer journal *Large-scale Assessments in Education.* She is honored to serve as a peer reviewer for several scientific journals on educational research, and many educational research networks (such as AERA and CIES).



Chapter 8
Quality Control During Data Collection: Refining for Rigor

Lauren Musu, Sandra Dohr, and Andrea Netten

Abstract Studies undertaken under the auspices of the International Association for the Evaluation of Educational Achievement (IEA) use rigorous quality control processes to help ensure high quality data and facilitate cross-country comparisons of study results. This chapter provides an overview of the quality control processes implemented during data collection including the production of detailed manuals to standardize data collection procedures and the monitoring of how these procedures are implemented through national and international quality control programs. National and international quality control procedures help to confirm the validity of the data by monitoring data collection efforts. National quality control programs are carried out by participating countries with specific guidance provided by IEA and the international study centers (ISCs). International quality control monitoring ensures that sampling procedures are followed at participating school, classroom, and student levels, monitors on-site data collection to check whether the test administration procedures and security guidelines set by IEA and the ISCs are met, and provides information on circumstances that occurred during data collection that could influence the data reliability and quality. This chapter provides a history of the development of these different quality control procedures and information on their implementation within IEA studies. The chapter concludes with a series of recommendations for potential improvements to consider for the future.

Keywords Data collection · Large-scale assessment · Quality assurance · Quality control · Test administration

L. Musu (✉) · S. Dohr · A. Netten
International Association for the Evaluation of Educational Achievement (IEA), Amsterdam, The Netherlands
e-mail: l.musu@iea.nl

S. Dohr
e-mail: s.dohr@iea.nl

A. Netten
e-mail: a.netten@iea.nl

H. Wagemaker (ed.), *Reliability and Validity of International Large-Scale Assessment*,
IEA Research for Education 10, https://doi.org/10.1007/978-3-030-53081-5_8

8.1 Introduction

Quality control is an essential component of international large-scale assessment. Rigorous quality control processes help ensure high quality data and facilitate cross-country comparisons of study results. Although quality control can and does exist at all stages, including before, during, and after data collection, the term is often used within IEA studies to refer to the monitoring of data collection procedures in schools and classrooms, especially at the international level. Quality control procedures during data collection serve several purposes:

- to standardize data collection by providing detailed procedural manuals for countries to follow;
- to check on-site whether the test administration procedures and security guidelines set by IEA and the international study centers (ISCs) and outlined in these manuals are followed;
- to ensure that sampling procedures are adhered to at participating school, classroom and student levels; and
- to provide information on any circumstances that occurred during data collection that could influence the data quality.

Deviations from the standardized procedures outlined in the manuals are a threat to reliability and validity. These procedures are put in place to ensure data collection occurs in a comparable way across countries with limited disruptions to the process. Therefore, national and international quality control procedures help to confirm the validity of the data by monitoring data collection efforts and ensuring appropriate participation from the sampled schools, classrooms, and students.

The terminology used to describe the process for monitoring data quality varies in the research literature, but quality control and quality assurance are the most commonly used labels. Some studies use these terms interchangeably, but they often refer to slightly different aspects of the processes for ensuring data are valid and reliable. Within IEA studies, as well as in other large-scale research, quality control procedures are often part of a larger quality assurance program. Quality assurance generally refers to the full spectrum of procedures that are implemented while quality control refers to the measurement of certain characteristics of data collection procedures to ensure that certain standards are met (Statistics Canada 2010). Various aspects of quality control and assurance are used throughout IEA studies. For example, response and range checks are used during data processing to look for implausible values or evidence of data tampering. These quality control measures are covered in other chapters, but this chapter focuses on quality control materials and procedures for test administration, that is, when the assessments and surveys are administered in the sampled schools and classrooms.

Quality control procedures for data collection and test administration consists of three major components:

(1) Production and distribution of standardized manuals managed by IEA and the ISC;
(2) National quality control procedures and monitoring managed by participating countries; and
(3) International quality control procedures and monitoring managed by IEA.

There are three major aspects of quality control during test administration, with various tasks and responsibilities for the organizations and entities involved in the data collection (Table 8.1).

All three elements of quality control during data collection play an important role in ensuring that countries follow standardized procedures and monitoring whether those procedures are implemented in a uniform way across countries. In describing ideal quality control procedures for large-scale assessment, Cresswell (2017) noted that quality management during data collection should include "the development and agreement of standardized implementation procedures, the production of manuals which reflect agreed procedures, the recruitment and training of personnel in administration and organization—especially the test administrator and the school coordinator, and the quality monitoring processes—recruiting and training quality monitors to visit schools and national centers" (p. 195). IEA studies follow these best practices with the use of collaboratively developed manuals, and national and international quality monitoring programs.

Table 8.1 Roles and responsibilities for quality control procedures during test administration

Procedure	IEA and the ISC	Participating countries
Manuals	IEA and the ISC develop standardized manuals to detail data collection procedures	Countries are responsible for translating and adapting the provided manuals to necessary languages and their national context
National quality control	IEA and the ISC develop and provide manuals for use in national quality control procedures	Countries have primary responsibility for implementing a national quality control program using guidelines and recommendations provided by IEA and the ISC
International quality control	Materials and training for IQCMs are developed collaboratively by IEA and the ISC. IQCMs solely report to IEA and the ISC	Countries coordinate with the IQCM to provide requested materials and select schools for monitoring visits

Notes ISC = international study center; IQCM = international quality control monitor

8.2 Manuals

To help ensure that national centers, schools, and test administrators are familiar with the required procedures for data collection, detailed manuals are developed and distributed to countries. The use of manuals is one of the earliest procedures implemented during IEA studies to ensure standardization across countries in the data collection. This section describes the development of these manuals and their implementation.

8.2.1 Early Quality Control Procedures and the Development of Manuals

The first IEA studies were conducted by a team of researchers in different countries who met at regular intervals to plan and implement the studies. The Pilot Twelve-Country Study, which took place in 1960 (IEA 2020a) is considered the first IEA study. The study was sponsored by the UNESCO Institute for Education and many of the founding members of IEA (e.g., Douglas Pidgeon, Benjamin Bloom, Robert Thorndike, and Torsten Husén) took part in the design, implementation, and analysis of the study. While the Pilot Twelve-Country Study was a milestone in the development of international large-scale assessment in general, there is little mention of quality control procedures in the reporting of the study results. In fact, the authors cautioned against overinterpretation of the results in the report and chose not to show total scores in a way that facilitated comparison between countries due to concerns about data comparability (Foshay et al. 1962).

Standardized procedures for data collection were developed in the form of detailed procedural manuals beginning with the next IEA study, the First International Mathematics Study in 1964 (IEA 2020a). Publications for this study also acknowledged the importance of standardization when describing the administration of the data collection procedures. "It was extremely important to ensure that as far as possible uniform methods of procedure were employed in the testing programme in all countries" (Postlethwaite 1967, p. 46). To accomplish this, a small committee of the researchers involved in organizing the study developed three manuals: one for the national centers, one for individuals coordinating data administration at each school, and one for the individuals administering the test. Additional information and instructions were also sent to participating countries in the form of circular letters and bulletins.

Similar to the First International Mathematics Study, individuals involved in planning the different assessments for the Six Subject Survey in 1970–1971 (IEA 2020a) produced detailed manuals that were distributed to the national centers, the school coordinators, and the test administrators. While closely scripted administrative procedures were described in the manuals and other documentation was provided to schools, oversight of the assessment administration was left to the discretion of

the participating schools under the assumption that the procedures outlined in these manuals were being followed.

Subsequent studies continued to produce detailed manuals and the content and structure of the manuals has expanded over time. These manuals form the backbone of IEA data collections in that they carefully explain the data collection and survey administration procedures that should occur within schools and classrooms.

8.2.2 Current Implementation of Manuals

Over time, IEA and the ISCs have developed increasingly detailed manuals for use by participating countries. These manuals include survey operations procedures (SOP) manuals, a school coordinator manual, a test administrator manual, and manuals for national and international quality control monitors. The manuals detail procedures for the test administration and data collection that have been agreed upon by IEA, the ISC, and the national research coordinators (NRCs). All manuals provided to national study centers are in the English language but can be translated to a national language as needed by the national study centers themselves.

SOP manuals outline the process of data collection from beginning to end. Each manual details a specific part of the study process such as sampling, preparing assessment instruments, and scoring items after the assessments have been completed. Manuals for school coordinators, test administrators, and national quality control monitors are included as supplementary materials with relevant sections of the SOP. To guide NRCs through the process, SOPs are released on a staggered basis to coincide with major data collection milestones.

School coordinators are responsible for ensuring that sampled classes, teachers, and students actually participate in the assessment. They also oversee the distribution, completion, and collection of testing materials and questionnaires. The manuals for school coordinators provide detailed instructions on the ways in which these tasks should be completed, allowing for some individual variation between countries due to contextual factors such as confidentiality laws. The manuals include extensive details on the role of the school coordinators, including completion of class and student listing forms and tasks for securing materials prior to testing, distributing them on the testing day, and returning them to the national center after testing is complete.

Test administrators are responsible for administering the assessments. Test administrators must ensure that each student receives their specific testing materials and that the assessments are given in a standardized way across countries. This includes following a specific script with instructions for students taking the assessment.

Manuals for national and international quality control monitors describe the roles and responsibilities of those positions. The manuals for national quality control monitors include a description of the roles and responsibilities and sample classroom observation forms that can be used during school visits. Manuals for

international quality control monitors also include a description of roles and responsibilities and observation forms. In addition, the manuals for the school coordinators and test administrators are provided so that international quality control monitors can ensure that these individuals are adhering to procedures when they do their classroom observations.

Since countries have varying degrees of familiarity with administering large-scale assessments, the different manuals are designed to provide all the details necessary to carry out the data collection procedures. All countries are asked to follow the procedures detailed in the manuals without significant deviation to ensure consistency, and training is provided for NRCs so that they understand the structure and content of the manuals and the procedures contained therein. National and international quality control monitoring also help to ensure that the procedures described in the manuals are carried out as specified.

8.3 National Quality Control Procedures

IEA and the ISC recommend that countries implement a national quality control program in order to monitor data collection efforts. Countries also want to monitor the quality of their data collection efforts so that they can intervene when problems are discovered and so that they have confidence in the data collected within their country. National quality control programs were developed individually in some countries before the international quality control program existed, but recent studies base guidelines for national quality control on the international program. Although similar in purpose and scope, the national quality control monitoring program and international quality control monitoring programs are designed to be separate but complementary to one another. For example, for IEA's Trends in International Mathematics and Science Study (TIMSS) 2015, NRCs were required to send national quality control monitors to a 10% sample of the schools to observe the test administration and document compliance with prescribed procedures. These site visits were in addition to the visits to 15 schools conducted by the international quality control monitors (Johansone and Wry 2016).

8.3.1 Development of National Quality Control Procedures

As studies became increasingly large and more complex, oversight of studies at the international level was helped by the establishment of ISCs and management at the national level was helped by the appointment of NRCs. The Second International Mathematics Study of 1980–1982 (IEA 2020a) was the first IEA study to explicitly mention the appointment of NRCs, then referred to as national technical officers, in each country (Garden 1990). While earlier studies used national study centers to

help coordinate the data collection, there was not always one individual person coordinating the work at the national level. NRCs were usually trained in how to perform their duties. In recalling the evolution of IEA, Alan Purves, Chair of IEA from 1986 to 1990, explained "there was on-the-job training for the national technical officers, as they were called. Usually [experts from IEA] visited each of the centers for several days" (Purves 2011, p. 546). Such training focused on the implementation of data collection procedures at the national level. However, training was not a requirement and was generally not given in a standardized way to all NRCs.

The use of international study centers (ISCs) and NRCs continued with the Second International Science Study in 1983–1984 (IEA 2020a) and the Reading Literacy Study in 1990–1991 (IEA 2020b). While international quality control was still lacking, some countries were implementing stricter independent quality control procedures for data collection at the national level. For example, for the Reading Literacy Study, the United States (US) chose to hire field staff with no associations with the schools themselves to serve as test administrators. This allowed the coordinating center to train the field staff and thus try to ensure more standardized procedures. As stated in the US technical report, "It was felt that data collected in this way would be far more comparable than that collected under an infinite number of differing conditions" (Binkley and Rust 1994, p. 41). While these procedures led to increased confidence that data were collected in a standardized way, they were also admittedly costly and were only implemented in this exact way in the US. Other countries implemented their own quality control procedures, but there were no checks implemented across countries to ensure that the standardized procedures were being followed.

TIMSS 1995 was the first study that explicitly laid out recommended procedures for quality control at the national level (Martin et al. 1996b). The recommendations for national-level procedures closely mirrored those that were being implemented at the international level during this same study. It was recommended that NRCs arrange for quality control observers to visit a sample of schools on the day of testing. To help facilitate this, IEA and the ISC developed a manual and accompanying forms based on the international materials that could be adapted for use at the national level. While NRCs could implement their own procedures for national quality control, they were encouraged to use the materials provided.

8.3.2 Implementation of National Quality Control Procedures

As part of the materials provided for participation in an IEA study, NRCs are given detailed information on how they can implement a national quality control program. Similar to the procedures in TIMSS 1995, IEA and the ISC produce detailed manuals on how to implement a national quality control program that will complement the international program. For example, TIMSS 2015 instructed NRCs to send national

quality control monitors to observe the test administration and document whether required procedures were followed in 10% of participating schools (Johansone and Wry 2016).

These national quality control monitor manuals are the primary resource provided by IEA and the ISC for national quality control. They are designed to train quality control monitors to observe test administration procedures in their country. For the most part, countries use the national quality control monitor manuals, but they are given flexibility in the best way to implement the program to meet the needs of their country. Some countries choose to implement altered national quality control procedures and sometimes countries are unable to implement the program as prescribed due to lack of funding. For example, in TIMSS 2019, one country with centrally trained test administrators who were totally independent of the sampled schools felt it was sufficient to observe a smaller percentage of these administrators in the field. Despite some difficulties or changes to procedures, the majority of countries implement national quality control procedures as specified in documentation provided by IEA and the ISC.

In addition to supporting quality control monitoring procedures at the national level, IEA and the ISC support standardized procedures at the national level by providing both online training and direct presentations to NRCs on appropriate procedures. As part of this training, the detailed manuals for test administrators and school coordinators described earlier in the chapter are provided and discussed with NRCs.

At the end of the data collection and submission process, NRCs are required to provide a summary report to IEA and the ISC describing their national quality control activities. In addition, NRCs provide feedback to IEA and the ISC through the survey activities questionnaire (SAQ). This questionnaire is meant to document study procedures at the national level, from sampling all the way through submitting the final data. NRCs were initially asked questions from the SAQ during a structured interview with the international quality control monitor (IQCM). In recent years, the SAQ has always been distributed electronically to NRCs by the ISC once all the data from a country has been received.

The purpose of the SAQ is to gather information from the NRC and other national center staff on how well procedures and materials worked and what can be improved in the future. The SAQ asks about sampling procedures and manuals and includes questions on contacting and recruiting schools, focusing on how schools were contacted and how school coordinators were trained. Subsequent sections of the SAQ include questions about how assessment materials were adapted and translated, how materials were distributed to schools, and whether there were difficulties in the actual administration of the assessments. In addition, there are sections asking about scoring the assessments and preparing and submitting the final data. This detailed set of questions allows the ISC and IEA to get a sense of national-level quality control procedures and identify areas where there may be potential issues or aspects that can be improved in the future. Information from the SAQ is often reported alongside information from the international quality control monitoring program in technical reports to provide a more in-depth picture of the data collection process as a whole.

8.4 International Quality Control

Quality control during test administration includes an international quality control monitoring component in which independent observers visit a sample of classrooms to ensure that standardized procedures and test security guidelines are being followed. Individuals who are independent of the national study centers are hired and trained by IEA to conduct these monitoring procedures. In recent years, international quality control has been implemented in a standardized way across studies, although differences exist because all quality control programs are tailor-made to accommodate the specific needs for each study. IQCMs (although the term international quality observers [IQOs] is used in some studies) are individuals hired in each country to observe (independently from the national center) the actual data collection in a subsample of all selected schools in their country and record whether the standardized procedures are followed.

8.4.1 Development of International Quality Control Procedures

As described in Sect. 8.2, the development and sharing of standardized manuals was the primary mode of quality control for data collection during the earliest studies. However, even with these manuals, challenges to ensuring uniform data collection and high data quality across countries were common in earlier IEA studies. When referring to the Six Subject Survey, Benjamin Bloom, one of the founding members of IEA, commented, "Inevitably there are difficulties in ensuring that the right tests, etc. get to the right students and that all understand exactly what it is they have to do. In surveys that cross country boundaries, especially where many different languages are involved, administrative problems are magnified and great care in planning is necessary if errors are to be avoided" (Bloom 1969, pp. 10–11).

Although the training and materials described earlier in this chapter helped to ensure additional standardization for data collection, there was still no explicit oversight at the international level to provide information about what happened within schools. This lack of oversight presented challenges. In one of the publications on the results of the Second International Mathematics Study, R. A. Garden, former NRC for New Zealand, commented, "During the study it was the negative aspects which dominated our lives - the National Research Coordinators (NRCs) who did not follow instructions, the postal delays, the misunderstandings, the unreadable data tapes, the miscoded data, and so on" (Garden 1990, p. 1).

In response to some of these issues, the Third International Mathematics and Science Study of 1995 was the first IEA study to implement a coordinated quality control program at the international level. Boston College was the ISC for TIMSS 1995 and oversaw the development of international quality control procedures within the study. Funding, always an issue in previous IEA studies, was resolved in TIMSS

1995 when Boston College received a grant from the US Department of Education to complete data collection for the study (Mullis and Martin 2018). Albert Beaton headed the study at Boston and brought with him a wealth of experience in psychometrics, data collection, and analysis from his many years at Educational Testing Service (ETS) and his work on the US National Assessment of Education Progress (NAEP). Two other experienced researchers also joined the Boston College team, namely Michael Martin and Ina Mullis. Michael Martin was a former NRC for Ireland's international studies, and Ina Mullis, like Albert Beaton, had worked with ETS and NAEP (Schwille 2011). This shared experience with national and international assessment and psychometrics helped shape the work on TIMSS 1995.

Although NAEP was administered within a national context, the study presented some challenges that were similar to those of large-scale international assessments. It thus provided a relevant example of ways to examine and implement quality control across countries. As the number of countries involved in IEA studies continued to increase, their level of comfort and familiarity with implementing large-scale assessment varied widely. The team at Boston College were able to provide leadership on this drawing on the experience of NAEP. In addition to the on-the-ground expertise from Boston College, the US funders for the study wanted to have confidence that the results were comparable across countries. They therefore requested that more rigorous oversight of quality control procedures be included at the international level as a condition of the funding.

While the data collection procedures themselves did not change significantly for TIMSS 1995, the level of oversight for these procedures and the amount of training provided to coordinators at the national and international levels did increase. Similar to previous studies, detailed manuals outlining standardized procedures for data collection were developed collaboratively and used to guide the data collection, although additional manuals were developed and, in many cases, they included greater levels of detail than that provided for prior studies. What was also unique for TIMSS 1995 was that IQCMs were employed and centrally trained to perform classroom-level observations of the data collection as it was taking place (Martin et al. 1996a). Boston College helped organize five different training sessions in various locations around the world so that all IQCMs had the opportunity to attend a session.

The duties of the IQCMs for TIMSS 1995 were standardized across all countries and communicated during the training sessions. NRCs and IQCMs prepared classroom observation tracking forms for each school and classroom under the guidance of Boston College (Martin et al. 1996a). In addition, IQCMs for TIMSS 1995 were asked to interview the NRC about all aspects of the data collection using a structured interview. The interview covered the topics of sampling, experiences working with school coordinators, translation of instruments, preparing test booklets (including sending them to schools and arranging their return), procedures for national-level quality control monitoring, coding of open-ended assessment items, and recording and submitting the final data (Martin et al. 1996a). The questions from this structured interview were later used to develop the survey activities questionnaire (SAQ) that is still in used across IEA studies.

Another new development in the monitoring and standardization of quality control came soon after TIMSS 1995 with the release of the technical standards for IEA studies, which were published in 1999 (Martin et al. 1999). The IEA technical standards focus on the international design and management of the studies, but also address aspects of national implementation that are important for collecting high quality, internationally comparable data.

Quality assurance and control are the primary focus of two of the technical standards. The first relevant standard is the "Standard for developing a quality assurance program." This standard specifies that operational documentation prepared by the ISC should emphasize quality control as integral to all aspects of a study, particularly data collection activities. The purpose of this technical standard notes that "[quality control] is particularly important for activities such as test administration, which may be conducted by school personnel and therefore outside the control of study staff" (Martin et al. 1999, p. 27). The guidelines recommend making visits to a sample of data collection sites. The data collection monitoring is described as essential for national centers to implement, but also highly recommended at the international level to ensure that unbiased and trained observers can report on the extent to which the sampled schools and classrooms follow the specified procedures.

The second technical standard to specifically address quality control is the "Standard for implementing data collection quality control procedures" and it states that, "Quality control should be an integral part of the study at both the national and international levels. Quality control encompasses both the internal mechanisms that are built into each stage of data collection to ensure that procedures are implemented correctly, and external reviews administered by staff members who are separate from the staff being evaluated" (Martin et al. 1999, p. 59). The implementation of this standard is important in ensuring that the data collection procedures meet the study requirements set by the ISC. The guideline for implementation of this standard emphasizes the ways in which quality control should be built into many steps of the data collection process, for example, hiring and training qualified quality control monitors to assist in observing the administration of the data collection. The guidelines further state that both the ISC and the national center should conduct separate quality control monitoring checks.

International quality control monitoring occurred in all of the IEA studies that followed TIMSS 1995 with a few exceptions. For the Civics Education Study of 1999, there was not enough funding to implement an international quality control monitoring program. However, the ISC (Humboldt University of Berlin) advised NRCs to implement broader national level quality control procedures. NRCs were asked to phone 25% of the tested schools to interview the school coordinator about how testing was done, whether there were any problems encountered, and whether there were any deviations from the testing procedure outlined in the manual (Humboldt University of Berlin, unpublished internal report 1999). The ISC provided formal guidelines for the telephone interviews, along with instructions for how NRCs could select a simple random-sample of the participating schools.

Understanding the development of international quality control procedures also requires an understanding of the development of IEA as an organization. Both structural and financial changes to IEA, as well as the countries involved in IEA studies, have led to changes and developments over time. One country representative provided the following anecdote in regard to early data collection. "This [visits to schools] was a very hard job, some researchers had to reach the schools on horseback since no other means of transportation existed" (Purves 2011, p. 552). The difficulties encountered in reaching and communicating with schools were a potential barrier to implementing a coordinated international quality control monitoring program in the earlier studies. Advances in technology, such as the increased use of email, video chats, and webinars, also helped to facilitate coordination of the studies and the international quality control program.

Changes within the organizational structure of IEA itself also contributed to the increased possibilities for implementing an overarching quality control program. In the early years of the organization, membership was voluntary and there was no formal structure for funding the various studies. Administrative costs for individual studies during the earlier years were often provided by a single organization and individual countries were responsible for funding the collection within their own country. While countries still fund their own individual data collections, a formal fee structure was implemented in the 1990s to help IEA cover the administrative costs of the various studies. Kimmo Leimu, a former NRC for Finland, summarized the impact of this development on project management and oversight: "With the number of actively participating systems increasing up to some 70, recent studies and years have witnessed IEA's development into a more comprehensive organization both nationally and internationally, despite the fact that national fees have become an indispensable condition for participation. At the same time the projects have become more carefully controlled from beginning to end, with ever-increasing formal procedures detailed at each stage through planning, development, fieldwork implementation, and reporting. In recent years, an international quality monitoring element of test management has been added. An efficient data processing unit enables smooth state-of-the-art analyses of the massive data sets" (Leimu 2011, p. 599). Although the international quality control for TIMSS 1995 was funded primarily by the US Department of Education, subsequent studies have been able to draw on the financial resources provided by the formal IEA funding structure that was implemented during the 1990s. The increasing numbers of participating countries also helps contribute to funding study oversight at the international level.

This expansion of involvement of different countries has added some additional challenges in more recent studies. As Hans Wagemaker, Executive Director of IEA from 1997 to 2014, commented, "[f]or IEA, the inclusion of the broader range of countries with distinctive local circumstances has meant the development of new ways of working to ensure that all countries can participate and that studies continue to achieve the highest technical standards" (Wagemaker 2011, pp. 268–269). Experience in administering large-scale assessment varied widely across countries, especially as countries joined IEA studies for the first time. It was therefore of increasing importance that standardized procedures were clearly documented in the

procedural manuals developed for the studies. Further, the use of independent quality control monitors at the international level helped to ensure that these procedures were being implemented in participating schools and classrooms. While the challenges of a larger and more diverse group of countries involved in the studies require close monitoring of quality, they also enable IEA studies to collect and report data on a broader spectrum of education systems around the world.[1]

8.4.2 Implementation of International Quality Control Procedures

The objective of the international quality control monitoring program is to document data collection procedures and to verify that NRCs, school coordinators, and test administrators are following the standardized procedures for data collection. In order to select IQCMs, NRCs are asked to nominate or recommend an individual to serve in this role for their country. All nominations are screened by IEA to ensure that each individual meets the criteria for being an independent observer. For instance, nominees should not be a member of the national study center or a family member or friend of the NRC. The IQCM is often a school inspector, ministry official, or retired school teacher. In many instances, IQCMs are retained across study cycles and continue to serve in this role for subsequent studies.

IQCMs are required to be fluent in both English and the main language of administration, and should have easy access to and experience working on a computer. Additionally, IQCMs need the flexibility to perform their tasks within the required timeline. This often results in a lot of work needing to be completed within a short time frame. To help accomplish this, IQCMs sometimes work together with assistants to help with the classroom observations. Assistants are most common in large countries or countries where the assessments occur on only one or a few days. This helps ensure that IQCMs or their assistants are able to visit the specific schools that are selected for quality control monitoring. Ideally, assistants come from different areas of the country so that a broader geographic spread of schools can be included in the observations.

IQCMs are trained in face-to-face sessions on the standardized procedures for conducting the observations. On average, the IQCM training sessions last between one and two days. Trainers, usually from IEA or the ISC, provide a detailed manual outlining the roles and responsibilities of the IQCM, providing information on the survey operation procedures and assessment design, and including copies of the international questionnaires in English. These materials help ensure that observations and interviews are conducted according to a defined protocol and that responses are

[1]The authors would like to acknowledge several individuals who provided their time and expertise on the history of quality control procedures during data collection. Specifically, special thanks are extended to Dirk Hastedt, Paula Koršňáková, Barbara Malak-Minkiewicz, Michael Martin, Ina Mullis, Tjeerd Plomp, Heiko Sibberns, Jack Schwille, and Judith Torney-Purta.

documented on standardized forms. It also helps to familiarize IQCMs with the procedures as they are supposed to be implemented.

Once IQCMs are selected and trained, they conduct school visits and classroom observations. There are three primary purposes of the tasks performed by IQCMs. The first purpose is to validate the sampling within the country. Specifically, it is important to know that the sampled schools, classrooms, and students are the ones actually participating in the assessment. The second purpose is to ensure standardized test administration and data security procedures set by the ISCs are being followed. The third is to provide information on occurrences during data collection that could have an influence on the data quality.

IEA uses rigorous school and classroom sampling techniques to include a representative group of students within each country (see LaRoche and Foy 2016, and Weber 2018 for recent examples). For the majority of studies, sampling is conducted in three stages. First, the countries are asked to provide an exhaustive list of all eligible schools, from which the number of schools to be sampled are selected (usually 150 schools). Selected schools must then provide a list of all the classes that contain students from the target population. From this list, a classroom is selected and all students in the classroom should be included in the study, with a few exceptions for students with disabilities or those that do not speak the language of the assessment. The reliability and validity of the data collected depend on countries closely adhering to the sampling frame that they complete in conjunction with IEA and the ISC. It is therefore essential for IEA and the ISCs to ensure that the agreed upon sampling plans are followed within each country and that any deviations are noted and accounted for. To this end, the international quality control monitoring program provides an opportunity for IQCMs to visit a sample of schools and check that the school name and location match the sampling plan.

As part of their duties, IQCMs also ask school coordinators for information to help validate the within-school sampling. For example, they ask for a list of classes in the target grade(s) at that school and ask whether there are any students at the school that would not be included in these classes. These questions help to validate that the sampled classrooms and students provide the actual data.

Another of the other main aims of the international quality control program is to ensure data comparability by monitoring whether test administrators and school coordinators are following standardized procedures for data collection that are detailed in the manuals. To ensure the manuals are being followed overall within a country, IQCMs are asked to visit samples of individual classrooms on the day of data collection to observe the procedures and note any deviations from the standardized protocols.

The classroom observations during the data collection process are the central and most time-intensive aspect of the IQCMs' duties. IQCMs are generally instructed to visit a sample of either 10% of schools or 15 schools per country. This differs slightly depending on the particular study. For example, the most recent administrations of TIMSS and PIRLS both specified that 15 schools should be selected (Johansone and Wry 2016, 2017). In studies where multiple grade levels are included (e.g., TIMSS), 15 schools per grade should be selected in each country. Further, when one

or more benchmarking participants from the same country participate in a study, five additional school visits are required for each benchmarking entity. The most recent administrations of ICCS and ICILS specified that 10% of schools should be sampled (Koršňáková and Ebbs 2015; Noveanu et al. 2018).

Much of the content of the classroom observation records has remained consistent over time. However, in recent years electronic assessments have become more common across the studies and additional questions have been added to account for this new administration method. The international quality control monitoring process for computer-based assessments is closely aligned with the process for observing the more traditional pen-and-paper assessments. However, some parts of the observation protocol are altered in order to account for the electronic medium of the assessments. For example, the PIRLS 2018 observation record for the computer-based PIRLS modules included questions on whether any technical issues occurred during the testing session (e.g., whether any of the USB sticks were defective, whether the class needed to be split into multiple sections due to the computer availability in the school, and whether any technological problems occurred during the testing session; Johansone and Wry 2016).

The information from the classroom observation records can help inform what may have happened during the data collection to impact the results if issues of comparability do arise at any point during the data management and reporting process. IEA and the ISC receive, compile, and analyze the information from IQCMs to establish whether procedures were followed both within and across countries. In this way, the program can both illuminate specific instances within a country that may need to be examined more closely and identify systematic issues that may be occurring across countries. Although cheating is rare, the program can also help to prevent cheating and incentivize close adherence to study procedures by ensuring that countries know that the data collection will be monitored. Another reason to ensure that procedures are being followed is to check that the assessments and the questionnaires remain under strict control. This helps ensure that items remain confidential so that they can be used for trend comparisons in future studies. It is also important that the individual responses to survey items remain confidential.

In addition to observing the test administration procedures, IQCMs conduct interviews with the school coordinators in the selected schools. IQCMs ask how and when test items were delivered and how they were kept secure prior to the scheduled test administration. Finally, IQCMs collect the final version of all data collection materials from the NRC. These materials include the final manuals for the school coordinators and test administrators, student and teacher listing and tracking forms, and final copies of all questionnaires and assessments. The student and teacher listing and tracking forms are used for sampling validation, while the other materials are used to check the translation of the materials that were actually used during data collection procedures.

While seemingly straightforward, the international quality control monitoring program is essential in ensuring the quality of the data collected. It helps to provide a complete picture of what is actually happening within the schools and classrooms

themselves. This would be difficult or perhaps impossible to capture in any way other than having independent on-the-ground observers record this information.

8.5 Future Directions

The establishment and development quality control procedures during data collection has been an ongoing process since the early IEA years and it is still evolving. This is necessary because of the changing nature of large-scale assessments, especially as technology evolves and more studies and more countries make use of computer-based assessments. While these changes add new layers of complexity, they also offer the opportunity to reflect on what is working well and where policies and procedures may need to be adapted. Currently, information on quality control procedures is disseminated through technical reports or used by IEA and the ISC as a check to ensure that data are valid and reliable. In some ways, these rigorous procedures contribute to the fact that data comparability and high quality data are taken for granted. The program generally uncovers very few issues, but there would be no way to know whether issues exist without the program. Although few issues are usually noted, it is essential to continue documenting adherence to standardized procedures to ensure that studies maintain the consistently high quality for which they have come to be known.

In addition, the context surrounding the assessments themselves has changed over time. The early IEA studies were conducted by researchers for research purposes. Over the years, policymakers and country leadership have taken an increased interest in the results in many of the participating countries leading to assessments becoming more high stakes in those countries. This is important because political pressure to perform could influence the behavior of NRCs, school officials, test administrators, and even the students themselves. In a high stakes environment it is even more important to ensure that there are fully independent observers monitoring the data collection process.

While the quality control procedures described in this chapter are important in ensuring data quality, individual components are regularly evaluated to ensure the quality control monitoring during data collection continues to accomplish the intended purposes. It is also vital to consider new ways in which the information from quality control procedures can inform researchers and study participants. Advances in technology offer opportunities to consider new ways to streamline and improve the process of quality control monitoring.

One issue that is not currently addressed in the international quality control monitoring is what may be happening in schools prior to the day of testing. Organizing the assessment administration within schools is a complex and time-consuming process. Therefore, schools know several months ahead of time whether they have been selected for inclusion in the study. While this information should not influence the educational activities within the school in the time leading up to the actual data collection, there is currently no way to monitor whether this is actually the case;

namely whether some countries are "teaching to the test" or coaching their students prior to the data collection. Some aspects of the assessment design mitigate attempts to provide students with direct answers to questions. This helps prevent efforts to give students the correct responses to individual questions but does not preclude coaching prior to testing.

One possibility to screen for this would be to have IQCMs monitor activities leading up to the data collection more closely. The logistics of organizing any type of pre-assessment monitoring could be complicated and costly, so careful planning would be needed before implementing this type of expansion to the quality control program. In addition, quality monitoring processes that occur once data submission is complete help to check for anomalies in country-level data. For example, sudden and dramatic changes in the mean level of performance within a country would be cause for concern. However, it could still be helpful to know about what is happening in schools at earlier points in time in order to best determine how to address potential situations where this may occur. Ultimately, these efforts are self-defeating for countries because they prevent valid measurement of student performance, which in turn precludes countries from accurately evaluating their education policies and practices for potential changes or improvements.

Another potential issue is that no international quality control monitoring occurs during field testing. The data for field testing is not disseminated externally, but field testing can be seen as a trial run of sorts for the main data collection. Thus, observing the field test could inform IQCMs of issues that need to be resolved prior to the main assessment. This would give the ISC and IEA time to consult with NRCs to ensure that corrections to the procedures can be made in time for the main data collection. National quality control is recommended during field testing, so increased communication in regard to national quality control procedures so that IEA and the ISC are aware of issues that arise during the field test can help countries problem solve before the main study.

In addition to potential pre-assessment monitoring, the procedures implemented during the actual data collection could be enhanced. As mentioned earlier, computer-based assessments are becoming more common. While some of the quality control monitoring procedures have been adapted to account for this medium, there has been little use of the computer-based assessments themselves as a way to collect data on quality control. For example, questions about assessment start and end times could be answered using data stored when students begin and end the assessment. The use of log file data is currently under investigation as a source of information that would further enhance the data already being collected as part of the international quality control program, so this is an area of active development. In addition, some of this data is already used to monitor response patterns for anomalies during the data cleaning phase.

Electronic platforms could also be used to streamline and improve the process of receiving information from IQCMs. Currently, IQCMs fill out paper forms as they observe the data collection within the classrooms and interview the test administrators and school coordinators. IQCMs are then asked to enter that data electronically at a later time and mail the hard copies of the paper forms to IEA or the ISC once

all of their duties have been completed. The current system adds additional steps and time to the process. A possible option for future studies would be to move the observation records to an electronic system that could be completed in real time while the IQCMs are in the schools. This would allow for better monitoring by IEA and the ISC and would cut down on the different steps IQCMs need to complete. It could also allow for more multimedia type information to be uploaded with the observation records, such as photographs of the testing facilities, being careful not to show actual test administrators, teachers, or students. Such technological advances could also be shared with countries for use during national quality control.

Quality control monitoring during data collection plays an important role in ensuring the overall validity and reliability of IEA data across studies. While few issues have emerged over the years, it is still important to continue to consider ways in which monitoring of data collection procedures can be streamlined or improved. These procedures can have a large impact on overall data quality and comparability. It is important that studies continue this type of monitoring to maintain confidence in the quality of IEA data.

References

Binkley, M., & Rust, K. (Eds.). (1994). *Reading literacy in the United States: Technical report of the U.S. Component of the IEA reading literacy study (NCES 94-259).*. Washington, DC: US Department of Education, National Center for Education Statistics, US Government Printing Office.

Bloom, B. S. (1969). *Cross-national study of educational attainment: Stage I of the I.E.A. investigation in six subject areas. Final report: Volume I*. US Department of Health, Education, and Welfare: Office of Education. Washington, DC: US Government Printing Office.

Cresswell, J. C. (2017). Quality assurance. In P. Lietz, J. C. Cresswell, K. F. Rust, & R. J. Adams (Eds.), *Implementation of large-scale education assessments* (pp. 193–204). Chichester, UK: Wiley.

Foshay, A. W., Thorndike, R. L., Hoytat, F., Pidgeon, D. A., & Walker, D. A. (1962). *Educational achievements of thirteen-year-olds in twelve countries: Results of an international research project, 1959-1961*. UNESCO Institute for Education. Hamburg, Germany. https://unesdoc.unesco.org/ark:/48223/pf0000131437.

Garden, R. A. (1990). *Second IEA mathematics study: Sampling report*. US Department of Education, Center for Education Statistics. Washington, DC: US Government Printing Office.

IEA. (2020a). Early IEA studies [webpage]. Amsterdam, The Netherlands: IEA. https://www.iea.nl/studies/iea/earlier.

IEA. (2020b). Other IEA studies [webpage]. Amsterdam, The Netherlands: IEA. https://www.iea.nl/studies/iea/other.

Johansone, I., & Wry, E. (2016). Quality assurance program for TIMSS 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 8.1–8.13). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/publications/timss/2015-methods/chapter-9.html.

Johansone, I., & Wry, E. (2017). Quality assurance program for PIRLS 2016. In M. O. Martin, I. V. S. Mullis & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 9.1–9.14). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-8.html.

Koršňáková, P., & Ebbs, D. (2015). Quality assurance of the ICILS data collection. In J. Fraillon, W. Schultz, T. Friedman, J. Ainley & E. Gebhardt (Eds.), *ICILS 2013 technical report* (pp. 127–142). Amsterdam, The Netherlands: The International Association for the Evaluation of Educational Achievement (IEA). https://www.iea.nl/publications/technical-reports/icils-2013-technical-report.

LaRoche, S., & Foy, P. (2016). Sample implementation in PIRLS 2016. In M. O. Martin, I. V. S. Mullis & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 5.1–5.126). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-5.html.

Leimu, K. (2011). 50 years of IEA: A personal account. In C. Papanastasiou, T. Plomp & E. C. Papanastasiou (Eds.), *IEA 1958–2008: 50 years of experiences and memories* (pp. 591–626). Nicosia, Cyprus: Cultural Center of the Kykkos Monastery. https://www.iea.nl/publications/iea-reference/iea-1958-2008.

Martin, M. O., Hoyle, C. D., & Gregory, K. D. (1996a). Monitoring the TIMSS data collection. In M. O. Martin & I. V. S. Mullis (Eds.), *Third international mathematics and science study: Quality assurance in data collection*, (pp. 3.1–3.12). Chestnut Hill, MA: Boston College. https://timssandpirls.bc.edu/timss1995i/TIMSSPDF/QACHP3.PDF.

Martin, M. O., Mullis, I. V. S., & Kelly, D. L. (1996b). Quality assurance procedures. In M. O. Martin & D. L. Kelly (Eds.), *Third international mathematics and science study (TIMSS) technical report volume I: Design and development* (pp. 11.1–11.12). Chestnut Hill, MA: Boston College. https://timssandpirls.bc.edu/timss1995i/TIMSSPDF/TRCHP11.PDF.

Martin, M. O., Rust, K., & Adams, R. J. (1999). *Technical standards for IEA studies.* Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA). https://www.iea.nl/publications/iea-reference/technical-standards-iea-studies.

Mullis, I. V. S., & Martin, M. O. (2018). 25 years of TIMSS and PIRLS. In S. Finlay (Ed.), *60 Years of IEA (1958–2018)* (p. 24). Amsterdam, The Netherlands: IEA. https://www.iea.nl/publications/60-years-iea-1958-2018.

Noveanu, G. N., Kobelt, J., & Köhler, H. (2018). Quality assurance procedures for the ICCS data collection. In W. Schultz, R. Carstens, B. Losito & J. Fraillon (Eds.), *ICCS 2016 technical report* (pp. 67–86). https://www.iea.nl/publications/technical-reports/iccs-2016-technical-report.

Postlethwaite, N. (1967). *School organization and student achievement: A study based on achievement in mathematics in twelve countries.* Stockholm, Sweden: Almqvist and Wiksell. https://ips.gu.se/english/Research/research_databases/compeat/Before_1995/FIMS/FIMS_Postlethwaite.

Purves, A. C. (2011). The evolution of IEA: A memoir. In C. Papanastasiou, T. Plomp & E. C. Papanastasiou (Eds.), *IEA 1958-2008: 50 years of experiences and memories* (pp. 531–556). Nicosia, Cyprus: Cultural Center of the Kykkos Monastery. https://www.iea.nl/publications/iea-reference/iea-1958-2008.

Schwille, J. (2011). Experiencing innovation and capacity building in IEA research: A memoir, 1963–2008. In C. Papanastasiou, T. Plomp & E. C. Papanastasiou (Eds.) *IEA 1958-2008: 50 years of experiences and memories* (pp. 627–708). Nicosia, Cyprus: Cultural Center of the Kykkos Monastery. https://www.iea.nl/publications/iea-reference/iea-1958-2008.

Statistics Canada. (2010). *Survey methods and practices.* Catalog number 12-587-X. Ottawa, Canada: Statistics Canada. http://www.statcan.gc.ca/pub/12-587-x/12-587-x2003001-eng.pdf.

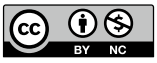
Wagemaker, H. (2011). IEA: International studies, impact, and transition. In C. Papanastasiou, T. Plomp & E. C. Papanastasiou (Eds.), *IEA 1958-2008: 50 years of experiences and memories* (pp. 253–272). Nicosia, Cyprus: Cultural Center of the Kykkos Monastery. https://www.iea.nl/publications/iea-reference/iea-1958-2008.

Weber, S. (2018). Sampling design and implementation. In W. Schultz, R. Carstens, B. Losito & J. Fraillon (Eds.), *ICCS 2016 technical report: IEA International Civic and Citizenship Education Study 2016* (pp. 43–52). Amsterdam, the Netherlands: IEA. https://www.iea.nl/publications/technical-reports/iccs-2016-technical-report.

Lauren Musu is a Senior Research Officer at IEA and has a background in educational psychology and statistics. She coordinates the International Quality Control Program for various IEA studies, which includes overseeing the program and developing materials for International Quality Control Monitors. She is also involved in data analysis and reporting for both internal and external purposes.

Sandra Dohr is a Research Officer at IEA and has a background in sociology and educational science. She is involved in managing International Quality Control Programs for various IEA studies, which includes recruiting and supporting International Quality Control Monitors and overseeing the implementation on a local level. In addition, she is involved in the coordination of Translation Verification procedures for IEA studies.

Andrea Netten is the Director of IEA Amsterdam. She has a background in child psychology and was the PIRLS NRC for the Netherlands for PIRLS 2006 until PIRLS 2016. She holds managerial responsibilities for the IEA headquarters and helps ensure that the responsibilities of IEA Amsterdam in international projects are met. She also maintains contact with IEA partners and responds to membership requests.



Chapter 9
Post-collection Data Capture, Scoring, and Processing

Alena Becker

Abstract The International Association for the Evaluation of Educational Achievement (IEA) collects large-scale assessment data from multiple countries and in many languages. Three important processes are data capture, scoring and coding, and data processing, which are supported by rigorous quality control mechanisms that IEA uses to ensure data comparability across countries. This chapter describes the steps that IEA takes to ensure the reliability of the data. Depending on the scope of the study, IEA uses different data capture systems during or after data collection. While data from multiple choice items can be captured directly, other data (e.g., responses from open ended questions) must be coded or scored uniformly to ensure comparability across countries and safeguard quality. Data processing starts only after data capture, scoring, and coding of data are finalized. As part of its data quality measures, IEA provides training for national representatives, instructional manuals, and study-specific software products to participating countries. Check protocols included in the software facilitate adherence to technical standards and minimize errors. Some data, while collected according to national conventions, ultimately needs to conform to IEA international formats. Moreover, the data must be assembled in a way that enables three linkages: schools with countries, teachers and classes with schools, and students with their teachers and parents. Data processing also detects and recodes any deviations from international formats or insufficient links between respondent levels.

Keyword Coding · Data processing · International database · Manual data capture · Online data collection · Paper data collection · Quality control · Reliability · Scoring

A. Becker (✉)
International Association for the Evaluation of Educational Achievement (IEA), Hamburg, Germany
e-mail: alena.becker@iea-hamburg.de

H. Wagemaker (ed.), *Reliability and Validity of International Large-Scale Assessment*,
IEA Research for Education 10, https://doi.org/10.1007/978-3-030-53081-5_9

9.1 Introduction

Reliability is critical when judging the adequacy and quality of an assessment. International assessments, because of their complexity and design, place additional demands on developers and test administrators to ensure that the concerns related to the reliability of the data are addressed. The assessments developed by IEA share many common features; the more complex assessments, such as the Progress in International Reading Literacy Study (PIRLS) and Trends in International Mathematics and Science Study (TIMSS), require response data to be captured from school principals, teachers, students, and sometimes parents or guardians. These data need to be captured systematically and accurately across multiple countries, multiple languages, and often, multiple populations within countries. Some data, while collected according to national conventions, ultimately need to conform to the international formats and have to be assembled in a way that links schools to countries, teachers and classes to schools, and students with their teachers and parents. Moreover, each student receives one specific cognitive instrument out of a pool of multiple instruments. To reduce the length of the cognitive instruments, items are arranged in blocks, which are rotated through multiple different instruments. To achieve a balanced number of responses across all items the cognitive instruments are evenly assigned to students by the IEA Within School Sampling Software. As IEA transitions from paper-and-pencil testing to computer-based assessment (CBA), data capture, processing, and scoring are evolving to reflect the new modalities. However, because the readiness of countries to adopt CBA differs, paper-and-pencil testing and CBA often have to operate in parallel.

To meet the challenge of ensuring high quality and comparable data, uniform methods of data capture and scoring have to be applied, first by country representatives at the national centers and later during the international data processing undertaken by IEA (Fig. 9.1). To facilitate these processes, IEA has developed training and standardized data capture and scoring procedures, together with specialized software; these include the IEA Data Management Expert (IEA DME), the IEA Windows Within School Sampling Software (IEA WinW3S), IEA Coding Expert, IEA eAssessment System, IEA Online SurveySystem, and the data processing programs used by IEA to process the data (Table 9.1). These procedures and software products are continuously reviewed to ensure the best quality possible.

An essential component of IEA training for participating national research coordinators (NRCs) involves an initial field trial that follows the same procedures as the subsequent main data collection. The field trial is a smaller version of the main data collection, collecting data from a small sample of respondents. It helps to identify any potential weaknesses in the procedures and rules, enabling these to be managed prior to the main data collection and thus further improving the reliability of the data management processes.

The transition from paper-and-pencil to online data capture began in 2004, with data from context questionnaires collected online via the IEA Online SurveySystem, software specifically designed for this purpose. As more countries expressed interest

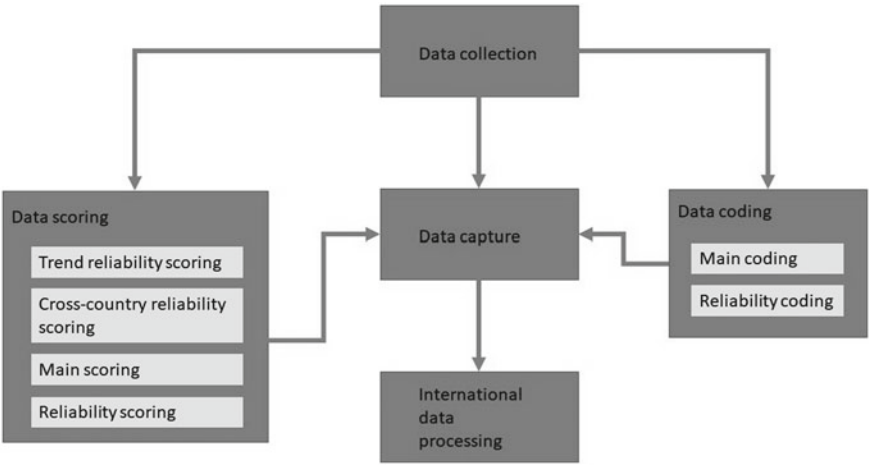


Fig. 9.1 Process overview of data capture, scoring, and processing

Table 9.1 Overview of IEA software used during and after data collection

Software	Purpose
IEA Windows Within School Sampling Software (IEA WinW3S)	Used to perform the within-school sampling of classes, teachers, and students according to study specifications. Tracks the participation status of respondents and monitors the response rates during administration
IEA eAssessment System	Used to create and translate context questionnaires and test booklets and to collect data online. Can be used to create paper instruments for paper administration
IEA Online SurveySystem	Used to create and translate context questionnaires and to collect data online
IEA Data Management Expert (IEA DME)	Used to manually enter data from paper instruments
IEA Coding Expert	Used to code and score open-ended questions
Windows Data Entry Manager (WinDEM)	Used to manually enter data from paper instruments

in transitioning from paper-and-pencil testing to CBA, IEA developed additional software to enable this (the IEA eAssessment System). The relatively new and exciting field of computer- and tablet-based testing has allowed item developers to construct more interactive items and to assess new sets of competencies.

However, the transition from paper-and-pencil testing to CBA and the expansion of item types present challenges for quality assurance, reliability, and comparability over time. Developers have to assess whether the different modes of data collection produce different results. For example, Fishbein et al.'s (2018) integrated mode effect

study for TIMSS 2019 demonstrated that the different modes of data collection in paperTIMSS and eTIMSS did not produce strictly equivalent results; they proposed modifications to the item calibration model used for TIMSS 2019 so that TIMSS trend measurements could be maintained.

Quality control occurs both at the national center and at the IEA. Some of the quality control procedures and checks that are implemented at the national center are repeated at the IEA using the same software and the same procedures. This ensures that the national quality control has been implemented as requested. Additional quality control procedures and checks with different software take place at IEA.

9.2 Manual Post-collection Data Capture and Management Training

For those assessments that rely on paper-and-pencil administration, data collected from respondents needs to be captured and transformed into an electronic format. All participating countries receive training from IEA on how to execute the data entry procedures at international data management training sessions. These training sessions are conducted twice per study cycle, once before the field trial and once before the main data collection. Training is intended to not only identify potential problems but also help the country representatives to familiarize themselves with the procedures and software. In addition, data entry manuals that are based on the IEA technical standards (Martin et al. 1999) supplement this training and provide an ongoing source of guidance for data entry. In these "train the trainer" workshops, mock materials are used in hands-on practice sessions and these are made available also for further in-country training of data entry staff that are organized subsequently by country representatives.

In order to meet the tight project timelines and to identify systematic data capture errors as early as possible, data capture starts during the data collection period. All sources of data, including the questionnaires and test booklets used, need to be stored in an orderly fashion to permit countries to consult original materials during data processing and verification procedures at the IEA. Any data inconsistencies that may arise (e.g., number of years worked at the school is higher than number of years worked in total) have to be manually verified to assure that they are not a result of mistakes during the data entry stage.

9.2.1 Data Capture from Paper-Based Instruments

Each country receives international versions of survey instruments and corresponding codebooks for data capture. When countries choose to or need to adapt some of the survey instruments to reflect national conventions, they need to reflect this in the

codebooks before entering data. To ensure that the codebooks follow the structure of the survey instruments, a test data capture of each survey instrument has to be done. Once the correct adaptation of the codebooks is verified, data entry can start.

The training provided and the common rules for manual data capture ensure that data is entered comparably by all personnel who are responsible for data entry. Data must be recorded exactly as listed in the survey instrument, no interpretation is allowed. When responses are omitted or cannot be interpreted, study-specific missing values have to be entered.

9.2.2 Software Used for Data Capture

IEA developed the Data Management Expert (DME) to improve the quality of data entry and standardize data capture procedures. The DME evolved from the previous Windows Data Entry Manager (WinDEM) system. The underlying structure file (codebook) ensures that all entered data is verified against all possible valid and missing values. Wild codes (codes that are not defined as valid or missing) or out-of-range values can either not be entered at all or have to be confirmed before they can be entered. At the beginning of the data processing, DME additionally checks that each record is entered only once, which decreases the chance of duplicate IDs. It also allows the adaptation of the structure files to account for any national deviations from the international survey instrument structure (national adaptation). In comparison to WinDEM, where multiple codebooks were needed and no checking between files was possible, the DME codebook can include all survey instruments and thus allows for additional consistency checks between these (e.g., it can be used to check if a student completed a context questionnaire but no data for a test booklet was entered). The checks are written in SQL (structured query language) and can be customized for each study.

9.2.3 Quality Control: Data Entry

To ensure data capture personnel are properly trained and adherence to the data capture rules, a certain proportion of survey instruments have to be entered twice; this is called "double punching."[1] A minimum agreement of 99% has to be reached for data to be accepted for submission. Countries are encouraged to enter data from paper survey instruments as early as possible during the data collection period so that possible systematic misunderstandings or mishandlings of data-capture rules can be identified quickly and appropriate remedial actions initiated, for example, further national center staff training.

[1]The term "double punching" dates from a time when physical punch cards were used to enter data; two cards were punched from the same data and were visually compared for any differences.

During and after data capture of paper survey instruments national center staff are required to run a number of checks using the IEA data capture software. This ensures that data submitted to the IEA for further processing fulfill the initial quality requirements.

The following checks are included in the DME software and NRCs are asked to run them regularly:

- Unique ID check: a unique ID check ensures that data for each respondent's questionnaire is entered only once;
- Validation check: a check validating all entered data against the structure files (codebooks); and
- Record consistency check: a number of study specific checks verifying data across different survey instruments.

Most countries capture their data manually using trained data entry personnel. When countries choose to scan their paper instruments, they are required to provide proof of their scanning reliability by scanning instruments twice and comparing the output. This corresponds to the double punching of manual data capture.

In addition, the IEA Windows Within School Sampling Software (WinW3S), which operates the participation tracking database, offers another set of checks that must be run before data is submitted. Depending on the study design, all or a subset of the following checks are available and must be undertaken. WinW3S allows NRCs to check whether:

- The data is available in a different administration mode than that currently entered in the WinW3S database (online versus paper administration);
- Participation, exclusion, or questionnaire return status in the WinW3S database matches the data availability in the DME database, the Online SurveySystem data tables, or the eAssessment database;
- The teacher subject code in the WinW3S database is consistent with the Teacher Questionnaire data in the DME database or the Online SurveySystem data tables;
- The assigned Booklet ID in the WinW3S database differs from the booklet entered in the DME database;
- Data exists for a booklet, but this booklet has not been assigned to a student;
- There are inconsistencies between the booklets assigned for reliability scoring and data availability for these booklets.

9.3 Scoring Cognitive Data: Test Booklets

The cognitive (achievement) test booklets consist of both multiple choice and constructed-response items. In order to allow testing of a larger number of items, they are usually grouped into different blocks that are then rotated in the test booklets.

Multiple-choice items are questions in which respondents are asked to select the correct answer from a list of different, often similar answers. Multiple-choice items can be machine scored and do not need to be evaluated by trained scorers.

Constructed-response items are questions that require students to provide a written answer, give a numerical result, complete a table, or provide a drawing. At the national centers, scorers trained by the scorers that attended the IEA international scorer training evaluate responses to these questions based on scoring guidelines provided by the international study center (ISC), which identify specific criteria for assigning a particular score. All country representatives are required to attend a scorer training session prior to the field trial and main data collection. At these training sessions all items, possible student answers, and different scoring codes are discussed with attendees. This ensures a common understanding and interpretation of the scoring guidelines.

9.3.1 Process of Scoring Constructed-Response Cognitive Items

The scoring teams within countries are divided into two groups, each with one professionally trained supervisor. The scoring supervisor moderates and answers all questions from scorers and reads a sample of scored responses to monitor the scoring reliability. This process is also called "back-reading" and essential for identifying scorers who do not understand particular scoring instructions. In such cases, scorers may need to be retrained or replaced when necessary.

A critical component of the scoring procedures is monitoring the quality of the scoring and calculating inter-rater reliabilities both within cycle and across cycles for linkage. In each cycle of any IEA international large-scale assessment, a certain study specific amount (ranging between 15 and 35%) of these items have to be scored by two different scorers. To simplify the administration process of reliability scoring of paper instruments, whole booklets (instead of single items) are randomly selected by the software that is also used to assign booklets to the students (IEA WinW3S). Within these booklets all constructed-response items are reliability scored. In order for the reliability scoring to be blind, the reliability scoring is completed first, with scores recorded on a separate scoring sheet and not in the booklets. The main scoring is completed after that, with scores entered in predefined fields in the booklets directly (see Table 9.2).

Table 9.2 Scoring responsibilities

Step	Team A	Team B
1	Scores reliability booklets from set B Records scores on separate reliability scoring sheets	Scores reliability booklets from set A Records scores on separate reliability scoring sheets
2	Scores all booklets from set A Records scores directly into the booklets	Scores all booklets from set B Records scores directly into the booklets

When scoring items that are available electronically, the IEA Coding Expert will display items without any previously assigned score. The assignment of items for reliability scoring does not have to be connected to whole booklets. The reliability items will be divided equally between two groups of scorers (team A and B) into set A and B. While scorer group A scores the items of set B, scorer group B scores the items of set A. When they are done, the sets of items are exchanged and scored a second time by the other group of scorers.

9.3.2 Software Used for Scoring Data

IEA has developed different software solutions to support the scoring process. The data capture software DME (and formerly WinDEM) provides two check procedures that help countries to assess the reliability of their scoring. One procedure checks if the same items for the same student are scored by two different scorers, while another compares the scores of the main scorer with that of the reliability scorer and provides the user with the agreement rate between the scorers. These two checks provide NRCs with information on two major quality requirements. Countries are urged to perform the checks continuously to identify as early as possible in the process whether any consistent misunderstandings exist among scorers. Should this happen, further training or replacement of scorers is necessary.

The IEA Coding Expert undertakes the same checks as the IEA DME application but offers additional features. When items are electronically available (e.g., scanned or imported from any CBA system), they can be displayed directly using the IEA Coding Expert. In this case, scorers enter their scores directly into the software and no additional manual data capture of scores is necessary. The IEA Coding Expert offers a scorer management system that creates unique logins for each scorer and assigns items for main and reliability scoring automatically to each scorer. Should scorers be trained to score only specific items, the IEA Coding Expert also considers this and assigns items accordingly. This ensures that the main scoring and reliability scoring of items are not undertaken by a sole scorer and also enables the scoring supervisor to monitor agreement rates in real time.

Before the development of the IEA Coding Expert, the IEA provided countries with separate software for the cross-country scoring reliability study (CCSRS) and the trend scoring reliability study (TSRS). These two software packages covered the same functionalities as the IEA Coding Expert. With the development of the IEA Coding Expert they became obsolete, since a single software package that can be used for all scoring tasks is more convenient.

Some constructed response items in the eTIMSS 2019 study were machine scored using SQL scripts. The machine scoring was undertaken after students completed the test, and not during the test.

9.3.3 Quality Control

The scoring quality is measured as the agreement rate between the main and reliability scorer (inter-rater reliability). An occasional error, or an understandable disagreement when the rules are not sufficiently clear is part of a judgmental scoring process and expected. However, consistent errors in categorizations arising from lack of understanding about the intent of the scoring guide are more serious. In such cases, all of the concerned constructed-response items need to be checked and corrected, not just the items selected for reliability scoring. If the error in understanding can be narrowed down to specific scorers, only items scored by those scorers need to be checked. Naturally, the goal is to have 100% or perfect agreement among scorers. An agreement between scorers above 85% is considered good and agreement above 70% is considered acceptable. Percentages of agreement below 70% are a cause for concern. The finally achieved agreement rate is also used during data adjudication, namely when reviewing data quality and making decisions about annotations for the reporting of data.

To ensure that there is continuity of scoring of the same items between the cycles of TIMSS and PIRLS, the TSRS has to be conducted by all countries that are participating in the study cycle that also participated in the previous cycle. The TSRS allows scorers of the current study cycle to score student responses collected during the previous cycle and is conducted using the IEA Coding Expert software. The responses are scanned by IEA following the previous cycle of the assessment.

All scorers who participate in scoring of the items of the current study cycle have to participate in the TSRS. Similar to the within-country reliability scoring, the TSRS blends with the main scoring procedure and is ongoing throughout the scoring process.

To verify consistent scoring between countries in the TIMSS and PIRLS studies, participating countries also need to perform a CCSRS. This gives an indication of how reliable the scoring is done across countries. Student responses included in the CCSRS are those related to items collected from English-speaking countries during the administration of the previous cycle of the assessment. Just like the TSRS, the CCSRS is conducted using the IEA Coding Expert software. The same set of student responses in English will be scored by all participating countries.

All scorers who participate in scoring of the items of the current study cycle should participate in the CCSRS. Similar to the TSRS, the CCSRS blends in with the main scoring procedure and is ongoing throughout the scoring process.

9.4 Coding Data

9.4.1 Process of Coding Data

Some IEA studies collect data on students' parental occupations using an open-ended text format. Before they can be internationally compared, they need to be transferred to an internationally comparable numerical code. For the coding of occupational data, IEA studies use a framework recommended by the International Labor Organization (ILO). The International Standard Classification of Occupations, ISCO-08 (ILO 2012) is a broadly accepted and accessible international classification of occupational data and a revised and improved version of its predecessor, ISCO-88, which was used in in previous study cycles.

Countries are encouraged to hire people with prior experience in this area of the coding. The coding team within countries is led by one supervisor. Supervisors review aggregated codes carefully and monitor coding progress. To further ensure high coding quality and common understanding of the coding rules, a study specific amount (ranging between 10 and 15%) of questions are coded twice.

Coding quality is proportional to the quality of student responses. The more detailed the information that students provide on their parents' occupations, the easier it is for coders to assign correct codes. Therefore, countries are advised to make administrators aware of the content of the questions that students will have to answer. If possible, students should be informed prior to the survey that these questions regarding their parents' occupation will be asked so that they have time to prepare their responses; if needed, schools should also explain to parents that they will need to help their children by providing the information that enables them to answer these questions. This would reduce the number of vague and omitted responses provided by students.

9.4.2 Software Used for Coding Data

To support the coding process, different software solutions have been developed by the IEA. The IEA data capture software DME supports two possible scenarios when coding questions. When questions are coded directly on the paper questionnaires and reliability coding sheets (occupation double coding sheets), the final ISCO-08 codes are entered directly into the codebooks with DME software. The software provides two check procedures that support countries in determining their reliability of the coding. One check procedure compares the values of the main and reliability codes and provides the user with the agreement rate between the coders. Another checks if the same question for the same student is coded by two different coders as required. These two checks provide NRCs with information on two major quality requirements at any time in the process.

Alternatively, it is also possible to enter the text responses from the occupational coding questions into the DME software. Once all data has been entered this way, the occupation data can be extracted to a specially designed Excel file that can be used for coding. The advantage of this procedure is that the Microsoft Excel file can be sorted by job title, which will make coding faster and more reliable. In the Microsoft Excel file the same check procedures are included as in the DME software.

Countries are urged to perform the checks continuously to identify any consistent misunderstandings among scorers as early as possible in the process and to initiate appropriate remedial actions when necessary, such as staff retraining.

The IEA Coding Expert covers the same checks as the IEA DME and offers additional features. When items are electronically available (e.g., scanned or imported from any CBA system) they can be displayed directly through the IEA Coding Expert. The items are coded in the software and no additional data capture of codes is necessary. The IEA Coding Expert offers a coder management system that creates unique logins for each coder and assigns questions for main and reliability coding automatically to each coder. This ensures that the same coder does not code both the main and reliability samples and it also enables the coding supervisor to monitor agreement rates in real time.

9.4.3 Quality Control

The coding quality is measured as the agreement rate between the main coder and reliability coder. An occasional error or an understandable disagreement when the rules are not sufficiently clear is part of a judgmental coding process. This is all part of the information about the coding reliability that is entered into the database. Consistent errors in classification of responses to occupational questions arise from misunderstandings. As for the other data coding activities, the goal is to have 100% or perfect agreement among coders. An agreement of more than 85% between coders is considered good and agreement above 70% is considered acceptable. Percentages of agreement below 70% are a cause for concern. The finally achieved agreement rate is also used during data adjudication, to review data quality and make decisions about annotations for reporting of occupational data. In total, about one sixth of the responses need to be double-coded. Reviews of coder discrepancies may indicate problems with the work of particular coders or a particular set of occupations that are difficult to allocate. In these cases, further training or the replacement of some coders may be advisable, thus IEA recommends coder agreement checks are undertaken at an early stage in the coding process. Where such data exist, it is also recommended that data are cross-checked against external/historical data sources.

9.5 International Data Processing

9.5.1 Processes in International Data Processing

The ultimate objective of data processing is to ensure the availability of consistent (reliable) and valid data for analysis. Once data collection within a country is completed and all data is available, participating countries submit their materials to the IEA. IEA then verify that all required materials that were sent are complete and fulfill all previously defined requirements. The receipt of materials is tracked in databases and confirmed to countries. In those cases where materials are missing, incomplete, or faulty, countries are contacted to resolve all issues. Materials are either sent via post, fax machine, email or uploaded to secure servers. Data collected during survey administration via IEA's servers (e.g., data from online context questionnaires) is already available and does not have to be resubmitted. If this is the case, countries have to confirm that data collection is finished and that access to online context questionnaires, the coding system, or the CBA system can be disabled. This ensures that no data is collected after the official data collection window and that data processing only starts when all materials are finalized. It remains, however, possible to reactivate access when requested by countries in agreement with the ISC to address data quality issues that may arise unexpectedly. This might be the case when there is reason to believe that additional time for data collection will improve low participation rates.

After all data are submitted to the IEA, final data processing commences. The objective of the process is to ensure that the data adheres to international formats, that information from respondents can be linked across different instrument files, and that the data accurately and consistently reflects the information collected within each participating country.

During the data processing, the IEA reports all inconsistencies that were found in the data back to countries to resolve all remaining issues. When all open issues regarding respondents' participation or reason for non-participation are resolved, the weighting of the sample data can start. Although the international sampling plan is prepared as a self-weighting design (whereby each individual ultimately has the same final estimation weight), the actual conditions in the field, non-response, and the coordination of multiple samples often make that ideal plan impossible to realize. To account for this, weights are therefore computed and added to the data.

Data submitted to the IEA come from a variety of different sources. Data for any individual respondent can come from any of the following: a tracking database, a database that is used for data entry and submission of paper questionnaires, a database that is used for data entry and submission of paper booklets, a database in which responses from the online questionnaires are stored, and a database in which responses from the online booklets are stored. The first step of the data processing, the data import, is to match-merge data from all sources. During this step, no specific checks are generated, but duplicate records are identified. Multiple data sources for a single respondent (i.e., those who responded to the same questionnaire on paper and

online, or those who responded to the same online questionnaires in two different languages) are flagged and need to be resolved. In most cases, the data pattern shows that a respondent changed their mind about their original choice of administration mode or language of questionnaire and very quickly abandoned their original choice to complete the questionnaire in another mode or language. In that case the data of the incomplete records are simply deleted. If, however, two equally complete data records with different data exist, the IEA asks the country to advise which record should be deleted. This issue of duplicate records is quite common for all studies, but usually affects only a few cases per country. Data processing can only continue with the next step once these issues of duplicate records are resolved.

In the next step after the data import, the structure of the national data is checked against the international data structure. The main goal of this is to document deviations from international standards both in data files and instruments, and to modify the national data according to the international design to ensure that the resulting data is internationally comparable. This is not only important for later analysis but also for all following cleaning steps, and ensures a consistent treatment of all data. National adaptations to the international survey instruments are agreed on and documented in special templates (national adaptation forms) before the data is collected. During the structure check, automated checks flag all deviations from the international format, which are then crosschecked against the adaptations documented in the national adaptation forms.

The main objective of cleaning is to identify any issues and deviations at the observation level. All deviations that appear for single observations or groups of respondents (e.g., students within one school) are reported. The data checks during the data processing can be divided into two major groups: ID and linkage cleaning, and background cleaning.

The automated checks of the ID and linkage cleaning compare the available data from the survey instruments with the reported available data in the tracking database. Checks ensure that the hierarchical ID system of the study was followed and thus a linkage between different respondent groups (e.g., students, parents, teachers, and principals) is possible, and that all tracking variables (e.g., student age and participation status) are assigned valid values and that the information in them is not contradictory to the actually available data from context questionnaires and test booklets.

The background cleaning checks verify the data in the context questionnaires. Depending on the item format certain data patterns can be expected. Answers to numerical questions are expected to fall within a certain range (e.g., student age), the sum of items of percentage questions (e.g., percentage of time the principal spends on the tasks in the school) is expected to equal 100%, the answers following a filter-dependent question are expected to be omitted when the filter question has been answered positively (e.g., does a school have a library vs. how many books are in the library), and the answer of logically dependent questions are expected to be consistent (e.g., enrollment of students in target grade is not expected to be greater than enrollment of students in the whole school).

The cleaning step produces findings for all checks. These findings can either be resolved with additionally available information (e.g., tracking forms or other supplementary documentation submitted by countries) or if the findings cannot be resolved directly, NRCs are contacted for further information and advice.

After all issues detected during the data processing steps are either resolved or confirmed by NRCs, the data processing commences with the next step, termed post cleaning. During post cleaning, the data undergo various major modifications.

To avoid discouraging respondents from answering questionnaires either incompletely or not at all, questions are never asked twice to the same respondent, For example, the teacher general background information would be the same when the mathematics teacher is also the science teacher. In those instances, data from overlapping questions between both questionnaires are copied over from one questionnaire to the other.

During post cleaning, a special missing code is assigned to questions that were deemed "not reached" to distinguish them from "omitted" responses. "Omitted" questions are those that a respondent most probably read, but either consciously decided not to answer or accidentally skipped; that is, the respondent started answering the questions but stopped answering before the end of the survey instrument, likely due to a lack of time, interest, or cooperation. "Not reached" responses are exclusively located towards the end of the survey instrument. To code as not reached, the last valid answer given in a survey instrument is identified. The first omitted response after this last answer is coded as omitted, but all following responses are then coded as not reached. Analyzing the frequency of the not reached code can give valuable information on the design of the survey instruments (e.g., length of context questionnaires, difficulty of test booklets). The not reached codes may be handled differently during the data analysis. In TIMSS, for example, the not reached items are treated as incorrect responses, except during the item calibration, where they are considered as not having been administered.

Not all processing steps necessarily lead to either automatic or manual recoding of data. Some data remained unchanged although a finding is reported for the affected records (e.g., contradictory/inconsistent filter usage). If a larger number of unresolved inconsistencies remains for a specific check, the data under consideration are carefully reviewed by IEA. For some cases, final actions or recodings during the post-cleaning phase, that is after all country processing feedback is taken into account, are agreed between the IEA and external stakeholders (such as NRCs and contractors).

In the final step, weights and scores that are computed based on processed data are merged with the processed data. All external data which is merged during the processing of the data undergoes its own rigorous checking procedures (see Chaps. 7 and 11).

The final check repeats all checks from the structure check and the cleaning checks. This is a quality measure to ensure that the data modifications during the post cleaning do not affect the data in any unintended way. However, the main goal of the final checks is to check the data that has been merged during the final merge.

The export step of the data processing produces data files in different formats. The data files produced during the export are frequently exchanged with partners, stakeholders, and countries. Data is available in various file formats (e.g., *.csv, *.sav, *.sas7bdat, *.dta) to facilitate further data analysis with different statistical software packages. Data can be exported at any time during the data processing period.

To protect the confidentiality of the respondents, certain disclosure avoidance measures are applied at the international level, which are consistent for all countries, and at the national level, which concern only specific national datasets. The most common measures across studies are scrambling of IDs and the removal of tracking or stratification variables. Measures at the national level can range from the removal of specific variables to the removal of complete datasets. Usually two versions of the international databases are created: a public-use file (PUF), available without any restrictions to any interested person, and a restricted-use file (RUF), available only upon special request by researchers. Researchers who want to use the RUF need to formally apply and this application is reviewed by IEA before access the restricted-use file is granted with requisite confidentiality rules. Unlike the PUF, the RUF includes confidential data (e.g., data identifying students' birth months and years).

9.5.2 Software Used for International Data Processing and Analysis

IEA uses three different systems to process all data from international studies. Regardless of the programs used, the processing steps and data checks are the same, and are system independent; differences are due to specific study designs and survey instruments. In theory, all three systems can be used for the data processing of all studies. While some are more powerful and convenient when processing large amounts of data, others are more easily adaptable by less technically-oriented staff. Finally, the timing of a study, the available budget, and the amount of data determine which cleaning program is used.

Originally, the data processing for all international studies was implemented using SAS (SAS Institute Inc. 2013). Due to the lack of readily available SAS programmers in many countries, IEA has also developed processing tools in SQL and SPSS (IBM Corp. 2017).

The international studies unit at IEA collects data processing requirements from all studies before they are implemented by programmers. This ensures that new developments between studies are exchanged and each new study benefits from the latest developments. When the data processing programs are developed, they are thoroughly tested using simulated data sets containing all the expected problems or inconsistencies. Providing data and programs in the different file formats maximizes the accessibility and utility of study data, further enhancing consequential validity.

9.5.3 *Quality Control*

To ensure that all procedures are conducted in the correct sequence, that no special requirements are overlooked, and that the cleaning process is implemented independently of those in charge, the data quality control process includes thorough testing of all data processing programs with simulation datasets containing all possible problems and inconsistencies. Deviations from the cleaning sequence are not possible, and scope for involuntary changes to the cleaning procedures is minimal. All systematic and individual data recodings are documented and provided to NRCs so that they can thoroughly review and correct any identified inconsistencies.

During the data processing, data is continuously exchanged with partners, countries, and any other stakeholders of the study. Before data updates are sent out, data are compared with previously sent data and any deviations are checked and verified. This ensures that only expected changes have been implemented in the data.

Univariate descriptive statistics are produced to help to review the content responses of the questionnaires. One file per sample and respondent level is created. Each presentation of univariate statistics provides the distribution of responses to each of the variables for each country.

9.6 Conclusions

Throughout the process of post-collection data capture, scoring, coding, and data processing, common quality control procedures ensure reliable data. These quality control procedures include a set of study-specific rules that all participating countries have to adhere to and customized software products that support both the NRCs and IEA in checking the adherence to these rules. This ensures that data collected by multiple countries, in multiple languages, and from respondents at different levels can be linked within countries and compared across countries.

At the conclusion of the study, IEA creates an international database; while most data is publicly available (the PUF), and the remainder is available only on request for formally approved research uses (the RUF).

References

Fishbein, B., Martin, M., Mullis, I.V.S., & Foy, P. (2018). The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-scale Assessments in Education, 6*, 11. Retrieved from https://largescaleassessmentsineducation.springeropen.com/articles/10.1186/s40536-018-0064-z.

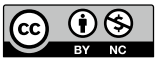
IBM Corporation. (2017). *IBM SPSS Statistics for Windows (Version 25.0) [Computer software]*. Amonk, NY: IBM Corp. Retrieved from https://www.ibm.com/analytics/spss-statistics-software.

ILO. (2012). *International Standard Classification of Occupations ISCO-08. Volume 1: Structure, group definitions and correspondence tables.* Geneva, Switzerland: International Labour Office. Retrieved from https://www.ilo.org/public/english/bureau/stat/isco/docs/publication08.pdf.

Martin, M. O. Rust, K., & Adams, R. J. (Eds.). (1999). *Technical standards for IEA studies.* Amsterdam, the Netherlands: IEA. Retrieved fromhttps://www.iea.nl/publications/iea-reference/technical-standards-iea-studies.

SAS Institute Inc. (2013). *SAS university edition 9.4 [Computer software].* Cary, NC: SAS Institute Inc. Retrieved from https://www.sas.com/en_us/software/university-edition.html.

Alena Becker is co-head of the International Studies Unit coordinating the data management activities for all international studies unit at IEA. Her position also includes supervising the international study centers at IEA Hamburg.



Chapter 10
Technology and Assessment

Heiko Sibberns

Abstract The International Association for the Evaluation of Educational Achievement (IEA) is moving toward computer-based assessment (CBA) in all its internationally conducted large-scale assessment studies. This is motivated by multiple factors, such as the inclusion of more comprehensive measures of the overall construct, increased use of online tools for instruction and assessment in IEA member countries, the need to develop assessments that are more engaging for students, and the hope that CBA will lead to increased efficiency and yield cost savings for participating countries. The transition to CBA has an impact on the design and procedures in both the international study centers and the national study centers. The very high standards in paper-based testing must be preserved throughout the transition steps. This relates not only to the layout, translation, and translation verification steps but also concerns the test administration and all associated procedures. The consistency, accuracy, and completeness of all collected data have to be achieve the same standards as paper-based testing, and the maintenance of the trend measures is of critical importance. The challenges arising from the transition to e-assessment are presented and discussed, including procedures developed to correct for any potential deviations that can be attributed to changes in the delivery mode. CBA provides potential new avenues for data analysis. Log-file and process data are of particular interest to researchers because such data provide an opportunity to increase the validity of the scales and minimize influences irrelevant to the construct being measured.

Keywords Computer-based assessment · Log-file data · Mode-effect · Process data · Transition from paper to computer assessment

H. Sibberns (✉)
International Association for the Evaluation of Educational Achievement (IEA), Hamburg, Germany
e-mail: heiko.sibberns@iea-hamburg.de

H. Wagemaker (ed.), *Reliability and Validity of International Large-Scale Assessment*, IEA Research for Education 10, https://doi.org/10.1007/978-3-030-53081-5_10

10.1 Introduction

Innovation and change has been an ongoing feature of the development of the International Association for the Evaluation of Educational Achievement's (IEA's) large-scale assessments. Nowhere has this been more evident than in the use and development of computer-based data capture and assessments. For the most part, this change has been triggered by changes in the environment. Internationally, there is greater use of technology for both learning and instruction. This coincides with the increased availability and use of computer technology outside of school. The transition to greater use of technology in assessment is based on a number of perceived benefits and initial successes and reflects the changing learning environment. However, there are still inequalities in information technology (IT) infrastructure between and within countries that need to be addressed if all envisaged benefits are to be achieved.

10.2 Technology in Education

In the majority of countries participating in IEA studies, instruction and assessment have gradually moved toward increasing use of technology in recent years (see, e.g., Bundesministerium für Bildung und Forschung 2019; UK Department for Education 2018). In many countries, ambitious goals for the provision of IT infrastructure in schools are not only formulated but large amounts of money have been made available to achieve them (see e.g., Bundesministerium für Bildung und Forschung 2019). As a result, there is a growing education technology industry. The central idea behind this is "that digital devices, software and learning platforms offer a once-unimaginable array of options for tailoring education to each individual student's academic strengths and weaknesses, interests and motivations, personal preferences, and optimal learning pace" (Herold 2016).

In addition to the provision of hardware and educational software that can be adapted to the needs of the teachers and students, a greater emphasis on the preparation of teachers in the effective use of technology is evident and demanded if the stated goals are to be achieved (again, see e.g., Bundesministerium für Bildung und Forschung 2019). Approaches that support the integration of private student computers (Herold 2016) have also advanced the use of the use of technology for learning and instruction.

IEA and its member countries recognized the need to make its studies available as computer-based assessments (CBAs). Today, whenever a new cycle of a study is presented at IEA's decision-making body, a computer-based version is immediately requested. Key arguments are based on major investment in IT infrastructure in schools, fears that a paper-based test would lack acceptance as the use of technology in schools advances, and a realization that students are increasingly using computers for

everyday tasks. However, among the diverse set of countries in which IEA operates, there are still a considerable number of countries that do not yet have the requisite infrastructure to implement CBA, and hence paper-and-pencil assessment is still important.

10.3 Promises and Successes of Technology to Reform Assessment Data Collection

Most programs for the introduction of CBA are accompanied by the promise to increase the efficiency of the survey. Other arguments are that the reliability and validity of the tests will be increased and new content domains can be explored.

10.3.1 Efficiency

CBA may lead to a decrease in costs. It is assumed that no printing and shipping of material to schools is required because the school can use their available resources for testing. Data entry is no longer required because student answers are directly entered into the computer. Finally, the scoring of constructed response items could be supported by machine scoring, provided the stored responses are comparatively simple, like numbers or fractions. However, gains in efficiency must also take into account changes in the overall cost structure and the redistribution of costs between national study centers and the international study center (ISC). When using paper-and-pencil testing, the participating country funds the necessary infrastructure and organization for many of the steps involved in the survey operations. For example, the layout of the national test instruments is undertaken within the country with the help of a desktop publishing program to eliminate deviations from the internationally-specified layout. The scoring of the constructed response items is also organized and carried out nationally. However, with CBA, the ISC must instead provide the technical infrastructure to revise the layout, as this requires programming knowledge and knowledge of the system used. The ISC must also provide the entire technical environment for scoring; this includes the display of all student responses, a tool for training the scorer, the assignment of the scorer to student responses, and a tool for monitoring the scoring process.

10.3.2 Increased Reliability: Direct Data Capture

When students enter their responses directly on the computer, this not only leads to a reduction in the costs of data entry but also leads to the elimination of expected

input errors. Errors that arise from illegible handwriting are also avoided, especially for constructed response items, since illegible items cannot be scored correctly or at all.

10.3.3 Inclusion of More Comprehensive Measures of the Overall Construct

A major advantage of CBA is the inclusion of more comprehensive measures of the overall construct. Aspects of the constructs that cannot be assessed via the traditional paper-and-pencil assessment are now accessible. Three examples illustrate the potential for enhancing measurement:

(1) The 2016 cycle of IEA's Progress in International Reading Literacy Study (PIRLS) included an electronic assessment (ePIRLS) of the reading competencies needed in the informational age, namely the non-linear reading skills needed to navigate webpages
(2) The 2019 cycle of IEA's Trends in Mathematics and Science Study (TIMSS) also included an electronic assessment (eTIMSS) that included innovative problem-solving and inquiry tasks (PSIs), extending the student assessment into areas that could not have been assessed by traditional paper-and pencil testing.
(3) The 2018 cycle of IEA's International Computer and Information Literacy Study (ICILS) included novel items measuring students' abilities in the domain of computational thinking.

10.3.4 Reading: Additional Competencies Needed in the Information Age

For ePIRLS 2016, CBA provided an engaging, simulated internet environment that presented grade 4 students with authentic school-like assignments involving science and social studies topics (see Mullis et al. 2017). An internet browser window provides students with a website containing information about their assignments, and students navigate through pages with a variety of features, such as graphics, multiple tabs, links, pop-up windows, and animation. In an assessment window, a teacher avatar guides students through the ePIRLS assignments, prompting the students with questions about the online information. Through this environment, non-linear reading of texts is introduced into the assessment.

The ePIRLS 2016 assessment consisted of five tasks, with each task lasting up to 40 min. Each student was asked to complete two of the tasks according to a specific rotation plan. The assessments were administered via computer (typically MS Windows-based computers) and students entered their answers by clicking on options or typing words. All input material for ePIRLS was tailor-made for the purposes of the study. The websites were hard-coded and therefore no templates had to be made available (Mullis and Martin 2015).

10.3.5 Mathematics and Science: Inclusion of Innovative Problem-Solving Strategies

To extend coverage of the mathematics and science frameworks, eTIMSS 2019 included additional innovative PSIs that simulated real-world and laboratory situations where students could integrate and apply process skills and content knowledge to solve mathematics problems and conduct scientific experiments or investigations. PSI tasks, such as designing a building or studying plant growing conditions, involve visually attractive, interactive scenarios that present students with adaptive and responsive ways to follow a series of steps toward a solution. Early pilot efforts indicated that students found the PSIs engaging and motivating (Mullis and Martin 2017). These PSIs provide an opportunity to digitally track students' problem solving or inquiry paths, and analysis of the process data, which reveals which student approaches are successful or unsuccessful in solving problems, may provide information to help improve teaching (see IEA 2020a).

It should be emphasized that the demanding criteria for PSIs made their development very challenging and resource intensive. Special teams of consultants collaborated both virtually and in person to develop tasks that: (1) specifically assessed mathematics and science ability (and not reading or perseverance); (2) took advantage of the electronic ("e") environment; and (3) were engaging and motivating for students (Mullis and Martin 2017).

10.3.6 Computational Thinking: Developing Algorithmic Solutions

> Computational thinking refers to an individual's ability to recognize aspects of real-world problems which are appropriate for computational formulation and to evaluate and develop algorithmic solutions to those problems so that the solutions could be operationalized with a computer (Fraillon et al. 2019, p. 27).

Following this definition, the computational thinking test modules in ICILS 2018 contained tasks developed to make use of elements originating in visual programming languages. Students were able to arrange blocks of code to solve real-world problems without knowing a particular programming language. More broadly, ICILS was developed to specifically measure computer skills. So CBA was essential in this case, and the process and sequencing data are of special interest. ICILS measures international differences in students' computer and information literacy (CIL): their ability to use computers to investigate, create, participate, and communicate at home, at school, in the workplace, and in the community. As mentioned, participating countries also had an option for their students to complete an assessment of their computational thinking (CT) ability, and approaches to writing software programs and applications. ICILS 2018 was administered by USB (universal serial bus) stick and local server mode (Fraillon et al. 2019).

10.3.7 Increased Reliability: Use of Log-File Data

The extended potential of computer-based testing is that the data record not only the student answers but also capture additional information about *how* students achieve those answers, such as their navigation behaviors, information about the time taken to reach an answer, or which tools they used (ruler, compass, or screen magnifier). This "log-file" data can be analyzed to reveal more about test-taking behaviors and strategies.

In an overview article, von Davier et al. (2019) outlined many of the options available for the additional use of log-file data. Timing information can not only be used to increase the reliability and validity of scales but also to discover rapid guessing or disengagement. When dealing with complex problem-solving tasks, the analysis of the processing steps and sequences makes it possible to distinguish successful processing strategies from less successful ones.

Ramalingdam and Adams (2018) showed that the reliability and validity of the scale in the field of digital reading can be increased by using log-file data. In the context of their study, only multiple choice items were considered. Log-file data helped to distinguish students who read the clues that were necessary to answer the item correctly from students who only guessed.

10.3.8 Development of More Engaging and Better Matching Assessments

More interactive test material and new item types that cannot be used on paper are perceived as an advantage of a CBA system because of the likelihood of increased student engagement. However, in cases where students are being assessed using test material that is also administered on paper, care must be taken to ensure that the CBA environment does not have a different impact on student performance; this is known as a mode effect (Walker 2017). It is assumed that, ultimately, the implementation of "adaptive testing" will enable student ability to be better matched to item difficulty, increasing student engagement, reducing student frustration, and, consequently, reducing the amount of missing data. Overall, this should result in more accurate assessment of student ability. However, like other parallel forms of test administration, mode effects need to be carefully monitored to ensure measurement invariance. Another potential advantage of CBA lies in the option to increase the accessibility of the tests by making supporting tools available for students with special needs. Many such opportunities are under investigation for implementation, such as applications for magnifying texts or graphics, or options to increase contrast between text and background. Text-to-speech solutions are another option; however, such initiatives have consequences for the test administration as earphones must be available if this option is used and all text elements in all languages would have to be available as voice recordings, which would be a considerable additional effort.

10.4 The Transition

10.4.1 Delivering Questionnaires Online

Before developing computer-based applications for the assessment of student achievement, IEA tackled the web-based administration of the background questionnaires. Since the content of the background questionnaires differs from the student assessment, the protocols that have to be followed for their administration are not as strict because the security requirements for the protection of questionnaire items are less critical. There is also no need for a controlled and standardized computer environment, thus the first step was to develop a comparatively soft system for surveying teachers, school leaders, and parents. The first system used for background questionnaires was the IEA Survey System, which was developed in 2004 and first used as part of the Second Information Technology in Education Study (SITES) in 2006 (Carstens et al. 2007). This was later renamed the Online Survey System (OSS). To be applicable in an international environment, the OSS had to provide adaptable questionnaires that enabled the necessary cultural adaptations (e.g., use of different metrics), and accommodated the addition or deletion of answering options. When, for example, asking about school types, certain countries may have a limited amount of school types while others, like federal systems, may have a larger variety. Equally, in other cases, questions may not be applicable and can therefore be deleted. National research programs may also wish to add optional national questions to the international questionnaire.

As questions need to be translated and the translations need to be verified, a tool to allow for adaptations, translations, and translation verification had to form part of the survey system, or, at the very least, the necessary interfaces had to be available. The system also had to work in a variety of cultures and for right-to-left languages, as well as languages with additional character sets.

Interfaces had to be programmed to guarantee the seamless integration of respondents' answers. Metadata, like codebook information from the IEA's existing data processing environment, also had to be included. Finally, the content had to be a standard hypertext markup language (HTML) format so there were no limitations on browser displays.

The OSS was always a purely online solution and did not support offline options like USB-delivery or local server mode. For this reason, from the very beginning, the amount of data that needed to be transmitted, in both directions, was kept to a minimum. Another prerequisite was that the OSS had to be easy to use, both in setting up the system and in enabling the adaptations and translations required by countries. OSS is still used for IEA studies, but ongoing maintenance and modifications to permit proper content display on current browsers means that new solutions are always under investigation. Over the years, OSS has proved to be a reliable and robust tool for carrying out surveys; slow internet connections or outdated browsers have not presented a problems. However, as technology as evolved the design has become increasingly dated, and complex surveys with many filter questions and branches require more modern solutions.

10.4.2 Computer-Based Assessment

IEA's move toward CBA has been gradual and careful. The OSS was initially developed for background information only, followed by the development of an assessment system supporting the assessment of web-based reading in 2014 for PIRLS 2016 and, shortly afterward, a full electronic assessment (eAssessment) system for TIMSS 2019. ICILS was based on a national study conducted in Australia. Consequently, ICILS international implementations in 2013 and 2018 made use of the CBA system that was developed for the Australian national study.

ePIRLS. Because internet reading is increasingly becoming one of the central sources that students use to acquire information, in 2016, PIRLS was extended to include ePIRLS, an innovative assessment of online reading (Mullis and Martin 2015). The main characteristics are non-continuous texts, such as those found on websites on which content is linked via hyperlink. These types of text were presented to the students in a simulated internet environment.

eTIMSS. TIMSS 2019 continued the transition to conducting the assessments in a digital format. eTIMSS provided enhanced measurement of the TIMSS mathematics and science frameworks and took advantage of efficiencies provided by the IEA eAssessment systems. About half the countries participating in TIMSS 2019 transitioned to administering the assessment via computer. The rest of the countries administered TIMSS in a traditional paper-and-pencil format, as in previous assessments.

To support the transition to eTIMSS, IEA developed eAssessment systems to increase operational efficiency in item development, translation and translation verification, assessment delivery, data entry, and scoring. The eTIMSS infrastructure included the eTIMSS Item Builder to enter the achievement items, an online translation system to support translation and verification, the eTIMSS Player to deliver the assessment and record students' responses, an online Data Monitor to track data collection, and an online scoring system to facilitate the national study centers' work in managing and implementing scoring of students' constructed responses.

eTIMSS also included new digital ways for students to respond to specific constructed response items, which enabled student responses to many items to be scored by computer rather than by human scorers. In particular, a number keypad enabled students to enter the answers to many constructed response mathematics items so that the answers can be computer scored. Other types of constructed response items that can be computer scored use drag-and-drop or sorting functions to answer questions about classifications or measurements.

eTIMSS was administered via USB delivery and through local server mode. Originally, the plan envisaged delivery via tablet devices, and it was expected that students could write down their answers with styluses. However, first trials with grade 4 and grade 8 students revealed that the interaction with the tablet device proved overly complicated and the stylus approach was dropped. On tablets, students could instead activate built-in keyboards and, since a significant number of countries had already invested in laptop computers for assessment programs, the

tablet approach was extended and delivery using personal computers (typically MS Windows-based computers) was implemented where students could make use of the physical keyboard and the mouse. Drawings were supported by a line drawing tool, and calculations through a built-in calculator (Mullis and Martin 2017).

ICILS. ICILS 2013 evaluated students' understanding of computers and their ability to use them via an authentic CBA of grade 8 students. ICILS 2018 was linked directly to ICILS 2013, allowing countries that participated in the previous cycle to monitor changes over time in computer and information literacy (CIL) achievement and its teaching and learning contexts (Fraillon et al. 2019). A third cycle is planned for 2023 (IEA 2020b).

10.5 Challenges

Regardless of the potential and possibilities of computer-based testing, there are many challenges and problems at all levels that need to be solved and which demand careful handling. Some of the main challenges are briefly considered here.

Most international large-scale assessments (ILSAs) generate trend measures enabling assessment results over time to be compared. Scores reported over time are established using the same metrics and represent valid and reliable measures of the underlying constructs; new scaling approaches may provide additional information. Differences in the properties of the items must be carefully considered. Items that show different behavior in the paper and the computer modes may have to be treated as different items, even if they are apparently identical (von Davier et al. 2019). The mode effect challenge remains an ongoing issue as the various assessments transition to become computer based. IEA continues to make considerable investments to ensure these risks are mitigated in all phases of the study implementation.

At the design stage, before significant resources are spent in the implementation of a CBA, one single small group of students is asked to work on practice versions of the items and "think aloud" when working through them. These qualitative student reactions are reported and analyzed. They inform the graphical design of the entire assessment system, as well as the arrangement of graphical and text elements in individual items. Equivalence studies, such as those undertaken in the development of eTIMSS 2019 are critical (see Fishbein et al. 2018). In the year prior to the field trial for eTIMSS 2019, participating countries were invited to administer certain trend items in both paper and electronic format. Students received two test forms, one was computer based, and the second was paper based. It was assumed that both tests measured the same construct and the differences in item behavior were only marginal. However, results showed that, while the same item administered in the computer environment and on paper showed similar behavior, students found items administered on computers slightly harder than the same items on paper. The effect was more prominent for mathematics items than for science items, and, although small, could not be dismissed. Thus, in order to safeguard the trend measure, each participating country had to take what was termed a paper bridging study. In each

eTIMSS country, a smaller sample of 1500 students took a paper test based on the trend items to establish the trend differences for each country. These paper tests were analyzed with the standard calibration methods established for previous cycles of TIMSS (see chap. 13 of Martin et al. 2016). Finally, the computer-based test scores were adjusted (so that they could be aligned with the trend results established by the paper bridging study).

Layout adjustments that are necessary after the tests have been translated are a particular challenge. They result when the translated text is longer than the original English text or when it contains different alphabets.

Graphics often contain text elements such as labels in coordinate systems or frequency tables, or descriptions in maps. All text elements must be translated and then indicated so that the graphic will continue to be correctly labelled. Since some languages require significantly more characters in translation than the English original, so dynamic adaptation of the text fields has to be supported. In extreme cases, manual adjustment is essential.

Often, text in the translated language is not displayed correctly because the space provided for it in the assessment system was too small and had to be expanded afterwards. In principle, this is also a problem in paper-based testing, but is relatively easier to solve when the commercial layout software used in paper-and-pencil design provides tools that enable quick and uncomplicated adjustments

Frequently, text boxes not only contain plain text but also graphics in the form of special mathematical symbols, currency symbols, or special symbols from the natural sciences. In part, these symbols must also be "translated" when used in another language, for example, in the English original the "/"is used as a sign for the division, while in German "÷" is used.

In particular, right-to-left (Hebrew and Arabic) written language systems, as well as some Asian languages, present a major challenge. Discussion of all possible special cases would go beyond the scope of this chapter. Suffice to say, a mathematical expression in a sentence is written in Arabic from left to right, while the rest of the sentence is written from right to left. Thus when writing "What is the value of x in the equation $x + 12 = 15$?" in Arabic it will appear as:

ما هي قيمة x في المعادلة $x + 12 = 15$؟

Irrespective of the language used, geometry items pose a particular challenge. In the paper version, geometrical figures must be constructed using a ruler or compass. A ruler is also needed to determine distances. When such items are transferred to a CBA system, technical tools must also be provided to allow the required operations to be performed. The use of the aids must be explained and practiced in an introductory section prior to the test so that students taking the test understand exactly how to activate the ruler and how it has to be handled to take the required length measurement.

A fundamental decision must be made as to which devices and which associated operating systems should be allowed to run the assessment. The planned assessment material and the expected interactions determine the type of devices. Screen size, screen resolution, the decision for or against touch screen, the type of keyboard (pop

up or physical keyboard) and the use of mouse or touchpad are set after the decisions regarding the assessment are taken.

For eTIMSS 2019, supporting tablet, laptop, and personal computer delivery turned out to be extremely demanding. Different technologies had to be used for Android®-based tablet devices and Windows®-based computers. As a consequence, digitalPIRLS was programmed only for personal computers. When using tablets, frequent updates of the operating system might pose a risk for the data collection. It may happen that accesses of the test software to certain system components of the operating system no longer work after an update and the assessment software stops functioning. The problem can be mitigated by resetting updates or by adapting the test software.

A challenge that is pertinent to tablet delivery is the interaction with the assessment system. In the initial phase of moving to CBA in eTIMSS, it was envisaged that student might use a stylus to record their answers on the tablet screen. However, early trials indicated that students were not used to this form of electronic interaction and the idea was quickly abandoned. Pop-up keyboards and number pads were implemented instead. For digitalPIRLS, physical keyboards are mandatory.

In a best-case scenario, the necessary (number and kind) resources are available in all schools and the national research coordinator (NRC) conducting a study has permission to use those resources. Even in this situation, schools need to be contacted prior to the testing in addition to the contacts necessary for the organization of students' testing. A diagnostic tool needs to be run on all the computers that the school plans to use for the test. The diagnostic tool checks screen resolution, available memory, processor speed, and other parameters that need to be met in order to run the assessment program smoothly. This diagnostic tool also has to reflect the delivery mode that is planned for a specific school. The assessment can only be conducted in the school if all machines pass the checks, and it may be that individual computers fail the requirements; in this situation the NRC has to devise a backup solution, like providing carry-in laptops or splitting test groups in schools.

Often, however, there may be no school resources available or access to them might be limited or denied because of school security policies. In this situation, laptops need to be delivered to the school purely to conduct the assessment, and, in most cases, this would impose a financial burden on national study centers because they would have to be purchased or rented for the assessment. This is generally only a viable approach if the computers are to be used regularly, for example for other international assessments or national testing programs. Such machines need to be especially robust because they have to be shipped to schools, and should be equipped with up-to-date hardware that allows a reasonable duration of use. For laptop delivery, a reliable and detailed logistic system has to be established in order to have a minimum number of laptops distributed among test administrators in a well-planned order.

Even assuming that equipment in most schools is ready for the type of CBA programs used in ILSAs, indications from NRCs are that schools tend to be more restrictive due to security policies. A further obstacle to the use of school resources is the paucity of dedicated information technology administrators or coordinators

within schools; these tasks are frequently outsourced to external contractors. Thus, changes to school contracts to grant permission to run an external assessment software often come with additional costs for schools and are therefore rejected. But the situation is constantly evolving, and there is evidence that countries are investing more in computer technologies in schools, including additional staffing (see for example, Bundesministerium für Bildung und Forschung 2019; UK Department for Education 2018).

Whatever situation is found, manuals have to be adapted to reflect the changes due to CBA. Unlike paper-based assessment, where paper instruments are distributed according to a strict protocol as manifested in the test administrator manual, computers have to be set up and prepared. The need for diagnostic tests of all computers to ensure proper functioning prior to the assessment has already been mentioned. On the test day, schools need to ensure that the testing software is either installed or accessible from all machines. When laptops are brought into the school, sufficient and properly fused power lines have to be available, and all computers have to start. Depending on the quality of computers, this could last for some time and may justify the presence of a second test administrator responsible for the proper functioning of the technology. This person could also be responsible for remedying problems. Students have to be provided with login and password information, and test administrators have to make sure that all students apply the information correctly.

During the first assessments making use of CBA, IEA noted shifts in the study cost structures. All processes that were needed in the paper-based assessment and that were, to a large extent, executed on paper or in word processors or layout programs, needed to be reproduced in a computer environment (see Sect. 10.3.1). The modules had to be developed, maintained, and updated or replaced if changes in computer technology required these changes. In addition, a move towards web-based delivery has significant demands on the system in use to guarantee confidentiality, integrity, and availability. This requires redundant systems. Identical server solutions around the world are needed to cope with delays due to long distances between the test taker's device and the server on which the data is stored and the system also needs to cope if significant numbers of respondents are accessing the system simultaneously when CBA is set up on a large scale and makes use of web-based testing. A shift of costs towards the international operations need to be properly accounted for in a revised fee structure. Conversely, significant savings can be expected at the country level because the central infrastructure is kept and maintained under the supervision and responsibility of the ISC.

10.6 The Future: Guiding Principles for the Design of an EAssessment Software

Based on IEA's experience with a variety of programs, including IEA's propriety software, the following principles can serve as a guideline for developments that secure high quality software components in the future.

10.6.1 Adaptive Testing

In adaptive testing, test tasks or blocks of tasks are put together during the test session and then presented to the student so that the maximum information about the student's performance is obtained. This provides a very precise and reliable estimate of student performance. In summary, it can be said that the difficulty of the test obtained in this way corresponds to the ability of the student and that the boundary conditions for the content of the test are also met. The possibility of adaptive testing must also be considered in future developments of ILSA (Wainer 2000). Different methods are available, starting with group adaptive testing where tests of different difficulty are created and groups of pupils of different ability are identified from available data before the test administration; student groups then focus on the test most appropriate to their abilities.

In a second form of multi-stage testing, each student initially receives a test of medium difficulty. The test is evaluated immediately in the CBA system. Depending on how well they managed the first test section, the student then subsequently receives a test section that may be more closely tailored to their ability; less challenging if the student had difficulty, or more challenging if the student achieved good results. This is also again evaluated immediately and another test section assigned depending on the previous result.

Finally, in the case of a fully adaptive design, the presentation of test items is a function of their success or failure on previously presented items. After each item has been processed, student skill is estimated and a new item of appropriate difficulty assigned.

All these test methods have in common that a lot of item material has to be developed and calibrated. In particular, many boundary conditions, such as a given distribution of the cognitive domains, the content domains, or the item types, must be taken into account when assembling a test form on-the-fly for particular students during the test session.

10.6.2 Translation

The translation system makes it possible to reliably translate the international survey instruments. Since many translators work with professional translation software that provides functionalities beyond what can be provided in an assessment system, it is necessary to implement a common interface to other systems. The XLIFF (XML localization interchange file format) is an XML-based format created as a common exchange format for translations and is also supported in IEA's propriety software (OASIS 2008). Further information about translations and the challenges created by a CBA system can be found in Chap. 6.

10.6.3 Printing

Very desirable is a functionality that allows the creation of paper instruments from the CBA system because items in common for both assessment modes are then maintained in only one system and an item need only be changed once avoiding deviances between items delivered on paper and on computer. Previous attempts for a paper export have not been satisfactory, although, as a rule, all text elements could be displayed correctly. Graphics, however, were often out of focus. Then, when the resolution of the graphic was increased, too much data load was generated for the computer-based delivery. Nonetheless an export function remains an important goal because the same assessment may be carried out both on paper and delivered by computer until all countries are able to make the step towards computer-based testing. With such an export function, the test development could take place in one environment only and, as well as avoiding inconsistencies between the paper and the computer environment, the additional costs for a double input with associated quality control would not be incurred.

10.6.4 Web-Based Delivery

A web-based test would have the great advantage that all information can be available in real time for both the national study centers and the respective ISC. No programs need to be installed and, given the webpages created are not too demanding, any web browser would work. If problems occured during test application, the change would be made only once, and a time-consuming and logistically difficult new rollout on USB sticks or server notebooks would be superfluous. If a suitable monitoring tool is available, the course of the study can be closely monitored and, if necessary any irregularities addressed, for example, if participation rates were lower than expected. It would also be possible to react to delays or irregularities in the sequence, such as an accumulation of tests in a certain period of time in the test window; the testing

sessions could be rearranged and spread more evenly across the testing window. A connection to a logistics system is also conceivable, with which the use of test administrators could be planned, controlled, monitored and, if necessary, accounted for.

This approach is only viable, if sufficient bandwidth for data transmission is available. Considerable efforts may be necessary to secure the availability, integrity, and confidentiality of a web-based delivery system: the test must be available at the required time during the testing session in the school, confidential information must be protected from unauthorized access, and the data must be complete and unchanged when transmitted or during data processing. Unlike a USB delivery, a system failure would cause massive damage because a malfunction could potentially endanger the entire assessment; a corrupted USB stick also leads to loss of data, however, only a single student is affected.

10.6.5 General Considerations

In all IEA's ILSAs, constructed response items are used. A significant number of these items have to be scored by people, as only short answers, such as intact numbers or fractions, can be machine-scored. For the scorers, it is important that they see not only the isolated response of the student but also the response embedded in the page the student has answered. The displayed page must be based on the translated version, as it is the only way to detect inaccuracies in the translation that influence the student's response.

In any future developments, the following factors should be considered. CBA systems and workflows should work under secure conditions, include a reasonable backup strategy, and fulfil data protection requirements. Countries with special data security concerns should be handled appropriately (e.g., through data encryption, as implemented already in various systems). What can be accessed within the different eAssessment system modules and what actions a user can and cannot undertake depends on the different user roles. Participating countries should only be able to access the version that was generated for the specific country and their own data; they should not be able to see data from other countries. CBA environments need to show only the specific study-related parts if countries and ISCs access the different modules and they should not provide access to data from other studies.

The designer tool should ideally be able to integrate third-party tools like GeoGebra (a free online mathematics tool; https://www.geogebra.org/) or import items that were generated using third-party programs and stored according to agreed industry standards (such as the IMS Global Learning Consortium's question and test interoperability (QTI) format, which defines a standard format for the representation of assessment content and results, supporting the exchange of this material between authoring and delivery systems; http://www.imsglobal.org/). The designer tool should also be used as an item bank, tracking items that were used in different studies. Items should be transferable between study cycles, and between one study

and another, as needed, especially questionnaire and released assessment items. The item designer tool should be able to track the item development workflow.

IEA has used three delivery modes for CBA in their studies. It is advisable to support at least two modes in a CBAs because conditions in countries vary considerably in terms of IT-infrastructure like hardware, connectivity to the internet, bandwidth, or access rights. Web-based delivery was described in Sect 10.6.4. Two other modes were USB delivery and local server mode:

In USB mode, the entire testing system, including the delivery and storage system, and any necessary browsers are stored on a USB stick, and the assessment is run from the stick. This technology is advisable when computers in schools do not have a solid internet connection or when there is only restricted bandwidth capacity for data transmission. Computers need permission to access USB devices and run programs from USB sticks. One disadvantage is the logistics necessary to implement the USB approach. USB sticks need to be duplicated and shipped to test administrators. After testing, the data need to be uploaded to a server.

In local server mode, the testing program is delivered via a local server that is located in the school. Computers need to be connected to the server, either through a wired network or through a wireless local area network (Wi-Fi), and a version of the test program supporting the local server mode needs to be installed. The test itself is run on the client server via a web browser. For this approach, access rights need to be granted and the browser that has to display the assessment needs to meet the standards defined by the study. Upon completion of the test, the data are uploaded by the test administrators.

10.7 Conclusions

The change from paper testing to computer-based assessment has far reaching implications for test design, implementation, and analysis, including test and item development, downstream procedures such as translation and layout control, test execution, scoring, data processing, and analysis. They are accompanied by changes in the workflow and the cost structure. The constant that remains is the importance to maintain the very high quality standards achieved in the paper-based test, and to improve the technology as far as possible. In particular, the preservation of trend metrics and the security of assessment material are of paramount interest. With the systems already in use, IEA has succeeded in keeping the path to computer-based testing on track without sacrificing quality standards, by introducing changes gradually and deliberately with appropriate monitoring of mode effects.

References

Bundesministerium für Bildung und Forschung. (2019). *Mit dem Digitalpakt Schulen zukunftsfähig machen* [webpage]. Berlin and Bonn, Germany: Author. Retrieved from https://www.bmbf.de/de/mit-dem-digitalpakt-schulen-zukunftsfaehig-machen-4272.html.

Carstens, R., Brese, F., & Brecko, B. N. (2007). Online data collection in SITES 2006: Design and implementation. In *The second IEA international research conference: Proceedings of the IRC-2006, Volume 2: Civic Education Study (CivEd), Progress in International Reading Literacy Study (PIRLS), Second Information Technology in Education Study (SITES)* (pp. 87–99). Amsterdam, the Netherlands: IEA. Retrieved from https://www.iea.nl/index.php/publications/conference/irc-2006-proceedings-vol2.

Fishbein, B., Martin, M., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-scale Assessments in Education, 6*, 11. Retrieved from https://largescaleassessmentsineducation.springeropen.com/articles/10.1186/s40536-018-0064-z.

Fraillon, J., Ainley, J., Schulz, W., Duckworth, D., & Friedman, T. (2019). *IEA International Computer and Information Literacy Study 2018 assessment framework*. Cham, Switzerland: Springer. Retrieved from https://www.iea.nl/publications/assessment-framework/icils-2018-assessment-framework.

Herold, B. (2016, February 5). Technology in education: An overview. *Education Week*. Retrieved from http://www.edweek.org/ew/issues/technology-in-education/.

IEA. (2020a). *TIMSS 2019: Trends in International Mathematics and Science Study 2019* [webpage]. Amsterdam, the Netherlands: IEA. Retrieved from https://www.iea.nl/studies/iea/timss/2019.

IEA. (2020b). *ICILS 2023 flyer*. Amsterdam, the Netherlands: IEA. Retrieved from https://www.iea.nl/publications/flyer/icils-2023-flyer.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). *Methods and procedures in TIMSS 2015*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/publications/timss/2015-methods.html.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2015). *PIRLS 2016 assessment framework* (2nd ed.). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/pirls2016/framework.html.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. http://timssandpirls.bc.edu/timss2019/frameworks/.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). Take the ePIRLS assessment. In I. V. S. Mullis, M. O. Martin, P. Foy, & M. Hooper (Eds.), *ePIRLS 2016 International Results in Online Informational Reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://pirls2016.org/epirls/take-the-epirls-assessment/.

OASIS. (2008). *XLIFF version 1.2*. Retrieved from http://docs.oasis-open.org/xliff/xliff-core/xliff-core.pdf.

Ramalingdam, D., & Adams, R. J. (2018). How can the use of data from computer-delivered assessments improve the measurement of twenty-first century skills. In: E. Care, P. Griffin & M. Willson (Eds.), *Assessment and teaching 21st century skills. Research and applications* (pp. 225–238). Cham, Switzerland: Springer.

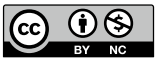
UK Department for Education. (2018). *New technology to spearhead classroom revolution* [webpage]. London, UK: UK Government. Retrieved from https://www.gov.uk/government/news/new-technology-to-spearhead-classroom-revolution.

von Davier, M., Khorramdel, L., He, Q., Jeong Shin, H., & Chen, H. (2019). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics, 44*(6). DOI: https://doi.org/10.3102%2F1076998619881789.

Wainer, H. (2000). *Computerized adaptive testing. A primer*. Abingdon, UK: Routledge.

Walker, M. (2017). Computer-based delivery of cognitive assessment and questionnaires. In: P. Lietz, J. C. Cresswell, K. F. Rust & R. J. Adams (Eds.), *Implementation of large-scale education assessments* (pp. 231–252). Wiley Series in Survey Methodology. New York, NY: Wiley.

Heiko Sibberns is a Senior Research Advisor at IEA. He has a teacher's degree for upper secondary education and extensive experience in the field of international comparative studies in education. He worked for many years at the data processing unit of the IEA Reading Literacy Study. In 1994, he became a co-director of the IEA Data Processing Center (later renamed IEA Hamburg), during which time he was responsible for all national projects conducted by IEA. In 2013, Mr. Sibberns became director of IEA Hamburg, a position he held until October 2019. Mr. Sibberns is a core member of the IEA Technical Executive Group, which provides guidance on the technical feasibility and best practice implementation of all IEA studies.



Chapter 11
Ensuring Validity in International Comparisons Using State-of-the-Art Psychometric Methodologies

Matthias Von Davier, Eugenio Gonzalez, and Wolfram Schulz

Abstract Researchers using quantitative methods for describing data from observational or experimental studies often rely on mathematical models referred to as latent variable models. The goal is to provide quantities that allow generalization to future observations in the same subject domain. A review of selected current and historical examples illustrates the breadth and utility of the approach, ranging from a worldwide used system for ranking chess players, to finding hidden structure in genetic data, to identifying common factors that can explain patterns of volatility of assets examined in financial modeling. This chapter describes how latent variable models are used in educational measurement and psychometrics, and in the studies of the International Association for the Evaluation of Educational Achievement (IEA) in particular. Within this domain, these models are used to construct a validity argument by modeling individual and system level differences as these relate to performance on large-scale international comparative surveys of skills, such as those commissioned by IEA.

Keywords Ability estimation · Educational measurement · Latent regression · Latent variable models · Psychometrics · Validity argument

M. Von Davier (✉)
Lynch School of Education and Human Development, Campion Hall, Boston College, 140 Commonwealth Avenue, Chestnut Hll, MA, USA
e-mail: vondavim@bc.edu

E. Gonzalez
Educational Testing Service (ETS), 44 Leamington Rd, Princeton, Boston, NJ MA 02135, USA
e-mail: egonzalez@ets.org

W. Schulz
Australian Council for Educational Research (ACER), Camberwell, Australia
e-mail: wolfram.schulz@acer.org

H. Wagemaker (ed.), *Reliability and Validity of International Large-Scale Assessment*, IEA Research for Education 10, https://doi.org/10.1007/978-3-030-53081-5_11

11.1 Introduction

International surveys of student learning outcomes, such as those conducted under the auspices of IEA, rely on advanced statistical methodologies developed in fields commonly called psychometrics or educational measurement. The goal of these fields is to provide mathematically rigorous methods for quantifying skills and knowledge based on observable data, mainly responses to survey questions.

The responses of students on tests of reading literacy or mathematical knowledge, or students' responses to questionnaire items designed to measure students' attitudes toward learning at school provide these observables. In international comparisons, there is a strong emphasis on selecting only those types of observables that tap into the comparable types of skills and attitudes across countries. In addition to the mathematical rigor of these methods, which enables researchers to check whether the responses are indeed reflecting a common construct or latent trait, expert knowledge regarding content, learning, child development, and skill acquisition and skill decline are important areas that have to be taken into account when using statistical methods to compare the performance of educational systems and their contexts across the world.

The analytic techniques used in IEA assessments follow best practices and use latent variable models developed over several decades. Their presentation in this chapter does not delve into the depths of the mathematical formalism, but uses equations and mathematical expressions whenever, in our judgement, a purely verbal description of important concepts would not be sufficient to provide an accurate representation.

In this chapter, we do not cover the full breadth of models used in educational measurement, but focus only on those psychometric models that were further developed and adapted for the application to data from international assessments of student learning outcomes.

The foundation of the approaches presented here are methods that aim at deriving quantitative information with regards to latent traits of interest (cognitive or non-cognitive) based on how respondents answer test or questionnaire items that are designed to assess the desired constructs. These constructs can be so-called cognitive constructs, such as reading literacy, or scientific literacy or mathematical skills, or non-cognitive constructs, such as mathematics self-efficacy, perceptions of classroom climate, or attitudes toward learning in the case of questionnaire-type surveys of attitudes or perceptions.

Typically, some foundational assumptions are made that enable quantitative measures to be derived from responses of students given to test or questionnaire items. These central assumptions can be summarized as follows:

(1) Each response can be scored to reflect a degree of correctness (in the dichotomous case as either correct or incorrect) or the degree of appraisal on a rating scale, which in a way reflects the amount of the constructs that are to be measured;

(2) Responses to all items are associated with a defined attitude or skill, i.e., the responses to the items are a non-random reflection or manifestation of the variable of interest, and not explainable by other (nuisance) variables; and
(3) The same variable of interest underlies the responses to the items, and affects the responses in the same way, across participating countries.

This chapter explains why researchers using quantitative methods for describing data from observational or experimental studies tend to rely on mathematical models that are usually referred to as latent variable models. A brief review provides a few current and historical examples in order to illustrate the breadth and utility of the approach, ranging from a worldwide used system for ranking chess players, to finding hidden structure in biological data, to identifying common factors that can explain patterns of volatility of assets examined in financial modeling. The examples are organized by complexity rather than chronologically, enabling latent variable models to be considered as a solution to a problem science faces on a regular basis: scientists aim to describe what is observable by relating observations to broader and more general principles that can be considered as underpinning the phenomena. We end this chapter with a more detailed description of how latent variable models are used in educational measurement and psychometrics. Within this domain, we focus on how these models are used for modeling individual and system level differences in performance on large-scale international comparative surveys of learning outcomes such as IEA's Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), International Civic and Citizenship Education Study (ICCS), and International Computer and Information Literacy Study (ICILS), but also, as these are using rather similar approaches, surveys such as the Organisation for Economic Co-operation and Development's (OECD's) Programme for International Student Assessment (PISA) and the Programme for the International Assessment of Adult Competencies (PIAAC).

A variable is called "latent" if it cannot be directly observed, although it may be possible to observe many of the effects, or at least correlates, of the hypothesized quantity on other related variables. While some may argue in favor of summarizing observed data without a reference to latent variables, there are striking reasons to use these types of models with variables. The assumption of "latent" variables (common in social and educational sciences) often raises concerns about assuming the existence of variables that cannot be measured directly. However, here we provide examples from natural sciences and economics to illustrate that many other disciplines facing similar problems use latent variable models. These are not only convenient but often turn out to be central to the solution of a scientific problem.

A simple example may clarify this: any person who has tried to calculate an average has in fact made an attempt to obtain a proxy to a latent variable. Whether the average is taken using test scores, ratings, or school grades, or whether batting averages in cricket, baseball, or softball are considered, the aim is the same: to obtain a more meaningful and stable summary of the underlying mechanism or cause (the average is also called a measure of central tendency) of outcomes in a certain domain.

At the same time most would agree that an average, while more stable than an individual score, is not to be confused with "true ability" (whatever that may be), as summary representations may be affected by particular circumstances at the time of measurement. A marathon runner will know a few stories to tell about which location, which time of year, and which temperature are more beneficial for faster finishing times. Usually this is not taken as evidence that the fastest runner in the world became consistently slower only because they happened to undertake the past few runs in extreme heat, in pouring rain, or at high elevations, conditions under which (almost) every runner is slower. Rather, people take task difficulty into account and assume that observed performance may vary based on these characteristics, while the underlying ability to run a fast marathon is virtually unchanged under constant and controlled conditions. This is the central tenet of latent variable models: different underlying mechanisms are assumed to affect observed performance. Some represent the agent's ability, some task difficulty, and potentially, some others represent additional situational characteristics such as time on task, or test taking motivation or engagement (see e.g., Rose et al. 2017; Ulitzsch et al. 2019).

Lewin (1939) described the dependency of observed behavior on personal characteristics and contexts in the following simple, semi-formal equation:

$$B = F(P, E)$$

Here, behaviors (B) are a function of person characteristics (P) as well as environmental factors (E). While it is uncertain whether this expression was already based on an understanding of the concept of broad behavioral tendencies in terms of latent variables, it appears that both the person P and the environment E are considered variables, and the behavior is considered a dependent variable, i.e., a different person P' may or may not show the same behavior B or B' when doing this influenced by the same environmental variable(s) E. Also, the same person P may show the same or different behavior if confronted with different environmental factors E'.

Interestingly, many latent variable models can be expressed in essentially the way Lewin (1939) described the dependency of observed behaviors on person variables and variables defining the situation in which a behavior is shown. The basic aim is to describe person attributes in a way that generalizes across items (environments) and, at the same time, to describe items in meaningful ways that allow comparisons and predictions about whether a certain person confronted with a certain environment is likely to behave in a certain way.

In Sects. 11.2 and 11.3, we describe how this basic aim was implemented in areas such as rating players in competitive games (chess, Call of Duty®, etc.), ordering or classifying biological observations according to their most likely genetic makeup, and finally describing factors that are responsible for group differences in skills, abilities, or affective variables measured in large-scale educational assessments. While on the surface chess player rating and quantitative genetics do not seem to have much in common with describing what students know and can do, we show that the methods used in these applications share a lot with what is done in educational

measurement. To illustrate this we use examples that share underlying assumptions about mechanisms operationalized as latent variables. Another common element is using controls for situational factors operationalized as variables, such as opponents, problem types, environments, time dependencies, or political events.

In Sect. 11.4, we put all these components together and show how gradual differences as well as systematic differences between groups, and the identification of common factors underlying a large number of observed variables are used in educational large-scale assessments. While these methods are broadly applied in educational measurement and other areas of data- and evidence-driven research, we focus on the quantitative methodologies as used in educational measurement in the context of international large-scale assessments (ILSAs) such as TIMSS, PIRLS, ICCS, ICILS, PISA, or PIAAC.

11.2 Modern Educational Measurement: Item Response Theory

Documented evidence of measuring human performance goes back at least to the second century BC in ancient China (e.g., Franke 1960). Humans have competed in sports and games for millennia (Murray 1913, 1952). Competitive, rule-based games such as chess, go, backgammon, and the ancient game of *hnefatafl* (Murray 1913, 1952) are most interesting when players are well matched, and opponents provide a challenge to the other player. In order to match players, a good match needs to be defined. One common approach is to assume that player strength can be quantified objectively, independent of the tasks or opponents this strength is measured against. Just as in the measurement of physical characteristics such as height and weight, the result of measurement should be (largely) independent of the scale or standard used to obtain the measure.

In the case of chess, rankings of players are well established. Elo (1978) came up with a method to adjust chess rankings of a player quasi on the fly, after each match, based on whether it had resulted in a win, loss, or a tie. The method Elo (1978) devised is an elegant way to provide players with an estimate of their strength (their ELO score) based only on the games they played. The mathematical foundations of this approach are based on what researchers describe as pairwise comparisons (Bradley and Terry 1952; Luce 1959). Under this paradigm, pairs of objects are compared, once or more than once, and either an observer declares a preference for one of the objects, or a rule-based system determines which of the objects has a higher ranking.

11.2.1 From Chess Ranking to the Rasch Model

The mathematical foundations of Elo's (1978) approach go back to Zermelo (1929), who developed a method for estimating player strength in chess tournaments where not all players compete against all others. In this approach, the probability of player A winning over player B is modeled as:

$$P(X = 1|\theta_A, \theta_B) = \frac{\exp(\theta_A - \theta_B)}{1 + \exp(\theta_A - \theta_B)}$$

where a data point $X = 1$ indicates a win of the first player over the second, and θ_A and θ_B denote the strength of players A and B, respectively. Zermelo (1929)[1] wrote the equation originally as:

$$P = \frac{w_A}{w_A + w_B}$$

where W_A and W_B denote the numbers of wins for players A and B respectively. This equation is mathematically equivalent to the above by setting $w_A = \exp(\theta_A)$. This form of the equation may seem more intuitive to some as it relates directly to repeated trials and winning and losing of players. If only two players are considered, and they compete 30 times, and player A wins 20 out of 30 while player B wins 10, the probability of player A winning becomes:

$$P_{AB} = \frac{20}{20 + 10} = \frac{2}{3}$$

However, player B is also matched with player C, where player C wins 15 times and player B wins 15 times, resulting in:

$$P_{BC} = \frac{15}{15 + 15} = \frac{1}{2}.$$

Intuitively, it could be inferred that players B and C are equally strong and it could also be assumed that in a match between A and C, the chances of A winning would be again 2/3. However, this does not need to be the case. This additional assumption may not be altogether solid, as C could be better at certain chess-related things (openings, for example) than B, while B could be better at endgames. Therefore, on average, the players may appear of equal strength, even though A is weaker than B at openings

[1] Economists tend to be familiar with Zermelo's name in the context of early game theory (Zermelo 1913; Schwalbe and Walker 2001) while mathematicians often know about Zermelo as one of the first scholars to formulate fundamental set theoretic results. His chess ranking model was independently rediscovered by Bradley and Terry (1952) and others. However, the estimation methods originally presented by Zermelo (1929) can be considered a more efficient maximum likelihood approach than that proposed by later authors who rediscovered the method independently (S.J. Haberman, personal communication 2016).

but extremely strong at endgames. In this case we would have an "endgame skill" as well as an "opening skill", and would need to devise two separate competitions to disentangle these skills with fidelity. Very often, however, chess players will be stronger or weaker than others in different kinds of chess-related situations. For example, a chess grandmaster tends to be excellent at openings, mid- and endgames, and will likely beat a novice in almost any match.

Related to the question whether we indeed measure just a single chess skill (as in assumption 2 in Sect. 11.1): how can we make sure we assess skills in a way that we can indeed generalize from seeing A and B playing as well as B and C to a hypothetical situation where A is confronted with a new task, say playing against C?

Much of the remainder of this chapter is concerned with how educational measurement is applied in ILSAs and builds models that allow these types of generalizations. Educational assessments assess individual students solving problems, for example when they answer questions about a text, or solve a mathematical problem. Models such as the ones introduced in Sects. 11.2–11.4 aim at deriving variables that allow these types of generalizations. In other words, these models aim at constructing proficiency measures that generalize beyond the specific set of test questions students see in educational assessments and permit more general statements about student's learning outcomes in different subject domains (such as mathematical or reading literacy) or other learning-related perceptions.

There is a striking similarity between the equations used in Zermelo's and Elo's chess ranking system and the model equation of the Rasch model (Rasch 1960; von Davier 2016). Georg Rasch was a Danish mathematician whose influence on educational measurement cannot be overstated (Olsen 2003). Many scholars have used the Rasch model as the basis for a variety of developments with broad impact on psychometric and educational measurement (see Fischer and Molenaar 1995; von Davier and Carstensen 2007). The Rasch model for dichotomous item responses is given by

$$P(X = 1|\theta_A, \beta_i) = \frac{\exp(\theta_A - \beta_i)}{1 + \exp(\theta_A - \beta_i)}$$

and θ_A, as before, denotes a strength, a skill, an ability or more broadly an attribute of person A, while β_i denotes the characteristic or difficulty of a task, indexed by $i = 1, \ldots I$. These tasks may be are a series of chess problems (as in: "checkmate in three moves") or a mathematics item ("solve for x: $3x + 8 = 20$" etc.) on a test, or some other exercise to check motor functions, or candidates selected by voters (Poole 2005). Fischer (1981) used the results provided by Zermelo (1929) to prove uniqueness and existence of maximum likelihood estimators of the Rasch model, and pointed out that the Rasch model is indeed a special case of the Zermelo (1929) model where two distinct sets Ω_θ and Ω_β of objects are always combined in a pairwise manner, while two objects from the same sets are never compared directly. More specifically, in the Zermelo approach all players can in principle be matched against all other players, the Rasch model assumes that human agents (test takers, respondents) are always paired with problems (tasks, test items) and, so to speak,

compete against tasks but not against each other. It is interesting to note that, apart from this particular feature, the two approaches are mathematically identical.

11.2.2 Characteristics of the Rasch Model

The Rasch model is one of the most successful item response theory models (IRT) (Lord and Novick 1968) and has been used for both large-scale international survey assessments as well as school-based and state assessments around the world. While there are more general models such as the two-parameter logistic (2PL) and three-parameter logistic (3PL) IRT models (Lord and Novick 1968), the Rasch model is considered one of the most elegant approaches as it has several mathematical properties that other models do not provide to the same extent (von Davier 2016). Among the applications of the Rasch model are the operational analyses and the reporting of proficiency scores from the initial reports from IEA's TIMSS 1995 grade 8 results, and those from PISA from 2000 until 2012, as well as those from IEA's ICCS and ICILS. While the Rasch model can be characterized as one of the most elegant approaches for measuring, it can be shown that it does not predict the observables as well as some more general models (see Sect. 11.2.3). TIMSS 1999 started using a more general approach, as did PIRLS, and PISA finally started using a more general model in 2015.

The reasons for this are best understood when comparing a test that assesses a multitude of topics in a variety of ways with the introductory chess example. A test, even one that a teacher may produce as a quick assessment of the students, contains different types of questions, requiring different types of answers. TIMSS and PIRLS items all assess a common domain (science, mathematics, reading literacy), but do so in a variety of ways. Chess matches are (to some extent) essentially always driven by the same objective: to achieve a checkmate (i.e., to capture the opposing side's king). Therefore, a model such as the Rasch model, which was originally developed to provide measures based on tests with extremely similar item materials, may need to be revised and extended in order to be suitable for broad, encompassing assessments such as TIMSS and PIRLS.

The Rasch model is not only one of the central tools of psychometrics and educational measurement but also an approach which is either reinvented frequently or highlighted as being particularly useful for complex situations which aim at deriving quantitative information from binary trials. The complexity often arises from the need to provide measures that enable comparisons even in situations where not all students are assessed on all tasks in a subject domain (which resembles the case of Zermelo's chess ranking model where chess players cannot play against all other players).

In many situations, measurement in education contexts cannot exhaustively assess all respondents on all possible tasks. Nevertheless, the aim is to make generalizable statements about the extent to which the task domain was mastered overall, and at what level this was the case. Ideally, the comparison between respondents should

not depend on what instrument they have been assessed with, just as the comparison of two chess players based on their ELO score should be independent of which opponents they faced in the past and, indeed, also independent of whether they ever played against each other.

In the Rasch model, this can be seen by calculating the logarithm of the odds (often referred to as log-odds) using the previous model equation. This provides

$$LO(\theta_A, \beta_i) = \log\left[\frac{P(X = 1|\theta_A, \beta_i)}{P(X = 0|\theta_A, \beta_i)}\right] = \theta_A - \beta_i$$

in the Rasch model. Further, when we calculate the difference

$$LO(\theta_A, \beta_i) - LO(\theta_B, \beta_i) = \theta_A - \theta_B$$

to compare any two respondents A and B, this difference turns out to be independent of which item β_i was used in the comparison. In order to compare any two tasks i, j we can use

$$LO\big(\theta_A, \beta_j\big) - LO(\theta_A, \beta_i) = \beta_i - \beta_j$$

which results in the same difference independently of which respondent was assessed.

In terms of practical comparisons, all respondents with the same total score receive the same estimate of the underlying latent trait θ under the Rasch model. In the situation of an educational test, for example, the probability of getting item i correct for respondent A can be estimated based on the relative frequency of getting the item correct based on the total group of respondents that has the same total score as respondent A. While the estimation methods used to generate optimal estimators are somewhat more sophisticated (Rasch 1960; Andersen 1970; von Davier 2016), comparisons can also be carried out across raw score groups and it is possible to calculate approximate differences (von Davier 2016) based on these conditional success probabilities in homogeneous score groups.

11.2.3 More General IRT Models

While the Rasch model can be considered the gold standard in terms of the ability to directly compare test takers independent of items, as well as items independently of the samples used to estimate item characteristics, this model puts rather strong constraints on how ability and the probability of successful task completion are related (von Davier 2016). There are more general IRT models that relax these assumptions and allow more flexibility when modeling the relationships between ability and task success. Among these, the models proposed by Birnbaum (1968) are the most commonly used ones. Since the 1999 cycle of the study, TIMSS has used the 3PL model for the multiple choice items, as does PIRLS.

As alternatives to the more constrained Rasch model, the 2PL and 3PL models are defined as

$$P_i(X = 1|\theta) = P(X = 1|\theta, a_i, b_i) = \frac{1}{1 + \exp(-a_i[\theta - b_i])}$$

and

$$P_i(X = 1|\theta) = P(X = 1|\theta, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{1}{1 + \exp(-a_i[\theta - b_i])}$$

where c_i is considered a guessing parameter (and assumed to be $c_i = 0$ in the Rasch model and the 2PL model) and $a_i > 0$ is a slope parameter and, as before, θ, b_i are ability and difficulty parameters, respectively. The 3PL model can be applied for multiple-choice test items in order to take the guessing probability.

It is important to note that all three models, the Rasch model, and the 2PL and 3PL models, assume a single ability variable θ underlying a measured learning outcome, and that with increasing ability, the probability of successful task completion (in the case of educational testing) increases. Similarly, both for the Rasch and 2PL model, with increasing endorsement of a particular affective dimension there is also a higher likelihood of providing higher ratings (in the case of measuring non-cognitive outcomes, for example when using items with response categories reflecting levels of agreement or disagreement). The respective item parameters (such as a_i, b_i, c_i) do not change this fundamental relationship, and it can be shown that all three models lead to rather similar results (Molenaar 1997). This is not a surprise as research has shown that all three unidimensional IRT models are mathematically equivalent to certain types of unidimensional item factor models (Takane and DeLeeuw 1987).

Perhaps the most important property of IRT, however, is the ability to provide estimates in circumstances in which not all respondents are administered all items, but are assessed with different subsets of tasks, which all measure the same subject domain. This is particularly important in the context of large-scale assessments, since these not only attempt to assess relatively broad domains with large numbers or items but also renew the sets of tasks used for a variety of reasons, including the need to cover new content while maintaining a strong connection to previous rounds of the assessments. In Sect. 11.2.4, we review two additional assumptions that are made when deriving customary IRT models that are very useful in this context.

As an example, TIMSS, like all other current IEA studies assessing student achievement, uses a design in which each student only receives a subset of the tasks. This allows the administration of many items to each of the participating countries in order to broadly cover the subject domains measured by TIMSS, without overburdening the students with endless testing sessions. In TIMSS 2015, there were a total of 28 blocks of items: 14 blocks of mathematics items and 14 blocks of science items (Fig. 11.1). Each booklet contained only two blocks of test items for each of the two domains, giving a total of four blocks per booklet. Each block appears in two different booklets, each time in a different location. This balances the exposure of

	Assessment Blocks			
Student Achievement Booklet	Part 1		Part 2	
Booklet 1	M01	M02	S01	S02
Booklet 2	S02	S03	M02	M03
Booklet 3	M03	M04	S03	S04
Booklet 4	S04	S05	M04	M05
Booklet 5	M05	M06	S05	S06
Booklet 6	S06	S07	M06	M07
Booklet 7	M07	M08	S07	S08
Booklet 8	S08	S09	M08	M09
Booklet 9	M09	M10	S09	S10
Booklet 10	S10	S11	M10	M11
Booklet 11	M11	M12	S11	S12
Booklet 12	S12	S13	M12	M13
Booklet 13	M13	M14	S13	S14
Booklet 14	S14	S01	M14	M01

Fig. 11.1 TIMSS 2015 student achievement booklet design at grades 4 and 8. *Notes* M indicates a block of mathematics items, S indicates a block of science items. *Source* Martin et al. (2013, p. 91). Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College

the items to the test takers and, at the same time, by having the blocks overlap with the others, allows to make inferences on items that were not taken by the students.

Just as in the chess example, using the Rasch model or more general IRT models allows these types of inferences if the same skill is measured by all the items within the domain. TIMSS and PIRLS employ a sophisticated set of statistical tools to check whether the items are indeed measuring the same skills across blocks, across positions in the assessment booklets, as well as over time and across participating populations.

11.2.4 Central Assumptions of IRT and Their Importance

IRT models are often referred to as item-level models that describe the probability of a correct response, given examinees proficiency θ and some item-specific parameters (such as a_i, b_i). However, this is not how IRT models are actually applied. Initially the item parameters and the proficiency θ are unknown, and all that analysts can rely on

is a series of answers to not one, but often 10, 20, or more items. What is needed is a model for many responses, and one that makes assumptions that enable inferences to be made about the unknown parameters given the observed responses. Three central assumptions underlying IRT modeling are unidimensionality, local independence, and measurement invariance.

11.2.5 Unidimensionality

We now assume that there are several items, and we denote the number of these items with I and the response variables with $X = (X_1, \ldots, X_I)$. Assuming unidimensionality means that a single quantity, the latent variable of interest, is sufficient to describe the probabilities of responses on each of the items, and that this is the same quantity regardless of the items, for a given person v.

So, for example, let P_{iv} and P_{jv} denote the probability of person v scoring 1 for items i and j, then, if unidimensionality holds, this can be re-expressed as

$$P_{iv} = P_i(X = 1|\theta_v)$$

and

$$P_{iv} = P_j(X = 1|\theta_v)$$

with θ_v being a real number.

Roughly, unidimensionality ensures that the same variable is measured with all the test items. This of course holds only if the assessment development aims at producing a set of items that indeed are designed to assess the same dimension. Unidimensionality would (very likely) not hold, for example, if half of the items in a skills test consisted of multiplication problems, and the other were assessing gross motor skills such as shooting a penalty kick, two seemingly unrelated skills.

In TIMSS and PIRLS, it is assumed that all items measure the same skill, or latent variable; in TIMSS this holds for mathematics and science separately, of course. At the same time, these assessments also allow the quantification of subscale or subdomain skills. This is a concept that relates to the fact that complex tests have multiple facets or topics that can be collected into distinct groups of items that tap into similar contexts or contents (Rijmen et al. 2014). While statistically there is good reason to report across these subscales using an overall mathematics, science, or reading score, there are situations where researchers may wish to analyze how groups of examinees perform in certain subdomains. One reason may be that these subdomains may "act" like they appear to be a single domain within each country while they also show distinct differences in performance across countries. This may be due to curricular differences across countries, and these differences can be studied in terms of their effect on subdomains. While here space limitations prevent a detailed explanation of this approach, the ideas have been developed and outlined in the TIMSS and PIRLS

technical reports (available for download on the IEA website; www.iea.nl). Further reading on the topic can be found in Verhelst (2012) and Feinberg and von Davier (2020).

11.2.6 Local Independence

The assumption of local independence states that the joint probability of observing a series of responses, given an examinees' proficiency level θ, can be written as the product of the item level probabilities, that is:

$$P(\mathrm{X} = x_1, \ldots, x_I|\theta) = \prod_{i=1}^{I} P_i(X = 1|\theta)^{x_i}[1 - P_i(X = 1|\theta)]^{1-x_i}$$

In particular, for any two items i and j, it can be assumed that

$$P\big(X_i = x_i \wedge X_J = x_j|\theta\big) = P(X_i = x_i|\theta)P\big(X_J = x_j|\theta\big)$$

While this assumption appears to be a rather technical one, it can be made more plausible by the following consideration. As already mentioned, the proficiency variable we intend to measure is not directly observable; it is only possible to observe behaviors that we assume relate to this proficiency, for example by means of the assumptions made in the IRT models. The assumption of local independence facilitates these inferences, in that it is assumed that once a respondent's proficiency is accounted for, all responses are independent from each other. That is, for example, knowing whether a respondent taking a test has correctly answered the previous question does not help predict their next response, assuming the respondent's true proficiency is already known.

This assumption can, of course, be inadequately supported by the data and the items in the assessment. While, in TIMSS, most of the items are independent units, there are cases where multiple questions are associated with a single item stem. Assessments of reading often contain multiple questions that relate to a single reading passage. This means that the assumption of local independence is potentially threatened by such a set of nested questions, but experienced item developers work to reduce these effects by making sure each question relates to an independent unit of knowledge or part of the text. In addition, statistical tools such as residual analysis and scoring approaches can alleviate the effect (e.g., Verhelst and Verstralen 1997).

The local independence assumption can not only be applied to the dependency of one item's responses on other items on the same test but also to other types of variables. Based on this more general understanding of the local independence assumption, if the model is correct, no other variables are helpful when predicting the next answer of a respondent, either given next on the test, or in three weeks' time. Only the underlying proficiency is mainly "responsible" for the probability of giving

correct item responses. In this sense, local independence is the assumption that it is only necessary to make inferences about the underlying cause of the behavior, not other observables and how they relate to test responses. If local independence holds, the latent variable provides all available information about the proficiency domain. It turns out that this expanded understanding of local independence is related to another assumption that is discussed next.

11.2.7 Population Homogeneity/ Measurement Invariance

One last central, but often only implicit assumption of IRT models is that the same relationships between item location (and other parameters), the respondents' latent trait, and item responses hold for all respondents. This seems to be a trivial assumption, but it turns out to be a centerpiece, one that cannot easily be ignored, particularly for the comparison of learning outcomes across countries or other types of population specifications. If the association between response probabilities and the underlying latent trait changes across groups, there would be no way to ensure that differences in their responses reflect actual differences in the measured learning outcomes, and not any other factors.

Formally, this assumption can be expressed as follows: if there are two respondents to a test, v and w, with $\theta_v = \theta_w$, denoting that both have the same proficiency, and $g(v) \neq g(w)$ indicating that the two respondents belong to different groups as defined by a grouping variable (such as gender, ethnicity, or country of residence), then population homogeneity holds if for these, and any two respondents v, w with $\theta_v = \theta_w$, we have

$$P_i(X = 1|\theta_v, g(v)) = P_i(X = 1|\theta_v) = P_i(X = 1|\theta_w) = P_i(X = 1|\theta_v, g(w))$$

In other words, the response probabilities for these two respondents depend only on their proficiency levels $\theta_v = \theta_w$ and item parameters, but not on the grouping variable $g(\cdot)$ or their group membership.

The assumption of population homogeneity means that response probabilities are fully determined by item characteristics and the measured latent trait. This assumption is central to the possibility of comparing learning outcomes across members of different groups. Note that this assumption does not say anything about the distribution of the measured latent trait in the different groups. It may well be that the average performance of respondents differs across groups. This is a seemingly subtle difference: the probability of getting the item right might be (on average, across all members) low in one group and high in another group. However, if we pick two test takers from each group, and it turns out that both test takers have the same ability level, then their chances of getting the same item right are the same, independently of which group they belong to.

While we do not know the "true" latent traits of respondents, we can assume that for each population, proficiencies are distributed according to some statistical

probability distribution, and with the assumption of population homogeneity we can estimate the total probability of responses based on this distribution alone, without considering any other variables. While it is not the only option, it is customary to assume a normal (or Gaussian) distribution. For a population of respondents, it is possible to assume that

$$\theta \sim \phi(\theta; \mu.\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{[S-\mu]^2}{\sigma^2}\right)$$

where μ is the mean of the distribution of the measured latent trait while σ denotes its standard deviation. Further, we may also assume more complex distributions, namely mixtures of distributions that consist of multiple Gaussians. In order to produce a statistical model for all the responses on an assessment, it is now straightforward to calculate the marginal probability as:

$$P(x_1, \ldots, x_I) = \int_{\theta} \prod_{i=1}^{I} P_i(X = 1|\theta)^{x_i}[1 - P_i(X = 1|\theta)]^{1-x_i}\phi(\theta; \mu, \sigma)d\theta$$

which is the standard marginalized form for models that contain unobserved variables, such as the latent trait measured in an IRT model. Among other things, this marginal probability takes center stage in the actual estimation of item parameters (such as a_i, b_i, c_i as well as the parameters of the proficiency distributions, μ and σ in this case). This expression is also important in the evaluation of how well this model (which could be considered an arbitrary choice) actually succeeds in predicting the observed responses of test takers within and across countries.

11.3 Simultaneous Modeling of Individual and Group Differences

Sums of many randomly varying quantities tend to approximately follow a normal distribution (e.g., Feller 1968); this is referred to as the central limit theorem. The normal distribution is probably the first and most important distribution taught in introductory statistics classes, and is fully described by only two parameters, a mean and a variance.

If, for example, a respondent to a test attempts many tasks that carry the same success probability, say $p = 0.5$, the summed number of successes can be well approximated by the normal distribution once there are more than 50 attempts, and with more certainty once more than 100 tasks have been attempted (Hays 1981). As an example, let $x = 1$ denote a success, and $x = 0$ an unsuccessful attempt, then for 50 independent attempts there is an expected value of $S = \sum x_i \sim 25$ successful attempts, and a variance of $p(1-p) \times 50 = 12.5$. While the probability distribution of the total number of success and failure can be described exactly by the

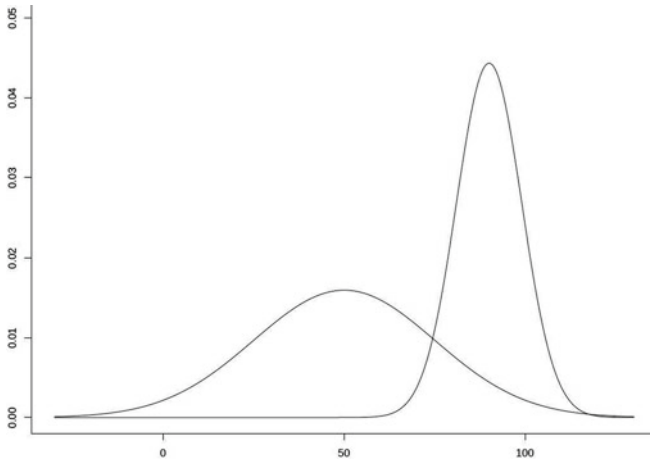


Fig. 11.2 Two normal distributions with means $\mu_1 = 50 = 0.5 \times 100$ and $\mu_2 = 90 = 0.9 \times 100$ and variances $\sigma_1^2 = 25 = 0.5 \times 0.5 \times 100$ and $\sigma_2^2 = 9 = 0.1 \times 0.9 \times 100$, respectively. It can be seen that for p = 0.9 a normal distribution does not provide a good approximation for the binomial with 100 trials, as substantial mass of the density is located at values larger than 100, while for p = 0.5 most of the mass of the approximation is located between 0 and 100

binomial distribution (e.g., Feller 1968; Hays 1981), the normal distribution provides a reasonable approximation by using

$$P(S) = \frac{f(S)}{Z} \text{and} f(S) \approx \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left[\frac{S-\mu}{\sigma}\right]^2\right)$$

with $\mu = 25$ and $\sigma = 12.5$ and $Z = \sum_{Q=0}^{50} f(Q)$ as a normalizing constant.[2] Consider two normal distributions (more accurately, normal densities) with expected values 30 and 90, and variances 21 and nine, respectively (see Fig. 11.2). These may correspond to two players who attempt gaming tasks (say playing against a chess program) with a constant success probability of 0.3 (weak player) and 0.9 (strong player), respectively.

However, there are many important cases where a normal distribution is not appropriate, even though many individual measures are averaged. One example was examined by Karl Pearson, who published, among many other things, several important pieces on a mathematical theory of evolution. Hence, Pearson (1894) is often also credited with having invented the mixture of normal distributions, a discovery that McLachlan and Peel (2000) traced back to Newcomb (1886). The relevance of this approach can be best understood by looking at an example. While Pearson (1894) was rather concerned with distributions related to the size of organisms (crabs, to be

[2] Strictly, the constant $\frac{1}{\sqrt{2\pi\sigma^2}}$ is not needed in the equation above if the normal is used to approximate a discrete distribution in the way described here.

specific) and how these depend on not directly observed genetic factors, education provides also examples of unobserved causes of differences resulting in distributions that are not necessarily symmetric.

Pearson (1894) studied a case in which an observed distribution is a composition (mixture) or two (or more) components, while each of these components can be thought of as a normal distribution with a component-specific mean and variance. Formally, if $F(\theta)$ denotes the distribution of a random variable θ and $f(\theta)$ the associated density, then the simplest case of a discrete mixture distribution can be written as

$$f(\theta) = p_1 \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2}\left[\frac{\theta - \mu_1}{\sigma_1}\right]^2\right) + p_2 \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2}\left[\frac{\theta - \mu_2}{\sigma_2}\right]^2\right)$$

The above equation writes the marginal distribution of a random variable as consisting of two normally distributed components with different means and standard deviations, here μ_1, σ_1 and μ_2, σ_2 that are specific to each component. Schork et al. (1996) reviewed the use of its general approach in human genetics research and pointed out that discrete mixture distributions are:

> … intuitive modelling devices for the effects of underlying genes on quantitative phenotypic (i.e. trait) expression. In addition, mixture distributions are now used routinely to model or accommodate the genetic heterogeneity thought to underlie many human diseases. Specific applications of mixture distribution models in contemporary human genetics research are, in fact, too numerous to count.

One of many examples from educational testing is the analysis of spoken English proficiency among medical students taking the United States Medical Licensing Examination Step 2 Clinical Skills exam described by Raymond et al. (2009), where the authors stated that the distribution of scores of more than 29,000 test takers did not appear to be well fitted by a normal distribution (see Fig. 11.3, which shows 3000 randomly drawn data points from such a distribution).

Even without a formal test it appears obvious that the histogram (Fig. 11.3) is not well approximated by a normal distribution. However, when splitting the sample into test takers with English as their first language, versus test takers who report English as their second language, each subsample can be well approximated by a normal distribution (similar to that seen in Fig. 11.2).

The differences observed using only the performance data may lead to the conclusion that there are disproportionally many respondents with very low skill levels relative to the average performance in the population. However, knowing that there are respondents with different language backgrounds may lead to a different conclusion, namely that there are group differences that should be considered. The examples (Figs. 11.2 and 11.3) imply very different realities. One starts from the notion of different populations with different skill distributions (Fig. 11.2), but this is not immediately evident in the second example (Fig. 11.3), which is based on the assumption that all test takers are random samples from the same distribution without considering the difference between test takers attending international versus

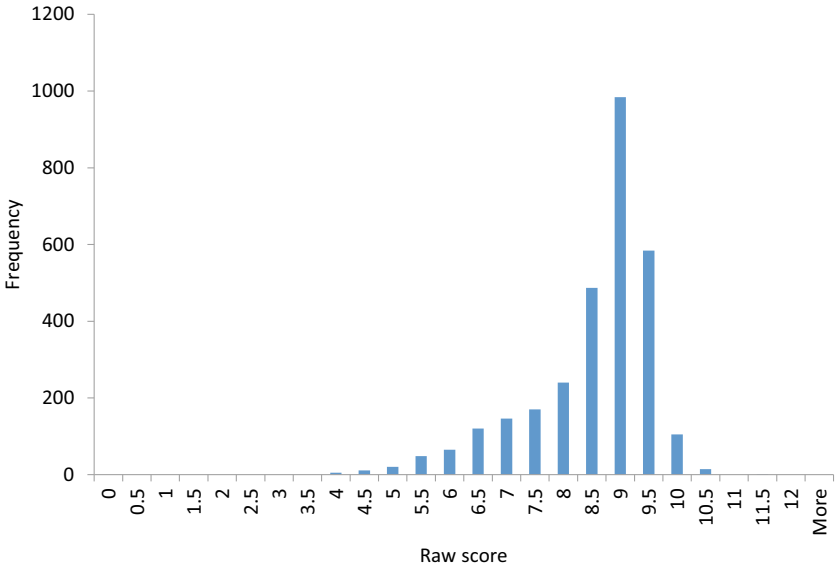


Fig. 11.3 Distribution of spoken language test scores obtained from 29,000 medical students taking the United States Medical Licensing Examination Step 2 Clinical Skills exam. The graph shows 3000 randomly drawn data points based on the distribution described by Raymond et al. (2009)

national schools when producing the graph that represents these as two separate but overlapping distributions.

Obviously, these types of group differences can be frequently encountered in populations that are composites of several subpopulations, for example those defined by regional, ethnic, or language differences. In TIMSS and PIRLS, these types of subpopulations can lead to important differences with respect to average performance as well as variability of skills in participating countries. Addressing this issue, and finding ways to incorporate the fact that group differences need to be assessed accurately has led to the types of multiple and mixture distribution models discussed in this section.

More generally, it is important to consider the existence of multiple populations, and that each population takes a translated and/or adapted version of the assessment in its source version, and that the outcome of the assessment depends on the underlying latent trait. This leads to several questions that we address in Sect. 11.4: how can we properly check whether the same latent trait is measured in all populations, how can we account for differences between populations with respect to trait distributions, and how can we examine whether the trait we are measuring is a valid representation of the construct we are interested in?

11.4 Statistical Modeling of Individual and Group Differences in IEA Survey Data

The famous statistician George Box is credited with the saying that "all [statistical] models are wrong, but some are useful," an insight which Box and coauthors have referred to in a variety of places (e.g., Box et al. 1978). This statement reflects the observation that statistical models can never include all possible evidence, and that even the quantities that are included may not entirely describe the way they truly interact. Statistical models can only be considered simplified formalisms that aim at providing useful summaries of observed data.

Statistical models used in ILSAs are certainly no exception. While similar models are used in national survey assessments (e.g., von Davier et al. 2007), the specific features that distinguish ILSAs are related to the added complexity due to the analyses of data from multiple countries surveying respondents with the same instruments in multiple languages (e.g., von Davier and Sinharay 2014; von Davier et al. 2009). This increased level of complexity is reflected in the following list of desiderata for statistical modeling:

(1) To account for and measure the extent to which groups differ on variables educational researchers are interested in;
(2) To specify how performance on a range of test items relate to underlying proficiency variables;
(3) To examine whether the assumed relationship between proficiency and observed responses is the same across countries;
(4) To integrate different sources of information such as responses and background information to facilitate reporting; and
(5) To provide variables for a database aimed at secondary analyses of proficiency data.

We will discuss how these desiderata are addressed in the approach used in large-scale assessments in the following subsections. International studies of educational outcomes translate these into a modeling approach in providing outcome variables that are comparable across countries and that facilitate secondary analyses as well as primary reporting. These reports aim at comparing the association of outcomes with policy-relevant contextual variables collected using student, teacher, parent, school and/or national context questionnaires.

11.4.1 Comparability as Generalized Measurement Invariance

When evaluating statistical modeling approaches for multiple populations, it is important to recognize that estimates of central quantities in the model, such as item parameters, as well as the estimates of distributional parameters, such as means and

standard deviations, may differ across populations. While distributional differences across populations are expected, and a focus of interest in terms of cross-country variance requires explanation, item parameter differences are undesirable with regard to the comparability of outcomes, and therefore should ideally be eliminated or at least reduced to negligible levels.

To illustrate this, we present a simple example with only three test items and two populations. We use a very much simplified IRT model with only two ability levels $\theta \in \{-1, +1\}$ with probabilities $P(\theta = -1) = 0.7 = 1 - P(\theta = +1)$ and one single binary item that differs with respect to how well it measures the ability variable θ in two groups A and B. This can be regarded as an item response model with a discrete latent variable (e.g., Follman 1988; Formann 1992; Haberman et al. 2008; Heinen 1996; Macready and Dayton 1977; von Davier 2005). Models of this type are studied in the context of mastery modeling, and more recent variants of these models are commonly referred to as diagnostic classification models (e.g., von Davier and Lee 2019).

Let us assume that the probabilities of a correct response in group A are given by $P_A(X = 1|\theta = -1) = 0.15$, and $P_A(X = 1|\theta = +1) = 0.85$, and in group B by $P_B(X = 1|\theta = -1) = 0.3$, and $P_B(X = 1|\theta = +1) = 0.70$.

The question is, what inferences can be drawn from observing a test taker who solves the single item on this test? We know that the test taker has an ability level of $\theta = +1$ with probability 0.3 and $\theta = -1$ with a probability of 0.7. Let us assume that the test taker succeeds in solving the item. Then we can apply Bayes theorem (Bayes 1763), one of the most fundamental equations in probability theory:

$$P_A(\theta = +1|X = 1) = \frac{P_A(X = 1|\theta = +1)P(\theta = +1)}{P_A(X = 1|\theta = +1)P(\theta = +1) + P_A(X = 1|\theta = -1)P(\theta = -1)}$$

Applying this theorem to our example then yields

$$P_A(\theta = +1|X = 1) = \frac{0.3 \times 0.85}{0.3 \times 0.85 + 0.7 * 0.15} \approx 0.71.$$

This can be translated to an estimated ability by using the expected-a posteriori (EAP) value, and in this simple mastery model, we obtain $\theta_{EAP|A} = -1 * 0.29 + 1 * 0.71 = 0.42$.

This means that our prior knowledge that mastery, $\theta = +1$, is only observed in 30% of population A is updated through the observation of one correct response by a test taker from population A and leads to a posterior probability of this test taker being in the group of masters of 71%. It appears that observing a correct response changes the likely mastery state of test takers in population A considerably. When looking at the same calculation in population B we obtain

$$P_B(\theta = +1|X = 1) = \frac{0.7 \times 0.3}{0.3 \times 0.7 + 0.7 \times 0.3} = 0.5.$$

As an EAP estimate we obtain $\theta_{EAP|B} = -1 * 0.5 + 1 * 0.5 = 0$ for population B. The change from identical prior knowledge (prevalence of masters with $\theta = +1$ is 30% also in population B) to posterior likelihood of mastery state is not quite as large among test takers in population B and we only would expect with a probability of 50% that they belong to the group that masters the material.

By drawing inferences about how likely it is that a test taker responds in a certain way, this constructed example illustrates that differences between groups in terms of item functioning can have profound consequences. In this example, our estimated probability that a test taker masters the material by solving an item decreases from 71 to 50% when comparing group B with group A. The reason is of course that the probabilities of correct responses on this item are very different for groups A and B.

What we implicitly assumed to be "the same item", on closer inspection, turns out not to function in the same way across the two comparison groups. Obviously, this was caused by the different probabilities of correct responses in the ability levels within groups, 0.7 versus 0.85 for respondents mastering the tasks and 0.3 versus 0.15 for those who fail to do so. The difference between those who fail and those who master the material is much larger in group A than in group B and, hence, if Bayes' theorem is applied, the information gained from observing a correct response in group A leads to a different adjustment of the posterior probability than in group B.

Therefore, one central commonly stated requirement is that items should work the same across all groups that are to be compared on the basis of a particular set of items. This is equivalent to the assumption of population homogeneity introduced earlier. Similar assumptions have been introduced in applied statistical modeling; this is referred to as measurement invariance (e.g., Meredith 1993; Millsap and Meredith 2007).

In IEA assessments, a careful review of items is conducted to ensure observed variables have the same relationship to the variable of interest, the skill and attitude variables that are intended goal of the assessment. The reports available for TIMSS and PIRLS, for example, allow insight into how translation accuracy is ensured, how items are reviewed, and how statistical tools are used to ensure invariance of the items across participating countries.

Even after items have been reviewed for appropriateness by all countries, and a translation into all languages of the assessment has been conducted, the quality control is not finished. As a hypothetical example, if an item was found to be impossible to solve in one specific country at the time of the first field testing of the item, this would trigger a series of investigations into, for example, the quality of translation, the review by countries, or the scoring guide. As a result, an item may be revised or eliminated, or the scoring guide adjusted for the affected country in order to ensure that the item that was found to violate invariance is either eliminated or changed in order to ensure integrity of the measurement.

11.4.2 Multiple-Group IRT Models

The idea behind IRT models is sometimes misunderstood as implying that the way items function, namely the way in which the measured trait determines item response probabilities, is independent of the groups that are being assessed. This is a prevalent misconception that appears to be based on a confusion of the concepts of a "population" versus a "sample (from a population)". It is sometimes said that IRT provides "sample-free" (e.g., Wright 1968) estimates (which is of course not true, a sample of observations is needed to estimate parameters). IRT (and in particular the Rasch model) are known to provide parameters that are (in expectation) invariant under sampling from the same population. Rasch (1966) spoke of specific objective comparisons: item difficulties can be compared independent of the respondents that were sampled to estimate the parameters.

In Sect. 11.3, we discussed how these types of misconceptions find their way into practice by implicit assumptions as a case of population homogeneity or measurement invariance. There is absolutely nothing in the models discussed here that will prevent parameters from varying from population to population. Population invariance is a feature of a test or an item that content experts have to work towards. That is why IEA, among many other quality control measures for curriculum-referenced studies such as TIMSS or PIRLS, performs curriculum coverage analyses to ensure that all items that become part of TIMSS or PIRLS have been vetted as appropriate for the student populations that are being tested. Statistics and psychometrics cannot enact population invariance, but rather they provide tools to test for invariance or approximate measurement invariance (e.g., Glas and Jehangir 2014; Glas and Verhelst 1995; Muthén and Asparouhov 2014; Oliveri and von Davier 2011; Yamamoto and Mazzeo 1992).

A customary approach to checking whether item parameters can be assumed to be invariant is estimation of multiple population versions of the statistical model under consideration. In IRT, these types of models have come to be known as multigroup IRT models (e.g., Bock and Zimowski 1997; von Davier 2016; von Davier and Yamamoto 2004). The basic assumption is that there is a finite number of populations denoted by $g \in \{g_1, \ldots, g_G\}$ and that the probabilities of correct response $P_{ig}(X = 1|\theta) = P_i(X = 1|\theta, g) = P\big(X = 1|a_{ig}, b_{ig}, c_{ig}, \theta\big)$ *may* depend on the group g as well as the ability variable. The same applies to the group specific ability distributions that can be mathematically described as:

$$\phi_g(\theta) = \phi(\theta|g) = \phi\big(\theta; \mu_g, \sigma_g\big).$$

While the model allows for deviations across multiple groups, ideally item parameters should be equal (invariant) across groups, or at least very similar (approximate invariance) in order to compare the latent trait estimates, so that situations like the one illustrated in the example in Sect. 11.4.1 do not occur. Note that, as already pointed out, $P_i(X = 1|\theta)$ should only depend on θ (which reflects the same latent trait across groups) and not on any other variables in order to meet assumptions of

population homogeneity or (strict) measurement invariance. This means the item response should only depend on the skill measured, not on the language version of the item, or the country where the item is administered. While this may be easier in chess and mathematics, as solving a system of linear equations is the same task no matter which language was used to ask the student to do this, it should also be possible to ensure in science and reading. Asking for the central agent in a reading passage should be possible, and should lead to the same response depending only on reading skills in the language of administration, and not on other variables such as cultural context.

The assumption of invariance would entail $a_{ig} = a_i, b_{ig} = b_i, c_{ig} = c_i$ for all population groups g. In terms of international studies, this would mean that the goal is to aim for the same shape of item functions across countries. If the items have the same item functions in all countries, the targeted skill has the same relationship to how the item is likely answered across participants from different countries.

Note that in cases of data collections where respondents come from multiple countries, the fact that each respondent was sampled in their country of (current) residence can be used to augment the data. Instead of only obtaining responses to items, we now have at our disposal the combined data, item responses x_i and group (country) membership g,

$$d_n^{IG} = \left[x_{n1}, \ldots, x_{nI}, g(n)\right]$$

where $g(n)$ represents the group (country) in which test taker n was tested. To estimate the marginal probability of the responses in the test takers group $g = g(n)$, we obtain

$$P(x_{n1}, \ldots, x_{nI}|g(n)) = \int_{\theta} \prod_{i=1}^{I} P_{ig(n)}(X = 1|\theta)^{x_{ni}} \left[1 - P_{ig(n)}(X = 1|\theta)\right]^{1-x_{ni}} \phi_g(\theta) d\theta$$

which, when assuming that all respondents complete their assessment independently can be used to define the likelihood of the data of all respondents as

$$P\left(d_1^{IG} \ldots d_N^{IG}\right) = \prod_{n=1}^{N} P(x_{n1}, \ldots, x_{nI}|g(n))$$

This is the basis for estimating the parameters μ_g, σ_g but also a_{ig}, b_{ig}, c_{ig}, typically starting by assuming that all item parameters are the same across groups, $a_{ig} = a_i, b_{ig} = b_i, c_{ig} = c_i$ for all g. While this is only a starting point, there exist elaborate procedures that allow items (and sub-groups or countries) to be identified for which this assumption is not met (more details about these procedures can be found in Glas and Jehangir 2014; Glas and Verhelst 1995; Oliveri and von Davier 2011; von Davier and von Davier 2007; Xu and von Davier 2006; Yamamoto and Mazzeo 1992).

Multiple-group IRT models are also used to link across cycles of international and national assessments where link items are included for the measurement of change

over time (Yamamoto and Mazzeo 1992) or where assessments are linked across different delivery modes of paper-based versus computer-based delivery (e.g., von Davier et al. 2019). The important feature of multiple-group IRT models in this context is the capacity to identify where there are item-by-country interactions that indicate that it is not possible to assume the same parameter across all countries or sub-groups. Technical reports on scaling assessments, such as those available online for TIMSS and PIRLS, show the linkage design both in terms of graphical displays and the number of link items involved. Link items may be used over two or more assessment cycles so that several data collections can indicate whether the items change over time or their retain measurement properties over multiple assessments (von Davier et al. 2019).

As a result of such analyses, some programs discard those items that do not meet the assumption of measurement invariance. Such an approach has the disadvantage that items are no longer used even though they may provide valuable information to increase the reliability of estimates for subgroups within countries. The fact that across countries these non-invariant items do not contribute to the comparability of scale means does not make them useless, so discarding items seems a rather extreme measure. However, there are more sophisticated approaches. It is also possible to maintain one set of item parameters that is used with common item parameters for all countries or sub-groups, while allowing parameters for some items to deviate in some countries or sub-groups. Such a procedure leads to a partial invariance model that maximizes the fit of the statistical model to the observed data, while maintaining the largest possible number of common parameters that meet criteria of measurement invariance (more details about a practical application of the approach can be found in von Davier et al. 2019).

11.4.3 Population Models Integrating Test and Background Data

Data collections in ILSAs tend to be fairly comprehensive. Aside from information about which items respondents completed in the assessment and how they scored, data from ILSAs also contain many additional variables about contextual variables collected in background questionnaires (from students, teachers, parents, and/or schools). These background data provide a rich context that allows secondary analysts to explore how students from different background perform on the assessment. Background variables further augment what is known about students participating in an assessment. We can write the complete data as

$$d_n^{IGB} = (x_{n1}, \ldots, x_{nI}, g_n, z_{n1}, \ldots, z_{nB})$$

where $z_{n1}, \ldots, z_{nB}$ represent the responses given by test taker n to the questions on the background questionnaire, g_n is the group membership (country of testing),

and $x_{n1}, \ldots, x_{nI}$ are the responses to the cognitive test items. The background data may contain additional variables from other sources (e.g., from school principal and teacher questionnaires) but, for simplicity, here we assume that we only make use of respondents' self-reports.

The background data can be assumed to be statistically associated with how respondents complete an assessment. What is measured by the assessment is quantified through the latent trait variable, θ, and readers are reminded that one of the central assumptions made in the previous sections was that the probability of observed successful performance is only related to θ, and no other variable. However, it was assumed that the distribution of this ability may depend on group membership and/or other (background) variables.

The population model consequently follows this line of reasoning by building a second level of modeling that predicts the expected value μ_n of the proficiency θ_n as a function of background variables $z_{n1}, \ldots, z_{nB}$:

$$\mu_n = \sum_{b=1}^{B} \beta_{g(n)b} z_{nb} + \beta_{g(n)0}$$

Furthermore, the proficiency variable is assumed as a normally distributed latent trait around this expected value, namely

$$\theta_n \sim N(\mu_n, \sigma)$$

Together, this provides a model for the expected proficiency given the background data $z_{n1}, \ldots, z_{nB}$. In other words, the expectation is that the distribution of proficiency depends on the background data used in the model. Such an assumption was mentioned in Sect. 11.3, when illustrating possible differences in learning outcomes between native speakers and second language speakers, and we also already mentioned the assumption of group-specific (e.g., across countries) latent trait means μ_g and standard deviations σ_g. However, the sheer amount of background data is much larger than the number of countries typically participating in an assessment. Therefore, if background variables are selected in such a way that suggests correlations with ability, it can be expected that the distribution around this expected value of $\sum_{b=1}^{B} \beta_{g(n)b} z_{nb} + \beta_{g(n)0}$ is somewhat narrower than a country-level distribution of abilities.

Formally, this is a multiple (latent) regression model that regresses the measured latent trait on background data collected in context questionnaires. The estimation problem is addressed separately within countries, since it cannot be assumed that background data have the same regression effects across different national contexts. Mothers' highest level of education, for example, is well known as a strong predictor of student performance, but this association can be moderated by other factors at the level of educational systems, so that in some countries it may be stronger than in others.

There are several ways to address the estimation of the regression parameters. In IEA assessments and other ILSAs, the latent trait is determined by an IRT model estimated across countries. Then the (latent) regression model is estimated using the item parameters obtained in the IRT estimation as fixed quantities. This ensures that the invariance properties that were determined through IRT estimation across countries will be applied equally to each national dataset (see, e.g., Mislevy et al. 1992; Thomas 1993; von Davier and Sinharay 2014; von Davier et al. 2007).

11.4.4 Group Ability Distributions and Plausible Values

The goal of the psychometric approaches described above is to produce a useful database that contains useful and reliable information for secondary users of the assessment data. This information comes in the form of multiple imputations or plausible values (PVs; see Mislevy 1991; Mislevy et al. 1992) of latent trait values for all respondents given all the responses to the assessment, as well as the knowledge about how respondents completed questions in the background questionnaire. Integrating the IRT model described in the first part of this chapter with the regression model introduced in the previous section, we can estimate the probability of the responses, conditional on background data, as

$$P_g(\mathbf{x}_n|\mathbf{z}_n) = \int_\theta \prod_{i=1}^{I} P_{ig}(x_{ni}|\theta)\phi\left(\theta; \sum_{b=1}^{B} \beta_{gb} z_{nb} + \beta_{g0}, \sigma\right) d\theta$$

This equation provides the basis for the imputation of plausible proficiency values for each respondent. To allow a more compact notation, we use $P_{ig}(x_{ni}|\theta) = P_{ig}(X = 1|\theta)^{x_{ni}}\left[1 - P_{ig}(X = 1|\theta)\right]^{1-x_{ni}}$

This model allows making inferences about the posterior distribution of the latent trait θ, given both the assessment items $x_i \ldots x_I$ and the background data $z_1 \ldots z_B$.

This Posterior Distribution Can Be Written as

$$P_g(\theta|\mathbf{x}_n, \mathbf{z}_n) = \frac{\prod_{i=1}^{I} P_{ig}(x_{ni}|\theta)\phi\left(\theta; \sum_{b=1}^{B} \beta_{gb} z_{nb} + \beta_{g0}, \sigma\right)}{\int_\theta \prod_{i=1}^{I} P_{ig}(x_{ni}|\theta)\phi\left(\theta; \sum_{b=1}^{B} \beta_{gb} z_{nb} + \beta_{g0}, \sigma\right) d\theta}$$

The posterior distribution provides an estimate of where the a respondent n is most likely located on the measured latent trait, for example by means of

$$E_g(\theta|\mathbf{x}_n, \mathbf{z}_n) = \int_\theta \theta P_g(\theta|\mathbf{x}_n, \mathbf{z}_n) d\theta$$

as well as the posterior variance, which provides a measure of uncertainty around this mean:

$$V_g(\theta|\mathbf{x}_n, \mathbf{z}_n) = E_g\left(\theta^2|\mathbf{x}_n, \mathbf{z}_n\right) - \left[E_g(\theta|\mathbf{x}_n, \mathbf{z}_n)\right]^2.$$

These two quantities allow PVs to be established (Mislevy 1991), quantities that characterize each respondent by means of a sample of imputed values representing their location on the latent trait measured by the assessment. PVs are the basis for group level comparisons and they contain information not only about the respondents' latent trait but also their group membership, such as a country or education system, as well as responses given to background questions (which may include attitude and interest scales, and self-reports about socioeconomic home background, such as books at home, parents' education, or parents' occupation).

PVs are random draws

$$\tilde{\theta}_{ng} \sim N\left(E_g(\theta|\mathbf{x}_n, \mathbf{z}_n), \sqrt{V_g(\theta|\mathbf{x}_n, \mathbf{z}_n)}\right)$$

that depend on response data x_n as well as background data z_n and group membership g, which in international assessments often relates to the country or education system where the respondent was assessed. That means two respondents with the same item scores but different background data will receive a different predicted distribution of their corresponding latent trait. This, on the surface, may appear incoherent when not considering the underlying statistical properties. The reason for the need to include all available (context or background) information into the model used for generating the PVs can be best understood when looking at the research on imputation methods (e.g., Little and Rubin 1987). The latent ability variable is not observed for any respondent, and must be inferred by imputation. When leaving out important predictors of ability, this imputation will lead to biased estimates of abilities as the relationship between abilities and context or background factors is ignored: Von Davier et al. (2009) illustrated this phenomenon in a simulation study that is modeled after large-scale assessments used by IEA and other organizations.

All available data needs to be included to derive quantities that allow unbiased comparisons of population distributions (e.g., Little and Rubin 1987; Mislevy 1991; Mislevy et al. 1992; von Davier et al. 2009). Importantly, PVs should never be referred to, used, or treated as individual assessment scores, because the term score commonly implies that the quantity depends only on the test performance of the individual respondent, and not on contextual data.

11.5 Conclusions

By design, although we only provide a cursory treatment of the psychometric methods underlying the scaling of large-scale assessment data as used when reporting statistics and building research databases for secondary analyses, we have illustrated the general principles and foundations of measurement used in ILSAs. Note that these methods have been developed in order to tackle the problem of estimating latent traits based on observed, but incomplete data, and that the goal is to provide quantities that allow generalization to future observations in the same subject domain.

Most of the major international student and adult skill assessments use methods that are closely related to the approaches presented here (von Davier and Sinharay 2014). The general principles used in international assessments also apply to many national assessments, however, national evaluations usually lack the complexity introduced by assessing multiple populations and in multiple languages. IRT models, the measurement modeling approach most commonly used, are also widely applied in high stakes testing, school-based testing, certification, licensure testing, and psychological and intelligence testing.

The latent regression model used to generate PVs is best understood as a general-purpose operational imputation approach that enables the integration of multiple sources of information while using computationally efficient preprocessing methods. At a conceptual level, the population model is best understood as a complex imputation model that allows complete data to be generated under certain conditional independence assumptions using incomplete data collection designs.

It is worth noting that while IRT was developed and has traditionally been applied when testing cognitive aspects of learning, it is also increasingly used to scale data derived from questionnaires, in particular in the context of international studies such as TIMSS, PIRLS, PISA, ICCS, and ICILS (see Martin et al. 2016, 2017; OECD 2017; Schulz and Friedman 2015, 2018). When applied to questionnaire data, IRT tends also to be used for analyzing measurement invariance (see Schulz and Fraillon 2011), often in combination with other analytical approaches, such as multiple-group analyses (see Schulz 2009, 2017).

Further information on the general modeling approach used to integrate background data, item responses, and invariance information to generate proficiency distributions for large-scale educational surveys can be found in von Davier and Sinharay (2014) and von Davier et al. (2007), and Rutkowski et al. (2014) described many aspects of the methodological approaches used in these studies. An accessible introduction to why grouping and background data is needed for unbiased estimates of proficiency distribution can be found in von Davier et al. (2009).

References

Andersen, E. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society Series B, 32,* 283–301.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society, 53*, 370–418. http://doi.org/10.1098/rstl.1763.0053

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York, NY: Springer.

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters: An introduction to design, data analysis, and model building*. New York, NY: John Wiley & Sons Inc.

Bradley, R. A., & Terry, M. E. (1952). The rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika, 39,* 324–45.

Elo, A. E. (1978). *The rating of chess players, past and present*. New York: Arco Publishing.

Feinberg, R., & von Davier, M. (2020). Conditional subscore reporting using iterated discrete convolutions. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/1076998620911933.

Feller, W. (1968). *An introduction to probability theory and its applications, Volume 1* (3rd ed.) New York, NY: John Wiley & Sons, Inc.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika, 46*(1), 59–77. https://doi.org/10.1007/BF02293919.

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models: Foundations, recent developments, and applications*. New York, NY: Springer.

Follmann, D. (1988). Consistent estimation in the Rasch model based on nonparametric margins. *Psychometrika, 53,* 553–562. https://doi.org/10.1007/BF02294407.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association, 87,* 476–486.

Franke, W. (1960). *The reform and abolition of the traditional Chinese examination system*. Harvard East Asian Monographs, Volume 10. Boston, MA: Harvard University Asian Center.

Glas, C. A. W., & Jehangir, K. (2014). Modeling country-specific differential item functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 97–115). New York, NY: Springer.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 69–95). New York, NY: Springer-Verlag.

Haberman, S. J., von Davier, M., & Lee, Y. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions*. ETS Research Report RR-08-45. https://doi.org/10.1002/j.2333-8504.2008.tb02131.x.

Hays, W. L. (1981). *Statistics for the social sciences* (3rd ed.). New York, NY: Holt, Rinehart and Winston.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences.* Advanced Quantitative Techniques in the Social Sciences, Volume 6. Thousand Oaks, CA: Sage Publications.

Lewin, K. (1939). Field theory and experiment in social psychology: Concept and methods. *American Journal of Sociology, 44*(6), 868–896. https://doi.org/10.1086/218177.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: John Wiley & Sons Ltd.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, NY: John Wiley & Sons Ltd.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2*(2), 99–120.

Martin, M. O., Mullis, I. V. S., & Foy, P. (2013). TIMSS 2015 assessment design. In I. V. S. Mullis & M. O. Martin (Eds.), *TIMSS 2015 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. http://timssandpirls.bc.edu/timss2015/frameworks.html.

Martin, M. O., Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., & Palazzo, L. (2016). Creating and interpreting the TIMSS 2015 context questionnaire scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 15.1–15.312). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. http://timss.bc.edu/publications/timss/2015-methods/chapter-15.html.

Martin, M. O., Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., Fishbein, B., & Liu, J. (2017). Creating and interpreting the PIRLS 2016 context questionnaire scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 14.1–14.106). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-14.html.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics. New York, NY: John Wiley & Sons, Inc. http://dx.doi.org/10.1002/0471721182

Meredith, W. M. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543.

Millsap, R. E., & Meredith, W. (2007). Factorial invariance: Historical perspectives and new problems. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and new directions* (pp. 131–152). Mahwah, NJ: Erlbaum.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196.

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

Molenaar, W. (1997). Lenient or strict application of IRT with an eye on practical consequences. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 38–49). Münster, Germany/New York, NY: Waxmann Verlag.

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology, 5,* 978. https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00978/full.

Murray, H. J. R. (1913). *A history of chess*. Oxford, UK: Oxford University Press.

Murray, H. J. R. (1952). *A history of board games other than chess*. Oxford, UK: Clarendon Press.

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics, 8*(4), 343–366.

OECD. (2017). *PISA 2015 technical report*. Paris, France: OECD. https://www.oecd.org/pisa/data/2015-technical-report/.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Journal of Psychological Test and Assessment Modeling, 53,* 315–333.

Olsen, L. W. (2003). *Essays on Georg Rasch and his contributions to statistics*. Ph.D. thesis. Institute Of Economics, University of Copenhagen, Denmark. https://www.rasch.org/olsen.pdf.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 185,* 71–110. https://doi.org/10.1098/rsta.1894.0003.

Poole, K. T. (2005). *Spatial models of parliamentary voting*. Cambridge, UK: Cambridge University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Volume 1 of Studies in mathematical psychology. Copenhagen, Denmark: Danmarks Paedagogiske Institut (Danish Institute for Educational Research).

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19,* 49–57. https://doi.org/10.1111/j.2044-8317.1966.tb00354.x.

Raymond, M. R., Clauser, B. E., Swygert, K. A., & van Zanten, M. (2009). Measurement precision of spoken English proficiency scores on the USMLE Step 2 Clinical Skills examination. *Academic Medicine, 84*(10 Suppl.), S83–S85.

Rijmen, F., Jeon, M., von Davier, M., & Rabe-Hesketh, S. (2014). A third order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys. *Journal of Educational and Behavioral Statistics, 38,* 32–60. https://doi.org/10.3102/1076998614531045.

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika, 82,* 795–819. https://doi.org/10.1007/s11336-016-9544-7.

Rutkowski, L., von Davier, M., & Rutkowski, D. (2014). *Handbook international large-scale assessment: Background, technical issues, and methods of data analysis.* London, UK: CRC Press (Chapman & Hall).

Schork, N. J., Allison, D. B., & Thiel, B. (1996). Mixture distributions in human genetics research. *Statistical Methods in Medical Research, 5*(2), 155–178. https://doi.org/10.1177/096228029600500204.

Schulz, W. (2009). Questionnaire construct validation in the International Civic and Citizenship Education Study. In M. von Davier & D. Hastedt (Eds.), *Issues and methodologies in large-scale assessments*, IERI Monograph Series Volume 2 (pp. 113–135). Hamburg, Germany: IERI. http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_05.pdf.

Schulz, W. (2017). Scaling of questionnaire data in large-scale assessments. In P. Lietz, J. Cresswell, K. Rust, & R. Adams (Eds.), *Implementation of large-scale education assessments* (pp. 384–410). Chichester, UK: John Wiley & Sons Ltd.

Schulz, W., & Fraillon, J. (2011). The analysis of measurement equivalence in international studies using the Rasch model. *Educational Research and Evaluation, 17*(6), 447–464.

Schulz, W., & Friedman, T. (2015). Scaling procedures for ICILS questionnaire items. In J. Fraillon, W. Schulz, T. Friedman, J. Ainley, & E. Gebhardt (Eds.), *International Computer and Literacy Information Study 2013 technical report* (pp. 177–220). Amsterdam, the Netherlands: IEA. https://www.iea.nl/publications/technical-reports/icils-2013-technical-report.

Schulz, W., & Friedman, T. (2018). Scaling procedures for ICCS 2016 questionnaire items. In W. Schulz, R. Carstens, B. Losito, & J. Fraillon (Eds.), *ICCS 2016 technical report* (139–243). Amsterdam, the Netherlands: IEA. https://www.iea.nl/publications/technical-reports/iccs-2016-technical-report.

Schwalbe, U., & Walker, P. (2001). Zermelo and the early history of game theory. *Games and Economic Behavior, 34*(1), 123–137.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408. https://doi.org/10.1007/BF02294363.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics, 2,* 309–322.

Ulitzsch, E., von Davier, M., & Pohl, S. (2019). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12188.

Verhelst, N. D. (2012). Profile analysis: A closer look at the PISA 2000 reading data. *Scandinavian Journal of Educational Research, 56*(3), 315–332. https://doi.org/10.1080/00313831.2011.583937.

Verhelst, N. D., & Verstralen, H. H.F. M. (1997). Modeling sums of binary items by the partial credit model. Measurement and Research Department Research Report 97-7. Arnhem, Netherlands: Cito.

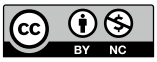
von Davier, M. (2005). *A general diagnosis model applied to language testing data.* ETS Research Report RR-05-16. https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.2005.tb01993.x.
von Davier, M. (2016). The Rasch model. In W. van der Linden (Ed.), *Handbook of item response theory, Volume 1* (2nd ed.) (pp. 31–48). Boca Raton, FL: CRC Press. http://www.crcnetbase.com/doi/abs/10.1201/9781315374512-4.
von Davier, M., & Carstensen, C. H. (Eds.). (2007). *Multivariate and mixture distribution Rasch models.* New York, NY: Springer.
von Davier, M., & Lee, Y.-S. (2019). *Handbook of diagnostic classification models: Models and model extensions, applications, software packages.* New York, NY: Springer.
von Davier, M., & Sinharay, S. (2014). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155–174). Boca Raton, FL: CRC Press.
von Davier, M., & von Davier, A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 3*, 115–124. https://doi.org/10.1027/1614-2241.3.3.115.
von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT Models: An extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*(6), 389–406. https://doi.org/10.1177/0146621604268734.
von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Eds.), *Issues and methodologies in large-scale assessments*, IERI Monograph Series Volume 2 (pp. 9–36). Hamburg, Germany: IERI. http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf.
von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2007). The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 1039–1055)., Psychometrics North Holland: Elsevier.
von Davier, M., Yamamoto, K., Shin, H.-J., Chen, H., Khorramdel, L., Weeks, J., et al. (2019). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assessment in Education: Principles, Policy and Practice, 26*(4), 466–488. https://doi.org/10.1080/0969594X.2019.1586642.
Wright, B. D. (1968). Sample-free test calibration and person measurement. In B.S. Bloom (Ed.), *Proceedings of the 1967 Invitational Conference on Testing Problems* (pp. 85–101). Princeton, NJ: Educational Testing Service.
Xu, X., & Von Davier, M. (2006). *Cognitive diagnosis for NAEP proficiency data.* ETS Research Report RR-06-08. https://doi.org/10.1002/j.2333-8504.2006.tb02014.x.
Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17*(2), 155–173.
Zermelo, E. (1913). On an application of set theory to the theory of the game of chess. Reprinted in E. Rasmusen (Ed.). (2001). *Readings in games and information.* Oxford, UK: Wiley-Blackwell.
Zermelo, E. (1929). The calculation of tournament results as a maximum-likelihood problem [in German]. *Mathematische Zeitschrift, 29,* 436–460.

Matthias Von Davier research focuses on developing psychometric models for analysing data from complex item and respondent samples and on integrating diagnostic procedures into these methods. His areas of expertise includes topics such as item response theory, latent class analysis, classification and mixture distribution models, diagnostic models, computational statistics, person-fit, item-fit, and model checking, as well as hierarchical extension of models for categorical data analysis, and the analytical methodologies used in large scale educational surveys.

Eugenio Gonzalez is a Principal Research Project Manager at Educational Testing Service (ETS), and director of the IEA-ETS Research Institute (IERI), a collaborative effort between the International Association for the Evaluation of Educational Achievement (IEA) and ETS that focuses on improving the science of large-scale assessments. IERI undertakes activities around three broad areas of work that include research studies related to the development and implementation of large-scale assessments; professional development and training; and dissemination of research findings and information gathered through large-scale assessments. Dr Gonzalez is also responsible for the technical documentation and international database training activities for PIAAC and PISA.

Dr. Gonzalez was formerly head of the Research and Analysis Unit at the IEA Hamburg (2007–2012), the director of quality control and field operations for the National Assessment of Educational Progress (NAEP) (2004–2006), and director of international operations and data analysis in the TIMSS & PIRLS international study center (ISC) at Boston College (1994–2004). In this last role, he oversaw the development and implementation of international operations, data analysis, and reporting procedures for the Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS). Since 1993, he has conducted database training activities for several research and governmental organizations, and has been a developer and technical lead of customized software for analyzing international large-scale assessment databases such as the IEA's IDB Analyzer and the Data Explorer. Dr. Gonzalez has also served as external consultant to several national and international large scale assessment programs, and has held teaching appointments at Boston College and the University of Massachusetts, Amherst. He has a PhD in Educational Research, Measurement, and Evaluation from Boston College, and an undergraduate degree in Psychology from the Universidad Católica Andres Bello in Caracas, Venezuela.

Wolfram Schulz is a Principal Research Fellow (formerly Research Director International Surveys) at the Australian Council for Educational Research (ACER) where he has worked on a large number of national and international large-scale assessment studies. He is International Study Director of the IEA International Civic and Citizenship Education Study (ICCS) and Assessment Coordinator for the IEA International Computer and Information Literacy Study (ICILS). He is also a member of the IEA Technical Executive Group (TEG).



Chapter 12
Publications and Dissemination

Seamus Hegarty and Sive Finlay

Abstract The International Association for Evaluation of Educational Achievement (IEA) has been a major source of scholarship and publication in international large-scale assessment from its early days in the 1960s until the present. This publication activity is grounded on IEA's large-scale assessment projects. IEA publications can be grouped in terms of: core project publications; project-related publications, including the IEA Research for Education series; the academic journal, *Large-scale Assessments in Education*; international research conference materials; and its occasional policy brief series IEA Compass: Briefs in Education. All IEA publications are open access to ensure the widest possible dissemination. Quality assurance is built into project activity, in keeping with IEA's focus on reliability and validity. A further level of independent scrutiny is provided by the Publications and Editorial Committee, whose remit is to ensure that all publications meet the highest standards of research and scholarship. IEA's extensive communications and dissemination activity serves both to take study findings to diverse audiences and to expose them to a further level of peer and public scrutiny.

Keywords Dissemination · IEA Compass briefs · IEA's international research conference · IEA Research for Education series · IERI journal · Promotional material · Publications · Publications and Editorial Committee · Quality assurance review

S. Hegarty (✉)
University of Warwick, Coventry, UK
e-mail: seamus@seamushegarty.com

S. Finlay
International Association for the Evaluation of Educational Achievement (IEA), Amsterdam, The Netherlands
e-mail: secretariat@iea.nl

H. Wagemaker (ed.), *Reliability and Validity of International Large-Scale Assessment*, IEA Research for Education 10, https://doi.org/10.1007/978-3-030-53081-5_12

12.1 Introduction

The International Association for the Evaluation of Educational Achievement (IEA) has a long history of publications and, over successive decades, has developed a dissemination strategy that is intended to maintain focus on the central objective of supporting educational progress around the world. Assessment studies run the risk of simplistic headlines. IEA's constant endeavor has been to ensure that study findings are used primarily to deepen understanding of the policies and practices that foster educational reform and, where necessary, to challenge the simplistic use of assessment data. This runs alongside the objective of helping educational systems to develop their own research and analysis capacities.

From the early ground-breaking research reports of the 1960s to the extensive contemporary output, IEA publications have been at the forefront of developments in international large-scale assessments (ILSAs). Besides making available unique datasets on student achievement worldwide, which have wide resonance nationally and internationally, IEA's ILSAs embody outstanding scholarship and have helped to shape and to advance the science of large-scale assessment. In addition, these datasets and publications underpin a vast number of journal articles and university dissertations by independent scholars. Most importantly, they have contributed, and continue to contribute, to educational reform and the enhancement of student learning around the world.

All IEA publication activity derives from the projects but, for purposes of this chapter, here we categorize the publication output as follows:

- Core project publications
- Project-related publications
- Academic journal
- International research conference
- IEA Compass briefs.

Each of these publication strands is subject to rigorous quality assurance reviews to ensure that they meet the highest standards of research and scholarship.

We also draw attention to IEA's dissemination activity, which serves both to take study findings to diverse audiences and to generate publications in its own right.

12.2 Core Project Publications

Each IEA study generates a set of open-access publications that are freely accessible for download (IEA 2020a). The starting point, naturally, is an assessment framework, a document that articulates the basic structure of each study and lays down the parameters for its design and execution. The framework sets out what is to be measured and how it will be measured. It details the background information that will be collected to enable analysis of the student achievement data. It also outlines

an assessment design, including information about the population(s) to be sampled, measurement instruments, and test administration logistics. Assessment frameworks build on previous cycles of the studies but also evolve: they maintain continuity with previous studies so as to permit the measurement of trends over time, but they must also take account of changes in the educational environment so as to incorporate new thinking and remain educationally relevant.

Assessment frameworks start by specifying the content of the assessment. The four most recent assessment frameworks for IEA studies serve as good examples.

IEA's Trends in International Mathematics and Science Study (TIMSS) 2019 presented two content assessment frameworks, one for mathematics and one for science (Mullis and Martin 2017). Each one was organized around two dimensions: (1) content, which specifies the subject matter to be assessed; and (2) cognitive, which specifies the thinking processes to be assessed. Thus, the content for grade 8 assessment in mathematics was grouped under number, algebra, geometry, and data and probability, and in science under biology, chemistry, physics, and earth science. For both, the cognitive processes assessed remain the same: knowing, applying, and reasoning.

The Progress in International Reading Literacy Study (PIRLS) 2016 assessment framework was described in terms of the major purposes of reading and the comprehension processes to be tested (Mullis and Martin 2015). The purposes of reading are: reading for literary experience, and reading to acquire information. Four processes were assessed: (1) retrieving specific information; (2) making inferences from text; (3) interpreting information in a text; and (4) evaluating content.

The International Computer and Information Literacy Study (ICILS) 2018 assessment framework (Fraillon et al. 2019) was developed in terms of two dimensions: (1) computer and information literacy; and (2) computational thinking. Meanwhile, the International Civic and Citizenship Education Study (ICCS) 2016 assessment framework (Schulz et al. 2016) was organized around four content domains: (1) civic society and systems, (2) civic principles, (3) civic participation, and (4) civic identities. In addition to this there were two cognitive domains: (1) knowing, and (2) reasoning and applying (seen as a single domain). In addition, and in recognition of the importance of students' attitudes toward civic engagement, the ICCS 2016 framework included an affective-behavioral dimension comprising two further domains: (1) attitudes, and (2) engagement.

The second component of the assessment frameworks relates to background information. This encompasses, as appropriate for each study, information on national context, national curriculum, school and classroom factors, teacher background, student perceptions, and home background. The final component expands on the study design, specifying the populations to be sampled, the instruments (tests and questionnaires) and item types to be used, and coverage of framework domains, along with practical details on test administration and questionnaire management. Increasingly, studies are moving some or all of their activities from paper to computer modality, and details are presented as necessary, including any linking arrangements that may be required.

For PIRLS and TIMSS, and sometimes ICCS, the assessment framework is supplemented by an encyclopedia (see IEA 2020a for some examples). These publications draw on public data, responses to study-specific national questionnaires and structured reports from national experts. The encyclopedias present demographic and economic data for participating countries and outline the structure of their education system. They describe the relevant curricula for the grade(s) being assessed, along with instructional practices and school organization. Other topics covered include teacher education, and specifically teachers' preparedness for the curriculum area in question, and national assessment and quality assurance procedures. These encyclopedias serve a two-fold purpose: they provide an additional lens for analyzing student achievement data within countries; and they constitute a rich source of information on school practices globally regarding reading, mathematics and science, and civics and citizenship education.

The key publication for each study is the international report. This presents overall study findings alongside background data that put student achievement results in context. Thus, for each participating country, the TIMSS 2015 mathematics report (Mullis et al. 2016) described students' achievement in mathematics, on average and at benchmark points, for both content and cognitive domains. It also reported on factors such as home background, school resources, school climate, teacher preparedness, and classroom instruction. The ICCS 2016 report describes how well young people are prepared for their role as citizens: it presents data on their civic knowledge, civic engagement and attitudes toward key issues in society, alongside school and broader contextual factors.

Once studies have been completed and the international report released, technical reports, the datasets, and supporting documentation are made publicly available. Technical reports are detailed documents which provide, for each study, an overview of the entire study. They document the development of test and questionnaire items, along with the rigorous translation verification procedures used and any national adaptations to the instruments. They give details on sampling design and implementation, field operation and quality assurance procedures, data collection and management, database construction, weighting procedures, and the construction of scales. Full data for IEA studies are freely available from the IEA Data Repository (IEA 2020b). Data are available in SPSS and SAS formats.

There are a variety of free software tools and additional resources which help in analyzing IEA datasets, all of which are available from the IEA website (www.iea.nl). They include the International Database Analyzer (IDB Analyzer; see IEA 2020c), which is a tool for combining and analyzing data from all IEA datasets (and from some other large-scale assessment datasets); user guides, which contain information on the structure of datasets and the variables contained within them, and how to conduct analyses using the IDB Analyzer (for a full overview of available user guides see IEA 2020a); the IEA Data Visualizer (IEA 2020c), which assists in visualizing trends over time and regional variations in TIMSS and PIRLS data; and the ILSA Gateway (ilsa-gateway.org/), which provides easy access not only to IEA studies but also to other international large-scale assessments. The IEA website also

highlights upcoming training opportunities and provides online video tutorials for the IDB Analyzer (IEA 2020c, d).

12.3 Project-Related Publications

International project reports are the starting point for more detailed scrutiny of student performance within individual countries. Many factors impinge on the latter, and in-depth analysis is necessary to undertand the patterns of student achievement within countries. Following the international release of study results, many national centers of participating countries publish national reports in the national language (see IEA 2020a). These present study findings within the context of the specific country and serve a valuable function in highlighting policy and practice implications at national level.

In 2016, IEA launched the open-access IEA Research for Education series (IEA 2020e) as an additional initiative to support analysis of the data and encourage dialogue focusing on policy matters and technical evaluation procedures. This series is dedicated to promoting in-depth analyses of IEA data and has a twin focus on significant policy issues and methodological innovation. IEA issues two calls to tender each year, and successful bidders are given financial support to carry out an agreed program of work. A wide range of topics has been covered to date, from gender differences in computer and information literacy and teaching tolerance in a globalized world, to the globalization of science curricula and the link between teacher quality and student achievement in mathematics.

IEA studies have led to a vast number of academic papers and dissertations. These papers can be found in many academic journals and testify to IEA's significant impact on assessment scholarship. Likewise, student dissertations in many universities draw on IEA datasets and publications. As a measure of support for this scholarship, IEA provides two annual awards (IEA 2020f): the Richard M. Wolf award and the Bruce H. Choppin award (Wolf and Choppin were well-renowned psychometricians who made significant contributions to the development of IEA studies). The Wolf award recognizes the author or authors of "a paper published in a refereed journal, monograph, or book that includes analysis of data from one or more IEA studies," and the Choppin Award recognizes outstanding "master's theses or doctoral dissertations that employ empirical research methods and use IEA data."

12.4 Academic Journal

IEA also publishes its own journal, *Large-scale Assessments in Education* (IEA 2020g), in association with Educational Testing Service (ETS; www.ets.org). This is dedicated to the science of large-scale assessment and publishes articles that utilize not only IEA data but also assessment data collected by any other similar large-scale

studies of student achievement, such as the ILSAs undertaken by the OECD and the US National Assessment of Educational Progress (NAEP). In addition to research articles, the journal publishes reviews and articles presenting methodological and software innovations relevant to the analysis of large-scale assessment data.

Large-scale Assessments in Education is published on an open-access basis by Springer, and all articles accepted for publication are made permanently available online without subscription charges. It also provides the opportunity to publish or link to large datasets, support static graphics or moving images, and display data in a form that can be read directly by other software packages so as to allow readers to manipulate the data for themselves.

12.5 IEA International Research Conference (IRC)

IEA's international research conference is a well-established biennial event that brings together researchers working with IEA data to present their findings to the wider research community. These can be cross-national studies or explorations of, for example, curriculum or teaching practice within particular countries, as well as an opportunity to explore different research methodologies. The conferences enable colleagues to share perspectives and enhance their professional development. In addition to the main conference activities, pre-conference workshops are held on specialized topics related to large-scale assessment.

These conferences have also generated a large number of publications, as IEA conference proceedings (IEA 2020h) and as subsequent articles in academic journals or policy papers. In 2017, IEA established the Constantinos Papanastasiou Poster Prize for the best poster submitted to the IEA IRC (IEA 2020f). The prize is awarded in recognition of Professor Papanastasiou's enthusiastic, long-term contributions to building and supporting the educational research community, including founding and hosting the inaugural IRC in 2004 at the University of Cyprus.

12.6 IEA Compass Briefs

IEA studies are complex, and project reports along with the analyses deriving from them are, necessarily, detailed and require close reading. This indeed is one of the reasons why media headlines can so often mislead. Technical language and methodological qualifications do have a role, and when they are set aside without understanding miscommunication becomes likely.

There is still, however, a need to communicate study findings beyond the research community. The studies have implications and raise questions of relevance to policy-makers, curriculum developers, teacher educators and teachers, as well as the general public. To address this need, IEA has developed the IEA Compass: Briefs in Education series (IEA 2020i). This is a set of brief documents that use IEA study data

to address issues of interest to various educational stakeholders. Each publication in the series aims to connect study findings to recurrent and emerging questions in education policy and practice. Recent briefs have covered topics as diverse as: how safe primary-aged students feel at school and how perceptions of safety affect their learning; Latin American students' support for dictatorships; and using TIMSS data to reform mathematics education.

12.7 Quality Assurance in Publications

Considerable effort is taken to ensure that all IEA studies are of high quality. It is important that they are methodologically rigorous and educationally relevant. The fundamental way of achieving this is to establish study teams which are highly competent and conduct their studies in a rigorous and transparent way. Each IEA study is supported by an array of expert committees, with regular reports to the IEA Standing Committee (this committee comprises six national representatives elected by the IEA General Assembly membership and serves as a board of directors for IEA) and to national research coordinators from all participating countries. When testing takes place within a country, quality monitors visit schools to help ensure that all procedures are followed correctly. In addition, the IEA convenes a Technical Executive Group (TEG), a small committee of expert psychometricians and methodologists, who advise on the technical aspects of all projects.

A further level of quality assurance is provided by the IEA Publications and Editorial Committee (PEC). This is a group of scholars from around the world who review documents prior to publication. PEC includes both IEA and non-IEA scholars to ensure an adequate measure of independent scrutiny. PEC provides expert, robust feedback to authorial teams and helps to ensure excellence in IEA publications.

12.8 Public Dissemination of IEA's Work

IEA is highly respected among the research community but often not well known outside of academic circles. To address this gap, since 2015, IEA has placed a key strategic focus on communicating IEA research, data, and study findings to diverse audiences, including educators, policymakers, media, and the public while also retaining researchers as a main audience for IEA.

The IEA website (www.iea.nl) is the primary source of information about IEA studies, new publications, activities, and opportunities. The news and events sections are updated regularly and website users have the opportunity to subscribe to a quarterly email newsletter, *IEA Updates*, for further information.

IEA has an active social media presence across Twitter, LinkedIn, Facebook, and YouTube (see www.iea.nl for links), using these platforms to share information about IEA, to support the network of IEA study participants (for example, by sharing

external publications based on IEA data), and to communicate IEA study results. In particular, IEA focuses on targeted social media posts to share findings that are related to current events. This helps to bring IEA data and resources to new audiences and to expand beyond the existing networks of people who are already aware of IEA's activities.

A range of promotional materials provide support for IEA activities, including the annual *IEA Insider* publication, study brochures, promotional flyers, infographics and short videos (IEA 2020a). These materials are shared by IEA staff and collaborators alike as part of wider networking and outreach initiatives.

Promoting IEA publications (study-specific publications, the journal, IEA Research for Education series and IEA Compass: Briefs in Education series) is another key area of activity. New publications are shared on the website, social media, and newsletter channels in addition to being promoted at conferences and other events. New IEA publications provide a basis for producing articles aimed at more general audiences. These include blog posts and contributions to external platforms.

IEA hosts information stands at international conferences to promote the wealth of data, publications, and training opportunities that are available to the education research community. In some cases, researchers may be aware of some of IEA's studies but unaware that the IEA is the organization behind those studies. Promoting all IEA activities together (studies, data, publications, and research services) helps to highlight the full range of what IEA can offer to the research community.

In addition to contributing to external events, IEA-specific events also promote the work. New international reports from IEA studies are released as part of large, media-focused events, usually held in partnership with external organizations such as UNESCO. IEA produces and disseminates press releases, infographics, and videos to promote the key findings of each study to diverse audiences. All materials are shared with participating countries so that they may be translated and adapted for national use. Any media outreach or interviews are always conducted in cooperation with the participating countries.

IEA's communications and dissemination activities shine a light on the wealth of high-quality, international data and research expertise available as valuable resources for diverse audiences. With these activities, IEA aims to bridge the gap between academic research, and education policy and practice.

12.9 Conclusions

IEA studies generate a vast body of publications: study findings and documentation, academic and policy papers based on the studies, training materials, and dissemination documents. IEA's own documents (IEA 2020a) are subject to rigorous review prior to publication, and papers in academic journals are subject to normal blind peer review. IEA's portfolio of publication activity, which is constantly growing, constitutes an enormous contribution to comparative educational research and the science of educational measurement.

References

Fraillon, J., Ainley, J., Schulz, W., Duckworth, D., & Friedman, T. (2019). *IEA International Computer and Information Literacy Study 2018 assessment framework*. Amsterdam, the Netherlands: IEA. https://www.iea.nl/publications/assessment-framework/icils-2018-assessment-framework.

IEA. (2020a). Publications [webpage]. Amsterdam, the Netherlands: IEA. https://www.iea.nl/publications.

IEA. (2020b). Data repository [webpage]. Amsterdam, the Netherlands: IEA. https://www.iea.nl/data-tools/repository.

IEA. (2020c). Tools [webpage]. Amsterdam, the Netherlands: IEA. https://www.iea.nl/data-tools/tools.

IEA. (2020d). Training [webpage]. Amsterdam, the Netherlands: IEA. https://www.iea.nl/research-services/training.

IEA. (2020e). IEA Research for Education series [webpage]. Amsterdam, the Netherlands: IEA. https://www.iea.nl/publications/series-journals/iea-research-education.

IEA. (2020f). IEA awards [webpage]. Amsterdam, the Netherlands: IEA. https://www.iea.nl/about/opportunities/award.

IEA. (2020g). Large-scale Assessments in Education journal [webpage]. Amsterdam, the Netherlands: IEA. https://www.iea.nl/publications/series-journals/large-scale-assessments-education.

IEA. (2020h). IEA international research conferences [webpage]. Amsterdam, the Netherlands: IEA. https://www.iea.nl/news-events/irc.

IEA. (2020i). IEA Compass: Briefs in Education series. [webpage]. Amsterdam, the Netherlands: IEA. https://www.iea.nl/publications/series-journals/iea-compass-briefs-education-series.

Mullis, I. V. S., & Martin, M. O. (Eds.) (2015). *PIRLS 2016 assessment framework* (2nd ed.). Chestnut Hill, MA: Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://www.iea.nl/publications/assessment-framework/pirls-2016-assessment-framework.

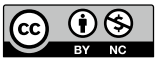
Mullis, I. V. S., & Martin, M. O. (Eds.) (2017). *TIMSS 2019 assessment frameworks.* Chestnut Hill, MA: Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://www.iea.nl/publications/assessment-framework/timss-2019-assessment-frameworks.

Mullis, I.V.S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics.* Chestnut Hill, MA: Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://www.iea.nl/publications/timss-2015-international-results-mathematics.

Schulz, W., Ainley, J., Fraillon, J., Losito, B., & Agrusti, G. (2016). *IEA International Civic and Citizenship Education Study 2016 assessment framework.* Amsterdam, the Netherlands: IEA. https://www.iea.nl/publications/assessment-framework/iea-international-civic-and-citizenship-education-study-2016.

Seamus Hegarty is Chair of the IEA Publications and Editorial Committee. He served as Chair of the IEA from 2005–12. He is a visiting professor at the University of Warwick. He was Director of the National Foundation for Educational Research for twelve years until his retirement in 2005. He has written and co-authored more than 20 books plus numerous reports and papers. He has edited the *European Journal of Special Needs Education* since founding it 35 years ago. He edited *Educational Research* for 21 years and has served on editorial boards for numerous international journals.

Sive Finlay led the IEA Communications team from 2018–2020. She was responsible for IEA's communications strategy and supporting IEA staff, members, and partner organizations with their outreach and dissemination activities, in addition to managing IEA's online presence and social media accounts.



Chapter 13
Consequential Validity: Data Access, Data Use, Analytical Support, and Training

Sabine Meinck, Eugenio Gonzalez, and Hans Wagemaker

Abstract The data collected by the International Association for the Evaluation of Educational Achievement (IEA) represent a vast resource of information on the functioning of the participating educational systems and provide a form of public accountability that should be made accessible to the wider research community. In keeping with IEA's mission to disseminate its research findings and develop the broader research community, its data is generally made publicly available. Furthermore, in order to address potential concerns regarding consequential validity (correct data use and interpretation), IEA has developed a training and capacity building strategy that includes several key components related to data access, training and capacity building, and the stimulation and support of secondary analysis. In brief, IEA data is accompanied by detailed technical documentation on how it was collected and processed. User guides provide additional information on the correct application of statistical analysis methods when using this data for in-depth analysis and, with the IEA International Database (IDB) Analyzer, IEA provides an analytical software tool that automatically accounts for the complexities of the data structure. Researchers are offered training in how to analyze data correctly as part of a broad portfolio of workshops designed for learning about fundamental and advanced concepts of large-scale comparative assessments of achievement in education.

Keywords Analysis · Capacity building · Consequential validity · Data access · Dissemination · Publication · Workshops

S. Meinck (✉)
International Association for the Evaluation of Educational Achievement (IEA), Hamburg, Germany
e-mail: sabine.meinck@iea-hamburg.de

E. Gonzalez
Educational Testing Service, Princeton, NJ, USA
e-mail: egonzalez@ets.org

H. Wagemaker
International Association for the Evaluation of Educational Achievement (IEA), Amsterdam, The Netherlands
e-mail: hanswagemaker@compuserve.com

H. Wagemaker (ed.), *Reliability and Validity of International Large-Scale Assessment*, IEA Research for Education 10, https://doi.org/10.1007/978-3-030-53081-5_13

13.1 Introduction

As outlined in Chaps. 1 and 2, concerns regarding validity and reliability do not end with the successful collection of data. One of the primary ongoing concerns for IEA is that the data collected for each of the studies is used to inform educational reform and improvement. While data are initially provided for and made available to the participating countries' analysts, these data represent a vast resource of information on the functioning of the participating educational systems and provide a form of public accountability, which can and (in IEA's view) should be made accessible to the wider research community.

Users of IEA data have access to not only the learning outcomes of the target populations that participate in the study but also a vast array of background information about the nature of the education systems in each country, the participating schools and the teachers that provided instruction, the students' home backgrounds, their demographics, and, in the case of some of the assessments, background information from students' parents or guardians.

In order to achieve its broader mission of educational improvement and the development of the community of educational researchers, IEA has taken a number of actions to not only facilitate the use and accessibility of the data it produces but also ensure that data are used appropriately.

The following sections outline some of the key strategies that IEA employs to support a positive outcome in terms of consequential validity arguments. They include the development of analytical tools, facilitating access to data, supporting training on the use of the data, and scientific collaboration and exchange.

13.2 Data Access

As noted, IEA explicitly encourages researchers around the world to use the data collected from its studies for further in-depth analysis, making significant efforts to provide data access along with comprehensive documentation. Data gathered in IEA studies is available via the IEA website (www.iea.nl) and can be downloaded from IEA's data repository (IEA 2020a). Comprehensive technical documentation, including data code books, accompany the data. This documentation enables researchers to navigate the contents of the international databases (IDBs), and understand the methods used to collect, process, and structure the data. The documentation also presents information on how achievement and background scales were calculated and how these should be used for analysis. The materials provide users with sufficient information to conduct secondary analysis that go beyond the mostly descriptive scope of the primary publications. By providing the data at individual and variable level, they also allow users to calculate new variables or scales. For example, researchers may want to recode variables by collapsing categories, or

create new indexes. Expert users may reproduce or engage in rescaling of achievement items, modifying the conditioning model by entering new variables merged from other data sources, or optimizing the item response theory (IRT) model for specific (groups of) countries.

IEA data are made available in both SAS (SAS Institute Inc. 2013) and SPSS (IBM Corporation 2016) formats. These formats can be easily converted into file formats usable for various other contemporary statistical software packages handling international large-scale assessment (ILSA) data such as R (R Core Team 2014), STATA (StataCorp LLC 2019), Mplus (Muthén and Muthén 2012), or WesVAR (Westat Inc. 2008), to name the most popular. To keep the structure clear, data is made available in individual files for different countries and data sources. Algorithms for merging and appending data for specific analysis are explained in the accompanying user guides. Codebooks provide detailed information on each variable in the database. These display names, labels, and valid ranges, describe field values and missing schemes, and variable allocation to specific domains. Even though many of the actual achievement items cannot be published (items that are to be reused in future cycles for trend measurement purposes have to be maintained secure), item responses and descriptive information are available for all items, sufficient for additional item or scale level analysis. Released items (i.e., those that will not be used in future cycles) can also be obtained from the IEA by completing and submitting a permission request form. To protect the identity of those participating in the studies, some variables are omitted from the public use data files, as they bear a low, but non-negligible risk of disclosing information about single schools or even individuals participating in an IEA study. As IEA assures full anonymity to participants and commits to data security and confidentiality laws, such as the European Union's General Data Protection Regulation (GDPR; see https://eugdpr.org/), suppressing information that would potentially identify participants is a priority. Researchers interested in using the variables not included in the public dataset for analysis can apply for access to the restricted use data. Applicants need to provide comprehensive information about how they intend to use the data and sign an agreement not to disclose details of any school or individual.

Additional support for analysts is provided through syntax in the format of ready-to-run programs that allow items to be scored or specific databases to be merged. Depending on the study and scope, additional information is accessible, such as national context survey data and encyclopedias, item almanacs presenting item parameters and descriptive statistics, or data on countries' curricula and how they match with the tests.

Meticulous study documentation supplementing the data is also freely available. Each IEA study produces its own technical report, detailing information about the instruments, sampling, survey implementation, and analytic procedures used to derive scale scores and indexes. User guides complement this information, giving users a comprehensive overview of the structure and content of the international database of each study, the main features, analytical possibilities, and limitations. The user guides also contain the international versions of the contextual questionnaires, together with an overview of national adaptations, example items, and information

on derived variables; the user guides are excellent resources to familiarize users with the data and studies.

Data files and technical documentation are also available from each study's webpages. For example, users can find the IEA's Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) data at their dedicated website (https://timssandpirls.bc.edu), along with results and technical information for all cycles.

IEA also supports another important website, designed as a central point of entry to international large-scale assessments in education: the ILSA Gateway (https://ilsa-gateway.org). Here, users can find comprehensive information on all contemporary ILSAs. The information available is not limited to IEA studies, but also includes studies commissioned by other international organizations. The ILSA Gateway includes direct links to data and documents on the various study websites, a database of study-related research papers, and announcements of recent news and upcoming events. The Gateway was designed for a worldwide audience, first and foremost researchers, but is also a useful tool for policymakers, decision makers, and all others interested in education research. It is intended to facilitate access to ILSA-related resources, encourage knowledge exchange and discussions, and inspire future research.

Finally, IEA provides researchers with support related to the use of data from IEA studies. The IEA website contains contact information on support options, and interested researchers are encouraged to make contact when they need guidance.

13.3 Facilitating Analysis: The IEA IDB Analyzer

The ILSAs conducted by IEA are also known as group-score assessments because they are designed with the purpose of reporting results for different groups of interest within a population, and not at the individual level. This population is generally a specified group of students within a country, but in some cases, it could be those within a jurisdiction of interest, such as a municipal school system or a subnational system.

The qualification "large-scale" refers to the broadness of the domains covered, and the size of the population represented by those participating in the assessment, but not necessarily to the size of the sample or the length of the instrument administered to those taking part in the assessment. As such, ILSAs rely on the evaluation of subsets of individuals from the population with subsets of items from the domain, in order to make inferences to the entire population about the entire domain assessed. This is what is called multiple matrix sampling (Gonzalez and Rutkowski 2010; Lord 1962; Shoemaker 1973). Using multiple matrix sampling results in great efficiencies in the assessment process since only some of the people have to be assessed using some of the items. This significantly reduces the burden on the individual participants (due to reduction in the assessment time) and on the organizations implementing the assessment (due to reduction in the number of individuals that need to be assessed).

But, as a consequence, these design features need to be accounted for, and sampling weights, complex variance estimation methods, and plausible values have to be used to analyze the resulting data with the corresponding adjustments.

The IEA International Database Analyzer (IDB Analyzer; IEA 2020b) is an application that can be used to combine and analyze data from IEA's large-scale assessments, as well as data from most major large-scale assessment surveys.

The IDB Analyzer creates SPSS or SAS syntax that can be used to combine files from across different countries and levels (e.g., student, parent, teacher, or school), and perform analyses with these international databases that take into account the sample and assessment designs. It generates SPSS or SAS syntax that takes into account information from the sampling design in the computation of sampling variance and handles the plausible values. The code generated by the IDB Analyzer enables the user to conduct a variety of analyses including, at the most basic level, descriptive statistics, and also allows users to conduct statistical hypothesis testing among groups in the population without having to write any programming code.

The IDB Analyzer is licensed free of charge and available for use in accordance with the terms of the licensing agreement. The IDB Analyzer consists of two modules: the Merge Module and the Analysis Module (Table 13.1).

When calculating these statistics, the IDB Analyzer is capable of using any continuous or categorical variable in the database, or can make use of achievement scores in the form of plausible values. When using plausible values, the IDB Analyzer generates code that takes into account the multiple imputation methodology in the calculation of the variance for statistics as it applies to the corresponding study.

All procedures offered within the Analysis Module of the IDB Analyzer make use of appropriate sampling weights and standard errors of the statistics that are computed according to the variance estimation procedure required by the design as it applies to the corresponding study. For more information about the workings of the IDB Analyzer, please refer to the Help Manual for the IDB Analyzer included with the software (IEA 2020b).

Table 13.1 Functions of the IDB analyzer

Module	Function
Merge module	Combines published SAS or SPSS data files from different countries, and, when necessary, merges data files from different sources (like student background questionnaires and achievement files, or student background files with teacher or school level files). Allows the user to select individual or sets of variables to create a smaller and more manageable dataset. The data files created using the Merge Module can be processed either with the Analysis Module of the IDB Analyzer or with any other analysis software that would account for the complex design and accepts SPSS or SAS files as input

(continued)

Table 13.1 (continued)

Module	Function
Analysis module	Provides procedures for the computation of several statistics, for any variable of interest overall for a country, and for specific subgroups within a country. Can be used to analyze data files from any IEA study, regardless of whether they have been preprocessed with the Merge Module of the IDB Analyzer. Can create code for several statistical procedures. Like the Merge Module, the Analysis Module creates SPSS and SAS code that computes the statistics specified by the user The following analyses can be performed with the Analysis Module of the IDB Analyzer: *Percentages and Means*: Computes percentages, means, design effects and standard deviations for selected variables by subgroups defined by the user. The percent of missing responses is included in the output. It also performs the computation of t-test statistics of group mean and percent differences taking into account sample dependency *Linear Regression*: Computes linear regression coefficients for selected variables predicting a dependent variable by subgroups defined by the user. The IDB Analyzer has the capability of including plausible values as dependent or independent variables in the linear regression equation. It also has the capability of contrast coding categorical variables (dummy or effect) and including them in the linear regression equation *Logistic Regression*: Computes logistic regression coefficients for selected variables predicting a dependent dichotomous variable, by subgroups defined by the user. The IDB Analyzer has the capability of including plausible values as independent variables in the logistic regression equation. It also has the capability of contrast coding categorical variables and including them in the logistic regression equation. Under the SAS version multinomial logistic regressions can be run *Benchmarks*: Computes the percent of the population in two modes: (a) as the percent within a group meeting a set of user-specified performance or achievement benchmarks, or (b) as the percent within a benchmark that belong to a specific group of the population. It computes the percentages within a group meeting the benchmarks in two modes: cumulative (percent of the population at or above given points in the distribution) or discrete (percent of the population within given points of the distribution). It can also compute the mean of an analysis variable for those within a particular group when the discrete option is selected. In doing this, it performs the computation of group mean differences and percent differences between groups taking into account sample dependency *Correlations*: Computes correlation for selected variables by subgroups defined by the grouping variable(s). The IDB Analyzer is capable of computing the correlation between sets of plausible values 1. *Percentiles*: Computes the score points that separate a given proportion of the distribution of scores by subgroups defined by the grouping variable(s)

13.4 Capacity Building: Workshops

Conducting ILSAs in education with the highest quality standards is always a collaborative endeavor, involving many actors at international and national levels. It can only be successful if all contributors have acquired thorough knowledge about their specific tasks within the project. From the early years, IEA trained staff from national teams to achieve standardized implementation of the assessments in all participating countries, thereby building research capacities that were often used to develop and implement national educational monitoring systems. However, IEA realized that many countries lacked the capacity to thoroughly analyzing the data collected from the assessments, given its complex structure and statistical constraints. Often, the studies' value to national policymakers and educational stakeholders was limited to translating the international reports into national languages, when the available data could instead be further exploited to address specific issues of national interest; in other words, IEA data was underused. This inspired IEA to offer a broad portfolio of learning opportunities by means of professional development workshops, conducted by experts in the field, and designed to cater all interested in learning about the concepts underpinning ILSA. The key objective of these workshops is to provide researchers with sufficient skills and knowledge to use the data collected optimally.

IEA workshops are arranged in thematic areas ranging from introductory courses for designing surveys, to the implementation of complex survey operations procedures. IEA also offers workshops on using various complex statistical methods, and report writing and publishing. Each theme consists of a series of both basic and advanced courses, often building on each other. In response to latest research developments in the field, workshop content is continually updated (see Table 13.2).

While workshops build research capacity, they also provide a forum for fruitful exchange of ideas and foster the development of research networks. The methods applied in IEA workshops are distinctively selected to accommodate this goal. All workshops are conducted as in-person training (i.e., participants and trainers meet in one location, as opposed to remote training methods, such as webinars). Each workshop combines lectures and tutorials with opportunities for participants to not only apply and practice what they have learned but also discuss emerging issues with each other. Experienced instructors are on hand to provide guidance and advice to participants and explicitly encourage exchange and collaboration. Participants share their ideas, approaches, and findings with the group, resulting in an inspiring atmosphere. Workshops benefit from the fact that the trainees come usually from multicultural backgrounds, bringing multiple perspectives on the topics. Catering to this audience, workshops are mostly conducted in English, but some are also available in German, Spanish, Arabic, and French (a workshop brochure can be download from the IEA website; IEA 2020c).

Workshops can be tailored to specific occasions, topics, studies, and/or audiences, and vary in length, from a 90 min introductory course to a one-week long intensive training course. Workshops are conducted at international conferences, at focused academies, or on demand; some are anchored within specific capacity-building projects.

Table 13.2 Examples of IEA workshops

Theme	Topics
International large-scale assessment: Methods, survey design, and implementation	Workshops conducted under this theme provide an overview of all aspects and challenges experienced in conducting a cross-national assessment in education. They cover topics such as item and questionnaire development, test design, sampling, and survey operation procedures. Workshops under this theme build up a fundamental understanding of all steps of an assessment cycle
Statistical methods	Workshops covered under this theme include basic training on quantitative analysis, analysis with large-scale assessment data, advanced modeling methods (such as multilevel modeling, structural equation modeling, and multilevel structural equation modeling), and statistical techniques to create scales and indexes (such as item response theory). These workshops are directed primarily to those intending to use the data for in-depth analysis
Effective writing and publishing	Workshops training participants in effective writing and publishing focus on writing comprehensive thematic reports, policy briefs, and academic papers based on the analysis of data from large-scale assessments. Participants are trained how to structure different types of publications, how to write efficiently, and how to address various audiences. The focus is on ensuring effective dissemination of the results to target stakeholders and using IEA data correctly to address local policy concerns

13.4.1 Promoting High-Quality Research Based on Large-Scale Assessment Data

In keeping with its stated aims and commitment to the development of the wider research community with a particular focus on ILSA, IEA has initiated a series of publications and actively supports other research related activities. IEA publishes its own journal, *Large-scale Assessments in Education*, in association with Educational Testing Service (ETS; www.ets.org) under the brand of the IEA-ETS Research Institute (IERI) (see Sect. 12.4), conducts its own focused biennial research conference (see Sect. 12.5), and encourages scholars in their work by distinguishing outstanding research with special awards (see Sect. 12.3). IEA also funds researchers to conduct comprehensive in-depth analysis of IEA data; the results are published in the Springer open access series, IEA Research for Education (see Sect. 12.3). Finally, with the IEA

Compass: Briefs in Education series (see Sect. 12.6), IEA established a special publication format addressing policymakers and other stakeholders in education outside the scientific research community. Each publication in the series aims to connect research findings to recurrent and emerging questions in education policy debates at the international and national levels. Using appropriate text formats and illustrations, the series intends to translate these findings for an audience with less technical backgrounds, making sure the results reach broader target audiences.

As reported in Chap. 12, all publications undergo rigorous review procedures at several stages, involving international renowned experts in the respective fields. One important review stage is conducted by IEA's Publication and Editorial Committee (PEC; see Sect. 12.7). In this way, IEA attempts to provide support and guidance to researchers in a way that helps to ensure the study data is being used appropriately. Good examples of methodologically sound or extraordinary innovative research are highlighted via IEA's social media platforms, by inviting authors for special publication formats, and by awarding prizes.

13.4.2 The IEA-ETS Research Institute (IERI)

In 2007, the IEA partnered with ETS to form IERI. IERI is a collaborative effort between the Research and Development Division at ETS and the IEA that focuses on improving the science of large-scale assessments. IERI undertakes activities around three broad areas of work that include research studies related to the development and implementation of large-scale assessments, professional development and training, and dissemination of research findings and information gathered through large-scale assessments. As part of the activities of IERI, a monograph series was started in 2007, which later became the online open journal *Large-scale Assessments in Education* (see Sect. 12.4).

This journal focuses on articles that contribute to the science of large-scale assessments, help disseminate state-of-the-art information about empirical research using these databases and make the results available to policymakers and researchers around the world. Submissions to the journal have to undergo and receive favorable technical, substantive, and editorial review prior to acceptance and publication. The journal uses a double-blind peer-review system, where the reviewers do not know the names or affiliations of the authors, and the reviewer reports provided to the authors that assess the quality of their manuscript are also anonymous. Independent researchers in the relevant research area are chosen to assess submitted manuscripts for originality, validity, and significance.

13.4.3 IEA International Research Conference

At the turn of the millennium, even though IEA had been active in the field of educational research for several decades, IEA study related research publications and presentations at notable international or national conferences were relatively rare. Early efforts to change this led to symposia initiated by smaller research networks,[1] but there was no dedicated platform for presenting and sharing IEA study findings. To foster collaboration among scholars using IEA data for their research, with the initiative and leadership of Constantinos Papanastasiou,[2] IEA launched its first International Research Conference (IRC). This first conference took place in Lefkosia, Cyprus, in 2004 and has followed a generally biennial cycle since.[3] The purposes and goals of the event were set high from the very beginning. According to Papanastasiou and Papanastasiou (2004, p. 202), "the aim of the IRC-2004 Conference was to provide an international forum for the exchange of ideas and information on critical … issues in education evolving from secondary analyses of data from IEA studies. With its objective to foster creative dialogue among scholars and researchers, the conference aimed at providing greater understanding of the numerous roles that education plays in the development of nations and in shaping individuals. Because of its international scope, the conference also aimed to examine issues in both a comparative and global context with the ultimate aim to enhance pedagogical knowledge and implement positive change. A final goal of this conference was to create a global network of researchers." These initial goals of the conference remain, however, the conference has significantly broadened its contents and scope over time.

Since 2008, the two days prior to the conference have been dedicated to capacity building. Experts in the field educate participants in various, often methodological topics related to ILSA. Statistical software and analysis approaches are introduced, and issues related to the dissemination of results to varying audiences are tackled.

Due to the increasing amount of high-quality proposals, a poster session format was introduced in 2017, together with a poster award, particularly designed to encourage early career researchers to present their work. Symposia were introduced in 2015, as well as panels on policy impact and on facilitating the use of IEA data for practitioners.

Another important development starting in 2015 was the inclusion of methodological focus sessions. New approaches related to the science of large-scale assessments are discussed in these sessions, fostering further development and implementation of new methods of analysis and stimulating critical exchange among international scholars about the applied methodologies used in ILSAs.

---

[1] For example, Network 9 (Assessment, Evaluation, Testing and Measurement) of the European Educational Research Association (EERA).

[2] Constantinos Papanastasiou was professor at the University of Cyprus, and Cyprus's representative at the General Assembly of IEA.

[3] An exception was made in 2012 (the conference was delayed to 2013), as the reporting phase for TIMSS and PIRLS resulted in high workloads for many of the stakeholders involved, and would have jeopardized the success of the conference.

There are some significant factors distinguishing the IRC from other research conferences in the educational field. Most of these factors are related to its highly focused nature, conditioned by its restrictive content (proposals must make use of IEA data or be directly related to IEA study methodology). This focus strongly enhances the opportunities for networking: all attendees are to some extent familiar with the studies, which sets the basis for efficient exchanges. Participants can connect with peers who may have encountered similar problems and can share solutions or find new partners for future projects. Moreover, data users are connected with data producers; the experts involved in designing the methodologies for data collection and analysis for the IEA studies are present at the conference. They can give users invaluable hints and insights about the studies and, at the same time, benefit from the outsider perspectives of the scholars making use of their data.

Information on past and upcoming IEA Research Conferences can be obtained from the IEA website (IEA 2020d).

13.4.4 Academic Visitors/Scholars

On a regular basis, IEA invites scholars to work within IEA's premises under its academic visitor programs. The programs are partly collaborative efforts with other internationally renowned institutions or organizations, and can be viewed as part of IEA's efforts to contribute to the development of a worldwide network of researchers in educational assessment and evaluation. Interested academics and early research fellows can apply for the programs with a research proposal using IEA data. If selected, they develop their own research project while benefiting from the individual support of IEA experts working in different fields related to ILSAs. More information on the visitor program can be found on the IEA website (IEA 2020e).

13.4.5 IEA Awards

One of the earliest attempts to stimulate the proper use of IEA data was the establishment of a series of research awards (see also Sect. 12.3). These include the Bruce Choppin award (which recognizes a Masters or Doctoral level dissertation that employs empirical research methods with IEA data; the Richard M. Wolf award (which recognizes the author or authors of a paper published in a refereed journal, monograph or book based on analysis of IEA data), and, more recently, the Constantinos Papanastasiou poster prize (which is presented at IEA's biennial research conference).

13.5 Conclusions

In order to address potential concerns related to issues of consequential validity, namely the correct use and interpretation of the data, IEA has developed a successful and elaborate strategy for addressing these concerns, which includes several key components. These relate to user-friendly data access, meticulous technical documentation, and a strong focus on training and capacity building. IEA also stimulates secondary analysis of the data through its awards, journal, policy briefs, and the research conference (see also Chap. 12) to ensure that the valuable information collected by IEA's ILSAs contributes effectively to educational policy reform and improvement.

References

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. In M. von Davier & D. Hastedt (Eds.), *Issues and methodologies in large-scale assessments*, IERI Monograph Series Volume 3 (pp. 125–156). Hamburg, Germany: IERI. http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_03_Chapter_6.pdf.

IBM Corporation. (2016). IBM SPSS Statistics for Windows (Version 24.0) [Computer software]. Amonk, NY: IBM Corp. https://www.ibm.com/analytics/spss-statistics-software.

IEA. (2020a). Data repository [webpage]. Amsterdam, the Netherlands: IEA. https://www.iea.nl/data-tools/repository.

IEA. (2020b). IDB Analyzer (Version 4) [computer software]. Hamburg, Germany: International Association for the Evaluation of Educational Achievement (IEA). https://www.iea.nl/data-tools/tools.

IEA. (2020c). Training [webpage]. Amsterdam, the Netherlands: IEA. https://www.iea.nl/research-services/training#section-6.

IEA. (2020d). IEA International Research Conferences (IRC) [webpage]. Amsterdam, the Netherlands: IEA. https://www.iea.nl/de/node/2629.

IEA. (2020e). Academic vistors [webpage]. Amsterdam, the Netherlands: IEA.https://www.iea.nl/about/opportunities/academic-visitor-program.

Lord, F. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement, 22*(2), 259–267.

Muthén, L. K., & Muthén, B. O. (2012). Mplus 7 [Computer software]. Los Angeles, CA: Muthén & Muthén. https://www.statmodel.com/.

Papanastasiou, C., & Papanastasiou, E. C. (2004). IEA international research conferences (IRC) from 2004 to 2008. In C. Papanastasiou, T. Plomp, & E.C. Papanastasiou (Eds), *IEA 1958–2008: 50 years of experiences and memories* (pp. 201–213). Amsterdam, the Netherlands: IEA. https://www.iea.nl/publications/iea-reference/iea-1958-2008.

R Core Team. (2014). The R project for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

SAS Institute Inc. (2013). SAS university edition 9.4 [Computer software]. Cary, NC: Author. https://www.sas.com/en_us/software/university-edition.html.

StataCorp LLC. (2019). Stata 16 [Computer software]. College Station, TX: Author. https://www.stata.com/why-use-stata/.

Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling*. Cambridge, MA: Ballinger Publishing Company.

Westat Inc. (2008). WesVar Version 5.1: Replication-based variance estimation for analysis of complex survey data [Computer software]. Rockville, MD: Author. https://www.westat.com/capability/information-systems-software/wesvar/download.

Dr. Sabine Meinck works for the IEA in Hamburg, Germany, being head of both its Research and Analysis Unit and Sampling Unit. Since 2006, she has been involved with the sampling, weighting, variance estimation, and analysis of nearly all contemporary large-scale assessments in education. Her experience as a member of the project management teams for IEA's TIMSS and PIRLS, with the consortia of the IELS and TALIS Starting Strong Surveys, and on the joint management committees of IEA's ICILS, ICCS, ECES, and TEDS-M, have enabled her to develop a diverse knowledge and expertise; she also serves on the board of the IERI Institute.

Dr. Meinck coordinates, guides and supports all research activities within the IEA. Her main research interest lies with the science of cross-national large-scale assessments, and the methodological challenges of complex survey data.

In support of the IEA's enduring commitment to knowledge dissemination, Dr. Meinck has conducted multiple workshops for international audiences designed to share her experiences, and teach best practices and methodologies in educational research. Topics taught range from basic to advanced statistical methods, survey design, and publication and dissemination strategies for diverse audiences. Further, she teaches a Masters Course at the University of Hamburg on "Quantitative methods in educational research." Dr. Meinck is associate editor of the Springer journal Large-scale Assessments in Education. She is honored to serve as a peer reviewer for several scientific journals on educational research, and many educational research networks (such as AERA and CIES).

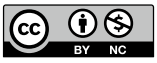
Eugenio Gonzalez is a Principal Research Project Manager at Educational Testing Service (ETS), and director of the IEA-ETS Research Institute (IERI), a collaborative effort between the International Association for the Evaluation of Educational Achievement (IEA) and ETS that focuses on improving the science of large-scale assessments. IERI undertakes activities around three broad areas of work that include research studies related to the development and implementation of large-scale assessments; professional development and training; and dissemination of research findings and information gathered through large-scale assessments. Dr. Gonzalez is also responsible for the technical documentation and international database training activities for PIAAC and PISA.

Dr. Gonzalez was formerly head of the Research and Analysis Unit at IEA Hamburg (2007–2012), the director of quality control and field operations for the National Assessment of Educational Progress (NAEP) (2004–2006), and director of international operations and data analysis in the TIMSS & PIRLS international study center (ISC) at Boston College (1994–2004). In this last role, he oversaw the development and implementation of international operations, data analysis, and reporting procedures for the Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS). Since 1993, he has conducted database training activities for several research and governmental organizations, and has been a developer and technical lead of customized software for analyzing international large-scale assessment databases such as the IEA's IDB Analyzer and the Data Explorer. Dr. Gonzalez has also served as external consultant to several national and international large scale assessment programs, and has held teaching appointments at Boston College and the University of Massachusetts, Amherst. He has a PhD in Educational Research, Measurement, and Evaluation from Boston College, and an undergraduate degree in Psychology from the Universidad Católica Andres Bello in Caracas, Venezuela.

Hans Wagemaker was the executive director of the International Association for the Evaluation of Educational Achievement for 17 years, responsible for the management of all IEA international research and assessment projects and activities. He helped develop IEA's Progress in International Reading Literacy Study (PIRLS) and oversaw the development and expansion of IEA's training and capacity building activities in low to middle income countries, and IEA's educational consultancy services. Together with Educational Testing Services (ETS), he established the IEA-ETS Research Institute (IERI), where he continues to serve as a Board member.

Dr. Wagemaker was a Senior Manager Research and International with the Ministry of Education, New Zealand, and represented New Zealand's interests in the APEC Education Forum, UNESCO's commissions, and the OECD, CERI, and the Education Governing Board. He has consulted for the Inter American Development Bank and UNESCO and worked extensively with the World Bank to advance a common interest in the uses of assessment for improving educational systems in developing countries. Most recently Dr. Wagemaker served as an advisor to the Minister of Education for the Sultanate of Oman. He is also a member of the Advisory Board for the Center for Education Statistics and Evaluation (CESE) for the government of New South Wales, Australia, the H Institute, Beirut, Lebanon, and continues in an advisory role with the IEA.

Dr Wagemaker holds BA and MA degrees from the University of Otago, New Zealand, and a PhD from the University of Illinois, where he was awarded a University Fellowship and, in 2009, the College of Education's Distinguished Alumni Award.



Chapter 14
Using IEA Studies to Inform Policymaking and Program Development: The Case of Singapore

Hui Leng Ng, Chew Leng Poon, and Elizabeth Pang

Abstract Singapore's participation in the Second International Science Study (SISS) in the 1980s marks the beginning of nearly four decades of involvement in IEA studies. International large-scale assessments (ILSAs) play an important role in complementing other information sources to inform policymakers about different aspects of the Singapore education system. The insights derived from these studies have at times served as reassurances to policymakers and program designers that progress has been made in some areas and, at other times, identified where improvements could be made. Participating in ILSAs has proved useful for Singapore. The Singapore Ministry of Education has used data from PIRLS and TIMSS for system-level monitoring and secondary analyses of the data have provided insights to inform policymaking and program development; here three actual use cases are used to illustrate the impact of ILSA in Singapore. These cases cover uses ranging from catalyzing curriculum redesign, to monitoring the implementation of a new pedagogical approach to learning science, to keeping tabs on any trade-offs from the bold, system-wide curricular and pedagogical shifts adopted. As a result of this long history of participation, the Singapore Ministry of Education has developed general principles guiding the use of data from large-scale assessments, which provide a useful framework for those interested in international efforts to improve education for all.

Keywords Curriculum redesign · Evaluation · Monitoring · Pedagogical change · Progress in International Reading Literacy Study (PIRLS) · Policymaking · Reforms · Trends in International Mathematics and Science Study (TIMSS)

H. L. Ng (✉) · C. L. Poon · E. Pang
Ministry of Education, Singapore, Singapore
e-mail: ng_hui_leng@moe.gov.sg

C. L. Poon
e-mail: poon_chew_leng@moe.gov.sg

E. Pang
e-mail: elizabeth_pang@moe.gov.sg

H. Wagemaker (ed.), *Reliability and Validity of International Large-Scale Assessment*,
IEA Research for Education 10, https://doi.org/10.1007/978-3-030-53081-5_14

14.1 Introduction

Singapore's participation in the Second International Science Study (SISS) in the 1980s marks the beginning of nearly four decades of involvement in studies undertaken under the auspices of the International Association for Evaluation of Educational Achievement (IEA). Because of the methodological and process rigor that goes into each IEA study (as detailed in the other chapters in the book), the high-quality data generated from these studies are valuable resources that Singapore uses in ongoing efforts to improve the quality of education.

In this chapter, we first discuss the value of large-scale international studies to Singapore, including IEA's Progress in International Reading Literacy Study (PIRLS) and Trends in International Mathematics and Science Study (TIMSS). We then describe how the Singapore Ministry of Education (MOE) has used PIRLS and TIMSS data for system-level monitoring and derived insights from secondary analyses to inform policymaking and program development. We illustrate such data use with specific examples. We conclude by distilling some guiding principles that underpin the ways in which MOE uses the data.

14.2 Why Singapore Participates in International Large-Scale Assessments

Education has always been highly valued in Singapore. For the people, education is perceived as a means to a better life. For policymakers, it is a key strategic lever to ensure the economic survival of a small nation with few natural resources other than its population. An urgent nation-building task in the post-independence years of the 1960s was thus to quickly build a public school system out of the largely disparate, generally community- and faith-based schools that existed at that time, driven largely by the mission to rapidly raise the basic literacy and numeracy of the people. That early phase of rapid expansion, termed the survival-driven phase by education historians (Goh and Gopinathan 2008), was characterized by high levels of central control by MOE, for efficiency reasons.

Over time, that early governance structure has evolved to one characterized by a close nexus between policymakers and practitioners, deliberately designed to achieve a balance between the centralization and decentralization of different aspects of the education system. Today, MOE is still responsible for setting national policies that affect access to education for all children (e.g., curriculum, funding rates, and school fees), for reasons of both efficiency and equity. But it also devolves significant autonomy and responsibility to school principals and teachers in administration and professional matters pertaining directly to their own schools (e.g., budget allocations in the school, setting of school policies, customization of the national curriculum, and pedagogical approaches for students with different learning needs). Such autonomy

for local customization is a key feature that now allows the Singapore education system to be nimble and responsive to student needs.

However, a key feature that remains even as the system evolves in governance structure is how MOE has always adopted an evidence-based approach in designing, developing, and reviewing the curriculum, programs, and policies since the early post-independence years in the 1970s. For example, what prompted the development of the Singapore model method for learning mathematics were the worrying results from a study conducted by MOE in 1975, showing that at least a quarter of the primary school graduates could not meet the minimum numeracy levels expected at the end of primary school (MOE 2009). This, set against the broader context of double-digit dropout rates even at primary-school level (29%; Goh et al. 1979), suggested strongly to policymakers then that something was clearly not right, either with the curricular design or the way in which it was implemented, or both. MOE introduced the model method to address the concerns raised by the study. Subsequently, after years of experimentation and refinements based on feedback from actual use in local classrooms, the model method has become a pedagogical approach integral to Singapore's primary school mathematics curriculum.

In fact, in those early post-independence years when Singapore was rapidly building and expanding the school system, MOE was particularly keen to learn best practices from all over the world. Participating in international studies in those early years (e.g., SISS and IEA's 1991 Reading Literacy Study) contributed valuably toward that purpose.

Today, the impetus to look outwards is even stronger, catalyzed by an increasingly interdependent and networked world. In this world, education systems no longer have the luxury to be insular or inward-looking, especially in terms of the skills and knowledge that they help students develop. This is particularly so for Singapore because of the open economy and society. Large-scale international studies therefore continue to play an important role in complementing other data sources to inform MOE about different aspects of the education system, serving at various times as reassurances that the system is progressing in the right direction in some areas and, on other occasions, providing insights into where improvements can be made. This is necessary to ensure that the education system is always forward-looking and responsive to dynamic changes in the external environment, in order to adequately prepare Singapore's students to thrive in an equally, if not more, dynamic future.

Singapore currently participates in five large-scale international studies: IEA's PIRLS and TIMSS, and three programs run by the Organisation for Economic Co-operation and Development (OECD), namely the Programme for International Student Assessment (PISA; OECD 2020a), Teaching and Learning International Study (TALIS; OECD 2020b), and Programme for International Assessment of Adult Competencies (PIAAC; OECD 2020c). MOE chose these carefully and purposefully, among the options available internationally, to form a suite of studies that would best address its knowledge needs. Together, PIRLS, TIMSS, and PISA allow policymakers to monitor and derive insights about the education system at different milestone grades (primary, lower-secondary, and upper-secondary) for various student developmental outcomes, and educator characteristics and practices.

TALIS, targeting only teachers and principals, provides in-depth insights into teacher characteristics, policies, and practices. PIAAC, involving adults aged 16–65 years old, provides information about longer-term continual skills development, employment, and other life outcomes beyond the formal education years. This last study is thus a rich source of information on both the progress of lifelong learning and how well the formal education system has prepared its students for life.

More broadly, these international studies share several features that make them useful to Singapore, as we now explain in more detail.

14.2.1 Participating in International Large-Scale Assessment Facilitates Benchmarking of Student Developmental Outcomes and Educator Practices

An important feature of large-scale international studies lies in the participation of many education systems, which is crucial to allow benchmarking and cross-national comparisons. On the student front, the international studies show how well students of different grade levels/ages are developing skills, competencies, and attitudes towards learning that are considered essential for thriving in the 21st century by the international education community, relative to their international peers. For example, PISA results show that Singapore's 15-year-old students not only do well in the "traditional" domains (reading, mathematics, and science), they are also competent collaborative problem solvers by international standards. Similarly, ePIRLS provides findings on how well the grade 4 students can navigate in an e-environment to select relevant information and integrate ideas across webpages. Such benchmarking is important because the students do not live in isolation, but have to compete in the global marketplace of skills in the future.

Similarly, apart from student data, the studies provide internationally comparable information about various policies and practices pertaining to educators (e.g., teacher and principalship preparations and teaching strategies), allowing MOE to understand the strengths and weaknesses of local practices beyond what can be gleaned from local data. For example, MOE already knew from local data that Singapore teachers worked long hours and spent a lot of time on out-of-class activities, such as running co-curricular programmes for students. But it was results from TALIS which showed that Singapore teachers put in some of the longest working hours among TALIS participants and that they spent proportionately more of those hours on "marking" and "administrative work" than their peers elsewhere.

The cyclical nature of the studies further means that benchmarking can be done not just at single time-points, but also over time, using scales deliberately designed for such trend analyses. This is particularly useful when there are changes to policies and curriculum over time. For example, trend data from these studies has enabled MOE to monitor changes in student outcomes and attitudes over time. It has also provided policymakers with some evidence of the impact of programs and policies, which

cannot be easily measured using local administrative or national examination data. For example, results from PIRLS, TIMSS, and PISA consistently assured MOE that cuts in curriculum content and the corresponding shifts towards emphasizing higher-order thinking skills since 1997 had not impacted student performance negatively, but were instead associated with increased levels of application and reasoning skills. Similarly, analyses using TIMSS data from multiple cycles before and after the launch of the 2008 science syllabi centered on the inquiry approach showed that there were more inquiry-based practices in grades 4 and 8 science classrooms after the implementation.

14.2.2 Participating in International Large-Scale Assessment Provides Additional High-Quality Rich Data Sources for Secondary Analyses

Besides benchmarking of outcomes and practices, another important feature of large-scale international studies is the availability of rich contextual data about each education system. This is useful for conducting secondary analyses to derive further insights about education systems, with implications for policymaking and practice.

In particular, the rich information about students' learning contexts in the classroom, school and home from these studies allows analysts to better understand the influence of these contexts on student outcomes. For example, using PIRLS and TIMSS data, it was found that grade 4 students whose parents engaged them more frequently in early literacy and numeracy activities during the preschool years did better in reading, mathematics, and science at grade 4, even after accounting for home socioeconomic circumstances. This means that, where parents are unable to provide the support, early intervention in preschool education and care centers is important, thereby supporting the government's investments in the sector.

In this regard, IEA's PIRLS and TIMSS share two distinguishing features not found in the OECD studies, which enhance their analytic value. First, the direct links of these two studies to the curriculum, across all three aspects of intended, implemented, and attained curricula underpinning the research frameworks of TIMSS and PIRLS, allow insights to be drawn for informing curricular review work. For example, information on the intended curriculum across different countries provides insights into both the broad common areas of curricular focus and the differences in emphasis between countries. These are useful as part of the external scans for regular syllabus reviews in Singapore. Similarly, information on the implemented curriculum (e.g., teaching practices) and the attained curriculum (e.g., student achievement scores and student attitudes towards learning) enables Singapore to monitor the enactment and impact of the curriculum, especially useful during curricular reviews.

Second, the direct links between students and their reading, mathematics, and science teachers in PIRLS and TIMSS open up additional analytic possibilities for discovering important relationships between teacher practice and student outcomes.

Although such estimated relationships are non-causal in nature because of the study design, they nonetheless allow some inferences to be made about hypothesized relationships (e.g., teacher inquiry-based practice and student outcomes), serving as a first step to further, more targeted studies aimed at uncovering any causal relationships where appropriate.

14.2.3 Participating in International Large-Scale Assessment Builds International Networks of Educationists and Experts

Over time, each of the international studies has built an entire ecosystem of parties interested, and in many cases, actively involved, in the work of educational improvements. These international communities comprise individuals with a diverse range of experiences ranging from research to policymaking to practice, but all driven by the common goal of providing quality education for the students. Participating in each study thus opens up opportunities to be plugged into international conversations about education with thought leaders and educationists from different parts of the world, exchanging views and learning from one another while working to improve our respective systems.

Another very useful, albeit incidental, benefit that MOE derives from participating in the international studies is that the staff learn and grow professionally in the specialized areas (e.g., sampling, design of computer-based assessment items, and measurement and analytical methods) from being directly involved in the projects. Over time, they have also built networks with the various experts, who can be readily tapped for advice in areas beyond work directly related to the studies.

14.3 How MOE Has Used Large-Scale Assessment Data

In this section, we illustrate how MOE has capitalized on valuable PIRLS and TIMSS data to inform both policymaking and program improvement, using three practical examples. We have chosen these cases to illustrate the range of typical uses, from serving as an external signal of areas for improvement, to monitoring the implementation of changes via a specific policy or program, to system-level monitoring of broader changes to policy or practice. The examples we have selected are all related to the curriculum because, as mentioned earlier, one distinguishing feature of PIRLS and TIMSS, which makes the data particularly valuable to Singapore, lies in their links to the curriculum.

14.3.1 STELLAR: "We Must and Can Do Better!"

The STrategies for English Language Learning and Reading program, better known as STELLAR, is the primary level English language instructional program, specially developed to cater to the learning needs of Singapore's children, taking into consideration the nation's multilingual environment. While English is the common language for government, business, and education in Singapore, Singaporean students do not speak only English at home. STELLAR was therefore created from a deliberate combination of first- and second-language teaching approaches. It was the first time that MOE articulated and operationalized a core set of pedagogies to guide the learning of the English language across six years of education (grades 1 to 6). In addition to applying principles gleaned from the research literature, the curriculum team behind STELLAR conducted a systematic review of English language teaching in Singapore, consulting teachers, observing classes of students, and speaking to stakeholders in the community, including employers. The team also conducted three study trips to learn from educators in Hong Kong, India, and New Zealand. During its implementation, the STELLAR team worked closely with English language teachers, influencing and changing classroom practices over time. Today, nearly 15 years since its launch, STELLAR remains the signature programme that builds a strong foundation for students in English language, not just as the common language for communication in multilingual Singapore, but also as a language for accessing further learning.

But what is perhaps less known about STELLAR is that Singapore's results from IEA's 10-year trend study of reading literacy, using data from the 1991 Reading Literacy Study and PIRLS 2001 (Martin et al. 2003), played an important role in galvanizing support for a fundamental redesign of the primary English language program that eventually became STELLAR. Specifically, on average, Singapore's grade 4 students had not made much progress in English language reading proficiency over the intervening decade, unlike four of the nine countries involved in the 10-year trend study, which had shown improvements. Moreover, Singapore was the only country with a widened gap between the highest and lowest performers. Further analyses showed that, while there was some progress made at the upper end of the achievement spectrum, with the 75th and 95th percentile scores being higher in 2001 than in 1991, students at the lower end of the achievement spectrum had largely stayed at the 1991 proficiency levels. These findings, especially the "long lower tail" phenomenon, led to a wide-ranging review of the curriculum, efforts that subsequently culminated in the launch of STELLAR in 2006, with a phased-in implementation approach that reached all schools at grade 1 in 2010. At the same time, for students at grades 1 and 2 who were assessed to need more support in beginning reading skills, an enhanced learning support program was implemented from 2007.

The use of PIRLS did not end with the launch of STELLAR. Instead, analyses using an external, stable benchmark, such as PIRLS, served as a useful complement to other evaluations that were conducted using local data. For example, data from

the four cycles of PIRLS (i.e., 2001, 2006, 2011, and 2016) showed that Singapore's grade 4 students had made steady progress over the 15 years from 2001 to 2016, with a growing percentage having acquired higher order reading skills. Importantly, there was a reduction in the proportion of students who could not meet the "low" benchmark in PIRLS, from 10% in 2001 to 3% in 2016. A six-year (from 2007 to 2012), quasi-experimental, longitudinal study in 20 schools conducted by MOE also found that students in the STELLAR program performed significantly better than a control group on a number of language and reading skills as they progressed through each grade level and at the end of grade 6 (Pang et al. 2015).

Aside from reading achievement, PIRLS data also showed how reading habits changed over time. For example, data indicated that the proportion of Singaporean grade 4 students reading silently on their own in school every day or almost every day had increased from 56% in 2011 to 62% in 2016. This was set against a decline in reading habits outside of school, with the proportion of students reading outside of school for at least 30 min on a school day falling from 68% in 2011 to 56% in 2016. This decline in reading habits outside of school was also observed in more than two-thirds of the education systems with trend data that participated in PIRLS 2016. For the STELLAR curriculum team, the increase in silent reading in school, set against a declining reading culture outside of school, affirmed the importance of school reading programs in promoting extensive reading.

The design of PIRLS, with its stronger links to the curriculum and its emphasis on robust trend data, has enabled the Singapore curriculum development team to monitor the impact of its programs over a long-term trajectory, contributing to efforts to improve the teaching and learning of English language in Singapore.

14.3.2 A New Pedagogical Approach to Learning Science: "We Tried a Different Method, Did It Materialize?"

From the mid-2000s, MOE made concerted efforts to move towards an inquiry-based approach to the learning of science at both primary and secondary grade levels. Specifically, the 2004 Science Curriculum Framework adopts inquiry as the central focus, both to (1) enable students to appreciate the relevance of science to life, society, and the environment; and (2) equip them with the knowledge, skills, and dispositions to engage in science meaningfully in and out of school (Poon 2014). This framework was operationalized through the 2008 primary and lower-secondary science curricula.

However, having the curricular documents is just a first step. To bring about real change in the classrooms, there must be deliberate and sustained efforts to help teachers not just to understand, but more importantly, adopt the new pedagogical approach. Towards this end, MOE mounted a series of professional development activities, in association with both pre-service and in-service training providers,

specifically targeted at helping science teachers develop the necessary skills to implement the inquiry-based approach in their classrooms in the initial years of implementation. This challenge of ensuring the alignment of every science teacher's training to the intent of inquiry science was surmountable because of the close partnership between MOE and the National Institute of Education, Singapore's sole teacher training institute.

Beyond professional development, regular school visits by MOE curriculum staff provided further support to the teachers during the actual implementation process, working hand-in-hand with teachers on designing lessons, observing their lessons, and providing feedback to further refine the lesson plans and practice. Such hands-on support, while labor-intensive, is critical; research has shown that understanding teachers' actual experiences when trying to adopt a new practice in their classrooms (especially their *in situ* challenges) and then providing appropriate support help teachers to successfully adopt an inquiry-based practice in science (Poon and Lim 2014).

Given the amount of effort required to effect a system-wide pedagogical change, one of the policy and practice questions of interest to MOE after the initial years of implementation was how much progress had been made in this systemic shift towards an inquiry-based approach to learning science. In particular, to what extent (if at all) had the espoused shift been translated into actual practice in science classrooms? In addition, to what extent (if at all) was teacher inquiry-based practice associated with different teacher characteristics?

Besides serving as an important support structure for teachers in the implementation process, the regular school visits by MOE curriculum staff also provided the curriculum designers in MOE with first-hand information about the enactment of the new approach through classroom observations and focus group discussions with heads of science departments and their teachers. Some local research studies also reported pockets of teacher use of hands-on and more open-ended scientific investigations aimed at encouraging students to move away from merely following instructions to more self-directed learning and creative thinking (e.g., Chin 2013; Poon et al. 2012). All these provided some answers to the MOE's questions, which were used for further fine-tuning of the curriculum and implementation process.

TIMSS supplemented these local sources of information, most of which were qualitative and small-scale rather than system-wide in nature, by serving as a large-scale, quantitative data source that enabled trend analyses because of its cyclical nature. Specifically, by analyzing data from TIMSS straddling the implementation of the 2008 science curricula (i.e., data from TIMSS 2007 and TIMSS 2011), MOE examined the extent to which the system-wide espoused shift towards inquiry-based approach had translated into teachers' actual practice in grade 4 and grade 8 science classrooms.

To do so, specialists from the TIMSS national research center at MOE first created an "inquiry approach" scale to measure the extent of inquiry-based practice in classrooms, as reported by teachers. For direct use in answering questions about the new science-inquiry approach implemented in local schools, this inquiry approach scale had to be aligned to MOE's definition of science as inquiry "as the activities and

processes which scientists and students engage in to study the natural and physical world around us…consisting of two critical aspects: the *what* (content) and the *how* (process*)* of understanding the world we live in" (MOE 2007, p. 11, emphasis added). The availability of item-level data in TIMSS made it possible for Singapore to create a fit-for-purpose scale, using selected items in the TIMSS teacher questionnaire assessed by an expert science curriculum team to be good proxies of inquiry-based practice based on MOE's conception.

Using the inquiry approach scale, the research team compared the average inquiry approach scores in TIMSS 2007 and 2011 to detect any shift in practice. Using other variables and scales that were available in the TIMSS datasets, the team also examined the association between teachers' inquiry-based practice with different teacher characteristics (e.g., years of experience, levels of preparation, and confidence in teaching science). The team found that, on average, proportionately more grade 4 and grade 8 students had science teachers who reported the use of inquiry-based pedagogies in TIMSS 2011 than in TIMSS 2007. In terms of relationships with teacher characteristics, the team found positive relationships between teacher inquiry-based practice and their use of instructional strategies that engaged students in learning. Similarly, it found a positive relationship between inquiry-based practice and teachers' levels of confidence in teaching science, potentially mediated by their levels of preparedness. These findings are useful because they suggest some potential levers that can be used to influence teachers' adoption of inquiry-based practice.

The analysis was subsequently replicated using TIMSS 2015 data when it was available, as part of MOE's continual efforts to monitor the situation in classrooms. More broadly, findings from such analyses using TIMSS also form part of the knowledge base that MOE has built over time about inquiry-based learning, not just in science but more broadly across different subjects, tapped by various parties involved in regular curricular reviews.

14.3.3 Bold Curricular and Pedagogical Shifts: "We Made Some Trade-Offs, What Did We Sacrifice?"

Policymaking is frequently about weighing trade-offs and choosing the best policy option available to maximize the positive impact and minimize the negative, based on the information that is available at the point when a decision has to be made. Subsequently, it is necessary to closely monitor the actual implementation of the policy option chosen to ensure that the predicted positive impact comes to fruition while the expected negative ones are minimized and mitigated where possible. Trustworthy data from various sources, both quantitative and qualitative in nature, are important to this ongoing post-implementation monitoring process. In our last example, we illustrate how MOE has used PIRLS and TIMSS data for this purpose.

Since the late 1990s, MOE has been progressively and systemically reducing the content of individual subjects across the grade levels to create space for greater

emphasis on other learning outcomes important to students' holistic development, including higher-order skills such as creativity, application, and reasoning skills (Gopinathan 2002; Poon et al. 2017). MOE embarked on this system-wide shift, which continues today, in part because of the recognition that, while a strong understanding of concepts in each subject domain remains important, being able to apply such understanding and knowledge to real-life situations, including novel ones, is increasingly important in a world where humans cannot out-compete machines in storing and retrieving voluminous amounts of information. As such, there is a need to ensure that students have enough opportunities to practice and develop these skills during their schooling years.

Effecting the shift requires deliberate and fundamental reviews of the curriculum of each subject across the grade levels, largely through a two-pronged approach. First, judicious cuts to the content materials (up to 30% of the original curriculum in some instances) have to be made to carve out adequate time and space from the precious curriculum hours each week to devote to developing skills beyond content learning. Second, teachers need to make pedagogical shifts in tandem in order to capitalize on the time and space available and design learning experiences that will foster higher-order thinking skills among their students. The move towards adopting an inquiry-based approach in learning science (described in Sect. 14.3.2) is an example of how that broader shift is manifested pedagogically in the specific subject of science.

A key policy question of interest to MOE is whether all the deliberate curricular cuts and pedagogical shifts at the system level have made a positive impact on students' learning and development. Most importantly, are students short-changed in any way by the bold move of trying to teach them less (by way of content) so that they can learn more (by way of other increasingly important skills)? PIRLS and TIMSS data are useful external, stable benchmarks that provide some answers for the reading, mathematics, and science curricula, in ways that cannot be answered using MOE's local examination data because these examinations change in tandem with changes to the curricula.

Results from TIMSS 2015 and PIRLS 2016 provide some assurance that the system-wide shifts are progressing in the right direction. The students continue to show a strong mastery of mathematics and science at grade 4 and grade 8, and of reading literacy at grade 4, performing well by international standards. More encouragingly, they have made steady improvements over the years, especially in terms of the higher-order thinking skills. In particular, Singapore students at both grades have demonstrated progress in their ability to apply and reason in both mathematics and science, as measured by two of the three cognitive domain scores in TIMSS, namely, "Applying" and "Reasoning." This is notwithstanding the decline in "Knowing" scores between 2007 and 2015 observed in science for grade 4 students, showing there are some trade-offs at play. Similarly, results from PIRLS and ePIRLS 2016 show that Singapore's grade 4 students are able to interpret and integrate ideas and information well, and evaluate textual elements and content to recognize how they exemplify the writer's point of view. ePIRLS 2016 also provided the opportunity to assess students' online reading skills for the first time on an internationally comparable scale. The results suggest that Singapore's grade 4 students do well not

only because they are able to transfer their reading comprehension skills in print to online reading but also because they are able to navigate non-linearly between different websites. These are important skills in an increasingly digitalized world, where information and knowledge reside across multiple online platforms.

These findings provide MOE with not only the confidence to continue with the system-wide curricular and pedagogical reforms but also concrete evidence to convince the different stakeholders that these reforms are on the right track, thereby garnering continued support for further reforms in the same direction.

14.4 Some Principles Underpinning MOE's Use of Large-Scale Assessment Data

Reflecting on Singapore's purpose for participating in international large-scale assessments (ILSAs) and how MOE has used the data from these studies, there are three key general principles guiding such use.

First, insights from ILSAs serve as only one of the sources, rather than the sole source, of information that feeds into the deliberation of policies and program design and development. This is because, despite the richness and robustness of the data that such international assessments provide, each study has its inherent limitations and can only support inferences within those technical limits. For example, due to the sampling design of PIRLS and TIMSS, the teacher respondents to the teacher questionnaires are not necessarily a representative sample of the teacher population; they are instead involved in the studies only because they taught the sampled students. As such, the data alone cannot support inferences that are directly generalizable to the teacher population. This means any deliberations that require such direct inferences about the teacher population have to involve other data sources, minimally as a form of triangulation for findings using data from the PIRLS and TIMSS teacher questionnaires. More importantly, ILSAs cannot be the sole source of information, because education policies and practices operate within a complex system for which no single data source, even when collected through a mixed methods design, will be adequate in painting a full picture (Jacobson et al. 2019). As such, insights from multiple sources are necessary to provide sufficient information at the point when a decision has to be made.

Second, and related, although the international assessments provide useful insights that can be used to inform curricular reforms (especially PIRLS and TIMSS, which have direct links to the curriculum), MOE does not intentionally align the Singapore curriculum to the objectives in the assessment frameworks of the different international studies. Instead, the curriculum designers and developers are guided by what they deem important for students to learn during their schooling years, and not all of these curricular objectives will necessarily fit into what is agreed upon for the assessments by the international community.

Finally, in order to ensure that data from ILSAs remain useful, MOE takes deliberate steps to ensure that the data are not at risk of being corrupted, which would render such material useless as an objective source of information. An important aspect of this is to ensure that the scores are not used for any accountability purposes with any stakes for individuals in the system, which may evoke undesirable behavioral responses leading to potential distortion of the assessment scores (e.g., Figlio 2006; Hamilton et al. 2007; Koretz 2017; Koretz and Hamilton 2006). MOE deliberately does not use results from the international studies to reward or sanction individual schools, teachers, students, or owners of specific policies/programs. Instead, the results are used only for system-level monitoring. Even when secondary analyses are done, the insights gained are used only for system-level decisions, in line with the inferences that data from the studies can properly support. More fundamentally, having a separate research arm comprising staff with the technical expertise to derive useable insights from the international studies (situated within MOE, but separate from policy and program owners) to oversee all aspects of each international study, including all secondary analyses needed, also helps to ensure independent and responsible use of the datasets from the international studies.

Adhering to these guiding principles, now and in the future, enables MOE to derive the greatest value from the ILSA data. International studies remain an important and valuable information source, complementing other sources of information that MOE regularly collects about different aspects of the education system, and contribute to the deliberations of policies and programs in MOE's ongoing efforts to improve students' schooling experiences.

References

Chin, T. Y. (2013). *Insights and lessons from teachers' initial professional learning and collaborative practices in questioning for higher-order scientific discourse in primary science classrooms.* Doctor in Education thesis, Nanyang Technological University, Singapore.

Figlio, D. (2006). Testing, crime and punishment. *Journal of Public Economics, 90*(4–5), 837–851. https://www.sciencedirect.com/science/article/abs/pii/S0047272705000708.

Goh, K. S. & Education Study Team. (1979). *Report on the Ministry of Education 1978*. Singapore: Ministry of Education. https://eservice.nlb.gov.sg/item_holding.aspx?bid=4082172.

Goh, C. B., & Gopinathan, S. (2008). The development of education in Singapore since 1965. In S. K. Lee., C. B. Goh, B. Fredriksen, & J. P. Tan (Eds.), *Toward a better future: Education and training for economic development in Singapore since 1965* (pp. 12–38). Washington, DC: The World Bank. http://siteresources.worldbank.org/INTAFRREGTOPEDUCATION/Resources/444659-1204656846740/4734984-1212686310562/Toward_a_better_future_Singapore.pdf.

Gopinathan, S. (2002). Remaking the Singapore curriculum: Trends, issues and prospects. In Hong Kong Institute of Education, *School-based curriculum renewal for the knowledge society: Developing capacity for new times: Proceedings of the 1st Conference of Asia Pacific Curriculum Policy Makers* (pp. 71–81). Hong Kong: Hong Kong Institute of Education.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., & Russell, J., et al. (2007). *How educators in three states are responding to standards-based accountability under No Child Left Behind*. Research Brief. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_briefs/RB9259.html.

Jacobson, M. J., Levin, J. A., & Kapur, M. (2019). Education as a complex system: Conceptual and methodological implications. *Educational Researcher, 48*(2), 112–119. https://journals.sagepub.com/doi/10.3102/0013189X19826958.

Koretz, D. (2017). *The testing charade: Pretending to make schools better*. Chicago, IL: University of Chicago Press.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: Praeger Publishers.

Martin, M. O., Mullis, I. V. S., Gonzalez, G. J., & Kennedy, A. M. (2003). *PIRLS: Trends in children's reading literacy achievement 1991–2001*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://www.iea.nl/publications/study-reports/international-reports-iea-studies/trends-childrens-reading-literacy.

MOE. (2007). *Science Syllabus Primary 2008*. Retrieved on 8 April 2015 from https://www.moe.gov.sg/docs/default-source/document/education/syllabuses/sciences/files/science-primary-2008.pdf.

MOE. (2009). *The Singapore model method for learning mathematics*. Singapore: EPB Pan Pacific.

OECD. (2020a). PISA. Programme for International Student Assessment [webpage]. Paris, France: OECD. https://www.oecd.org/pisa/.

OECD. (2020b). TALIS. The OECD Teaching and Learning International Survey [webpage]. Paris, France: OECD. http://www.oecd.org/education/talis/.

OECD. (2020c). Survey of Adult Skills (PIAAC) [webpage]. Paris, France: OECD. https://www.oecd.org/skills/piaac/.

Pang, E. S., Lim, F. V., Choe, K. C., Peters, C., & Chua, L. C. (2015). System scaling in Singapore: The Stellar Story. In C.-K. Looi & L. W. Teh (Eds.), *Scaling Educational Innovations* (pp. 105–122). Singapore: Springer.

Poon, C-L. (2014). Five decades of science education in Singapore. In A-L. Tan, C-L. Poon, & S. S. L. Lim (Eds.), *Inquiry into the Singapore science classroom: Research and practices* (pp. 1–26). Singapore: Springer.

Poon, C-L., & Lim, S. S. L. (2014). Transiting into inquiry science practice: Tales from a primary school. In A-L. Tan, C-L. Poon, & S. S.L. Lim (Eds.), *Inquiry into the Singapore science classroom: Research and practices* (pp. 139–164). Singapore: Springer.

Poon, C. L., Lam, K. W. L., Chan, M., Chng, M., Kwek, D., & Tan, S. (2017). Preparing students for the twenty-first century: A snapshot of Singapore's approach. In S. Choo, D. Sawch, A. Villanueva, & R. Vinz (Eds.), *Educating for the 21st century: Perspectives, policies and practices from around the world* (pp. 225–241). Singapore: Springer.

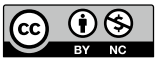
Poon, C. L., Lee, Y. J., Tan, A. L., & Lim, S. S. L. (2012). Knowing inquiry as practice and theory: Developing a pedagogical framework with elementary school teachers. *Research in Science Education, 42,* 303–327.

Hui Leng Ng is the Director of the Research and Evaluation Branch at the Ministry of Education, Singapore. She is also a Principal Research Specialist in International Benchmarking and Research. She received her B.Sc. and M.Sc. in Mathematics from Imperial College of Science, Technology and Medicine, MEd in General Education from the National Institute of Education, Nanyang Technological University, and EdD in Quantitative Policy Analysis in Education from Harvard University, specializing in quantitative research methods. She currently oversees Singapore's efforts in various large-scale international studies, and was Singapore's National Research Coordinator for TIMSS.

Chew Leng Poon is the Divisional Director of the Research and Management Information Division at the Ministry of Education, Singapore. She is also a Principal Specialist in Research and Curriculum. She earned her Master degree in Curricular Studies at the Ohio State University, US and her Ph.D. from the National Institute of Education, Nanyang Technological University, specializing in science inquiry pedagogy. She currently drives the Ministry's efforts in

research and evaluation to inform policymaking, program development and service delivery, and is Singapore's representative at the IEA General Assembly.

Elizabeth Pang is Principal Specialist, English Language, at the English Language and Literature Branch of the Curriculum Planning and Development Division, Ministry of Education, Singapore. She received her B.A. in English from the University of Oxford, and her M.A. and Ph.D. in Educational Linguistics from Stanford University. She currently oversees curricular and pedagogic development in language learning and literacy instruction at the Ministry of Education. She is a member of the PIRLS Reading Development Group and is Singapore's co-National Research Coordinator for PIRLS.



Chapter 15
Understanding the Policy Influence of International Large-Scale Assessments in Education

David Rutkowski, Greg Thompson, and Leslie Rutkowski

Abstract International large-scale assessments (ILSAs) require national governments to invest significant resources in both time and money. With such investment national governments expect the results of ILSAs to provide policy and research communities with evaluative information on their educational system. Armed with this evaluative information, policymakers in many participating countries have used the results to stimulate reform. However, systematically tracking this influence and judging the validity of the claims has proven difficult for both the policy and research communities. There can be an erroneous expectation in the wider education community, and sometimes amongst policymakers themselves, that ILSA data automatically suggests policy solutions. Because of this error, a means for better systematizing the policymaking process responding to ILSA results is required. The model developed here can assist the policy and research community to better understand whether ILSAs are providing valid evidence to support their use in policy formation enactment and can be used to analyze ongoing consequences of that influence. Two worked examples demonstrate the utility of the model.

Keywords Evaluation · Large-scale assessment · Policy impact · Policy influence

D. Rutkowski (✉) · L. Rutkowski
Indiana University, Bloomington, IN, USA
e-mail: drutkows@iu.edu

L. Rutkowski
e-mail: lrutkows@iu.edu

G. Thompson
Queensland University of Technology, Brisbane, Australia
e-mail: g6.thompson@qut.edu.au

H. Wagemaker (ed.), *Reliability and Validity of International Large-Scale Assessment*, IEA Research for Education 10, https://doi.org/10.1007/978-3-030-53081-5_15

15.1 Introduction

Countries and educational systems participate in international large-scale assessments (ILSAs) for a variety of reasons, including educational system monitoring and comparison. As taking part in an ILSA requires the investment of significant money and time (Engel and Rutkowski 2018), it is important that countries derive value and use from their participation. One way to justify participation is to demonstrate the ways in which ILSA results are used (or are claimed to be used) as a lever for policy formulation and as a means to change policy trajectories in order to improve educational outcomes. That is, successfully attributing policy changes to ILSA results can be seen as a rationale for continued participation, building a case for the further outlay of time and money. It makes sense that testing organizations want to argue to their stakeholders that the resources spent on their assessment tools are worthwhile. However, demonstrating how ILSAs impact systems and nations remains difficult. First, this is because policymaking itself is an emotive and politicized domain informed more by what can be sold to the public for electoral success than what research might suggest could be the best direction (see for example Barber et al.'s (2010) concept of "Deliverology"). Second, even if an association is identified, it remains challenging to establish the direction of the relationship or the amount of influence ILSAs had in any policy change that resulted. In other words, it is difficult to prove the counterfactual that the policy change would not have occurred in the absence of the ILSA. Third, evidence of the influence of ILSAs on policy can be inflated or misleading. For example, there are a number of cases demonstrating that governments made use of ILSA results simply to justify policy reforms that were already set to be implemented (Gür et al. 2012; Rautalin and Alasuutari 2009; Takayama 2008).

In this chapter, we explore how participation in ILSAs, and the subsequent results, could reasonably be said to "influence" policy. In other words, how can ILSA results, or any proposed policy attributed to those results, be shown to be the reason that a policy or policies change? In complex systems the attribution of singular causes that can explain an altered state of affairs is always difficult because of the multiple forces at work in that system. Further, few would argue that the ILSA-policy nexus is easy to understand given: (1) the complex social, cultural, historical, economic and political realities within each system; (2) the complexities between systems; and (3) the limitations of what the tests themselves can measure on any given topic and the difficulty in measuring policy change. This, then, leads to a key problem that confronts policymaking communities; how can system leaders properly understand and manage ILSAs' influence on their system?

A second question that drives this chapter asks what are the overall consequences of participating in an ILSA? This question works from the premise that there are always intended and unintended consequences when an ILSA has influence at the national level. When ILSA results are used to set policy goals or are the impetus for educational change this creates the conditions for a range of consequences. For example, implementing a particular kind of science curriculum as the result

of middling science performance on an ILSA will have consequences that might include money spent training teachers and abandoning other teaching approaches, and so on. Correspondingly, where a system's leaders become convinced that doing well on rankings will lead to better educational outcomes, a variety of perverse incentives can emerge, resulting in attempts by a range of stakeholders to "game" the test. This is evidenced through the multitude of high stakes testing cheating that took place in the USA (Amrein-Beardsley et al. 2010; Nichols and Berliner 2007) and the fact that some countries participating in ILSAs are removed from the results for "data irregularities," including being too lenient when marking open-ended questions, which resulted in higher than expected scores (OECD [Organisation for Economic Co-operation and Development] 2017). Stakes at the student and school level may remain low; however, at the national level there is growing evidence that the stakes of participation are high.

One challenge in the ILSA-policy nexus that encompasses both understanding ILSA influence and the consequences of that influence is that there is rarely, if ever, systematic analysis undertaken of the data in the context of the whole system. When ILSA data is released media, policymakers, and other stakeholders tend to sensationalize and react, often quickly, without a full accounting of the evidence (Sellar and Lingard 2013). To underline this problem, we present two cases that highlight the problem of simply claiming ILSA "influence." Subsequently, we describe a model as a means for better systematizing how influence ought to be attributed to policy processes as a result of participation in ILSAs and the publication of subsequent results. This model that can assist the policy and research communities to better understand whether ILSAs are providing valid evidence to warrant their influence on educational policy formation and debates and to analyze the consequences of that influence.

15.2 Impact, Influence, and Education Policy

To understand influence, we first differentiate between what we view as ILSA's impact on policy (which is hard to demonstrate) and ILSA's influence on policy (a concept that is still difficult but easier to demonstrate than impact). For the purposes of this chapter, we define impact as a *difference in kind* while influence is defined as a *difference in degree*. To show policy impact, we would have to isolate an ILSA result and proved that this caused a significant shift in a policy platform. We should expect to see clear evidence that there was a rupture, such as a new national policy direction being caused by ILSA data. However, making causal claims such as ILSA *X* caused Policy *Y* requires a methodological framework that may simply not be possible because of the complexity of most national systems. Moreover, many of the claims made in reports that attribute impact to an ILSA result exemplify what Loughland and Thompson (2016) saw as a post hoc fallacy at work rather than identifying a causal mechanism. They explained that, "when an assumption is made based on the sequence of events—so the assumption is made that because one thing occurs after

another, it must be caused by it" (Loughland and Thompson 2016, pp. 125–126). This is particularly true of ILSA results where the data often appear to be used to maintain current policy directions in the interests of political expediency, even where the data suggests this may be having unhelpful consequences. Finally, impact is notoriously difficult to demonstrate because education policy agendas are often politically rather than rationally decided (Rizvi and Lingard 2009). As such, even if ILSAs provided perfect information they will only be one factor among many that influence policymaking. For these reasons, when ILSAs are mentioned together with policy impact, we suspect that it would be better to frame this in terms of evaluating the influence that ILSAs have on policy agendas and trajectories within given contexts.

Policy influence can be viewed as the use of ILSA results to buttress or tweak policy settings that already exist. However, establishing exactly what constitutes influence remains difficult for a number of reasons. For example, similar to impact, much of the policy literature fails to define influence (Betsill and Corell 2008). The lack of a clear definition in the literature leads to (at least) three problems. First, without an established definition for influence, it is difficult to determine the type of evidence needed to demonstrate influence. This is a particular problem, as ILSA literature tends to report evidence of influence on the basis of the particular case at hand without consideration of wider application and with a pro-influence bias, rarely examining evidence to the contrary (e.g., Breakspear 2012; Schwippert and Lenkeit 2012). Second, to make a strong case that ILSAs influence policy, some consensus as to what data should be collected to mount a sufficient argument is needed. Finally, this lack of definition makes cross-case comparisons potentially unstable because different stakeholders risk measuring different things and claiming them as demonstrating influence. In other words, if each claimant develops their own ideas around influence and collects data accordingly, they may simply end up comparing different things.

In this chapter we borrow from Cox and Jacobson (1973), who defined influence as the "modification of one actor's behavior by that of another" (p. 3). With this definition we take a broad view and define ILSAs as policy actors that are as involved in creating meaning in a variety of contexts as much as they are created artefacts of organizations or groups of countries. As an actor, an ILSA represents multiple interests and ambitions, and intervenes in social spaces in a variety of ways. For example, in the case of OECD's Programme for International Student Assessment (PISA) study, the results are intended to serve at the behest of the OECD's and member countries' policy agendas (OECD 2018). However, the declared explicit use of ILSAs for policy modifications are less clear for the International Association for the Evaluation of Educational Achievement's (IEA) Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS), perhaps because of the IEA's history as a research rather than policy focused organization (Purves 1987). That said, ILSAs are evaluative tools partly sold to "clients" such as nation states with an assumption that the assessment will help judge the merit and worth of a system by measuring performance. Demonstrating positive impact and/or influence to those jurisdictions who have paid to have the

tests administered would make commercial sense regardless of the methodological concerns outlined above. Rightly so, testing organizations and countries alike want to know whether the tests they design and administer are having a positive influence on systems, at least partly because organizations then have a compelling narrative to sell to other potential participants and countries have a legitimate reason for participating.

We do not, however, want to ground our discussion of ILSA influence on policy in a naive caricature of ILSAs as authoritarian tools thrust onto nations by some evil council of neoliberals such that nations have little choice but to participate (the critical ILSA literature is full of this). In most cases, nations willingly sign up to testing regimes because they have come to believe that ILSAs offer their systems something that they either do not have or should have more of. If coercion is "the ability to get others to do what they otherwise would not" (Keohane and Nye 1998, p. 83), then "influence is seen as an emergent property that derives from the relationship between actors" (Betsill and Corell 2008, p. 22). Influence differs from coercion. There are obvious power imbalances in regards to intergovernmental organizations like the OECD, where more powerful countries have larger voices; however, that does not always result in coercive leverage towards less powerful actors. In other words, ILSAs may have the potential to be leveraged over educational systems to compel actor behavior, but that is not always the case and most national systems choose to participate, as evidenced by the growing number of participants that are self-electing to join the studies.

Here we present two cases that illustrate the policy influence of ILSAs. We chose these cases because they demonstrate possible overclaiming that ILSA results influenced change and evidence of ILSA influence that resulted in an unusual policy.

15.3 Policy Influence?

15.3.1 Case 1: PISA Shocks and Influence

Given the explicit goal of the OECD to inform national policy of member nations there is a considerable amount of research concerning the policy influence and impact of its flagship educational assessment PISA (Baird et al. 2011; Best et al. 2013; Breakspear 2012; Grek 2009; Hoplins et al. 2008; Kamens et al. 2013). PISA-inspired debates have resulted in a range of reforms including re-envisioning educational structures, promoting support for disadvantaged students (Ertl 2006) and developing new national standards aligned to PISA results (Engel and Rutkowski 2014), to name a few. The term "PISA shock" is now commonly used to highlight participating countries that were surprised by their sub-par PISA results and subsequently implemented educational policy reforms. Perhaps the most notable of these shocks occurred in Germany after its initial participation in PISA 2000. In response to lower than expected PISA scores, both federal and state systems in Germany implemented significant educational reforms (Ertl 2006; Gruber 2006; Waldow 2009). However,

Germany was not alone, and other countries, including Japan (Ninomiya and Urabe 2011) and Norway (Baird et al. 2011), experienced PISA shocks of their own.

In general, "shocks" attributed to ILSAs tend to be focused on PISA results rather than other studies. Notably, Germany, Norway, and Japan participated in the TIMSS assessment five years prior with similar results (in terms of relative rankings) to PISA (Beaton et al. 1996), yet this resulted in significantly less public discourse and little policy action. Of course, the perceived lack of a TIMSS shock could be for a variety of reasons. First, it is possible that the IEA simply does not have the appetite and/or political muscle to influence policy debates, leaving any discussions of results to academic circles. Second, the idea of a PISA shock may be misleading, representing an engineered discourse rather than a true social phenomenon. For example, Pons (2017) contended that much of the academic literature claiming that there is a PISA shock is biased because it contributes to a particular representation of what effect PISA "is expected to produce in conformity with the strategy of soft power implemented by the OECD" (p. 133). Further, similar to our discussion of the term influence, PISA shock is never fully conceptualized in the literature, making it difficult to compare and assess within and across systems. Pons (2017) further contended that assessing the effects of PISA on education governance and policy is difficult because the scientific literature on PISA effects are heterogeneous and fueled by various disciplines and traditions that ultimately lead to findings corresponding to those traditions.

15.3.2 Case 2: An Australian Example, Top Five by 2025

In 2012, the Australian Federal Government announced hearings into the Education Act 2012, which was subsequently passed and enacted on the January 1, 2014 (The Parliament of the Commonwealth of Australia 2012). This referred specifically to five agendas that were linked to school reform. These five reform directions were "quality teaching; quality learning; empowered school leadership; transparency and accountability; and meeting student need". These five reform directions were to improve school quality and underline the commitment of the Federal Government to have a system that was both high quality and high equity. The Act went on to outline that the third goal was:

> "…for Australia to be ranked, by 2025, as one of the top 5 highest performing countries based on the performance of Australian school students in reading, mathematics and science, and based on the quality and equity of Australian schooling" (The Parliament of the Commonwealth of Australia 2012, p. 3)

The Explanatory Memorandum that accompanies this Act includes in brackets, "(These rankings are based on Australia's performance in the Programme for International Student Assessment, or PISA.)" There are a number of curious things about this legislation that binds the Australian education system to be "top 5 by 2025." The first of these is that it shows, in the Australian context, that PISA and no doubt other

ILSAs have had an influence on policymakers. But the nature of the influence remains problematic, focused more on national rankings outcomes rather than considering what PISA tells Australia about its system and the policy decisions that have been made. While generic references to quality teaching and so on might work as political slogans, the reality is that they contain no specific direction or material that could ever be considered as a policy intervention or apparatus.

Second, it is curious that Australia legislated for a rank rather than a score or another indicator of the type that PISA provides such as some goal regarding resilient students. This would seem to suggest that this is how PISA is understood by policymakers in Australia, as a competitive national ranking system of achievement in mathematics, science, and reading. It would be easy to lay this solely at the feet of the policymaker, but it is probably not helped by the way that the OECD itself presents PISA as rankings to its member nations. Third, it seems fairly obvious that this use of ILSAs opens a system up to perverse incentives.

In the case of being "Top 5 by 2025," the influence of ILSAs falls short because it lacks intentionality. It also shows that those who are charged with making policy do not understand the data that they see paradoxically as: (1) determining a lack of quality and equity, and (2) clearly indicating what needs to be done as a result. In other words, without identifying what policy agendas in specific contexts could best respond to highlighted problems, ILSA data is often left to speak for itself as regards to what must be done within systems. The Education Act 2012 points to the impact and influence of PISA on Australian policymakers, yet paradoxically that impact and influence comes at the cost of policymaking itself.

Demonstrating the impact and influence of ILSAs on national policymaking is not the same as demonstrating that ILSAs are having a positive, or beneficial, impact or influence on policymaking. Legislating to be "Top 5 by 2025" is a prime demonstration of impact on policy that consequently opens the test up to perverse incentives. It is an absurd example, but should New Zealand outperform Australia on PISA, then invading them and taking over their country necessarily brings Australia closer to its goal. This neatly illustrates the problem of influence: how can society think about making the influence that ILSAs are having more useful than an obsession with rankings? This begins by considering how ILSAs might be used to hold policymaking to account, particularly at a time where "top down" accountability in most contexts seems to be about protecting policymakers from repercussions based on their policy decisions (see Lingard et al. 2015).

These two cases are illustrative in two ways. The first case of "PISA shock" shows that while influence is easy to claim, it is invariably linked to pre-existing interpretations and expectations. In other words, it appears that ILSAs are often used to buttress the preconceived policy frames rather than interrogating them. The second case shows that even where influence can be demonstrated, this does not necessarily improve policymaking nor does it improve the understanding that policymakers have regarding their system. Both cases exemplify the problem of influence. ILSAs may or may not influence change in systems and, where they do, the resultant change may be artificial, superficial, or downright silly. What is needed, then, are better tools to inform decision making through understanding, predicting, and evaluating influence.

This may go some way to help policymakers become more purposive in their use of ILSAs as evaluative tools.

15.4 A Model for Evaluating Influence

Oliveri et al. (2018) developed a model to assist countries in purposeful, intentional ILSA participation. Although the model was originally designed as a means for countries to evaluate whether their educational aims can be met by what an ILSA can deliver, it is generalizable for other uses. The model encourages intentionality by carefully considering whether claims that are made about what an ILSA can reasonably be expected to do are valid in a given country. Further, it helps establish a set of more valid interpretations of ILSA data for policymakers to use in their decision making. In our retooling of Oliveri et al.'s model, we use the same general structure (Fig. 15.1).

Using the model begins with a matching exercise between the influence attributed to ILSA results and what evidence the ILSA in question can provide to motivate changes that are said to be directly caused by ILSA results. In this step, the national system or other stakeholder must clearly articulate all the ways ILSAs have influenced or are anticipated to influence their educational system and the policy process. As an example, assume that country X scores lower than desirable results on the TIMSS grade 8 science assessment and that, as a consequence, policymakers in country X propose a policy that requires all science teachers to obtain an advanced degree

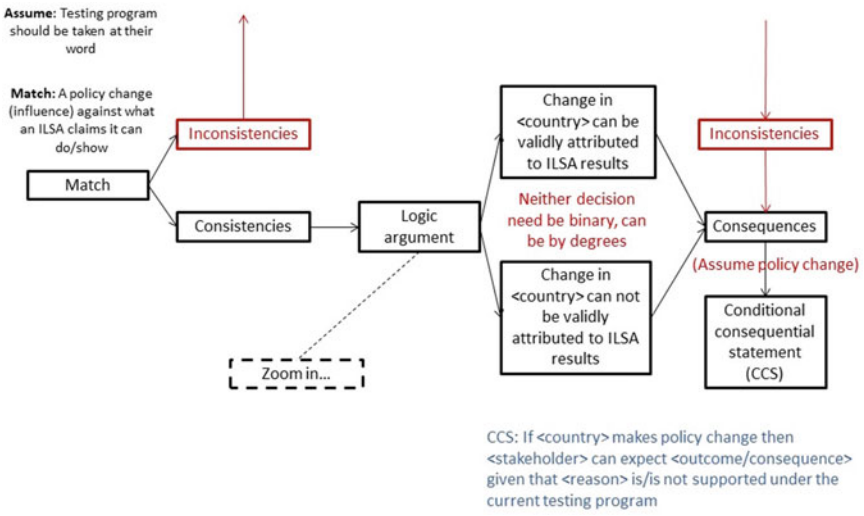


Fig. 15.1 Model for systematically understanding international large-scale assessment (ILSA) influence and the associated consequences for policymaking

(e.g., master's degree) by some set date. The matching analysis involves querying whether this policy can be attributed to TIMSS results. One line of argument might go like this: TIMSS provides results in science and teachers are asked about their level of education. Thus, initially, TIMSS appears to show evidence that science teachers with master's degrees on average teach classes with higher performance. This causal claim, then, is said to be initially consistent with the evidence that TIMSS can provide, setting aside formal causality arguments.

The next step in the process is a formal logic argument. The logic model (Fig. 15.2) is a derivative of Toulmin's (2003) presumptive method of reasoning. The process involves a claim supported by a warrant and additional evidence, which is often provided through a backing statement. In contrast, rebuttals provide counterevidence against the claim. The process allows for an informed decision to be made on whether and to what degree an ILSA can be reasonably said to have influenced a proposed or enacted policy. In the case of requiring master's degrees, the logic argument could proceed as follows. TIMSS was said to influence policymakers' decision that all science teachers should have a master's degree. The warrant for this decision is that better educated teachers produce higher average student achievement. The backing could be that in country X (and maybe other countries), teachers with master's degrees teach in classrooms with higher average TIMSS science achievement. Then, a possible rebuttal could be multifold. First, TIMSS does not use an experimental design. In the current example, teachers are not randomly assigned to treatment (master's degree) and control groups (education less than a master's degree). Thus, assuming that the teacher sample is strong enough to support the claim, a plausible explanation for this difference in country X is that teachers with master's degrees command a higher salary and only well-resourced schools can afford to pay the master's premium. A second plausible explanation is that only the very highly motivated seek a master's degree, which, rather than serving as an objective qualification is instead a signal of a highly motivated and driven teacher. A final decision might be that, given the observed associations (e.g., more educated teachers are associated with higher performance), country X decides to pursue the policy, thereby ignoring the evidence in the rebuttal.

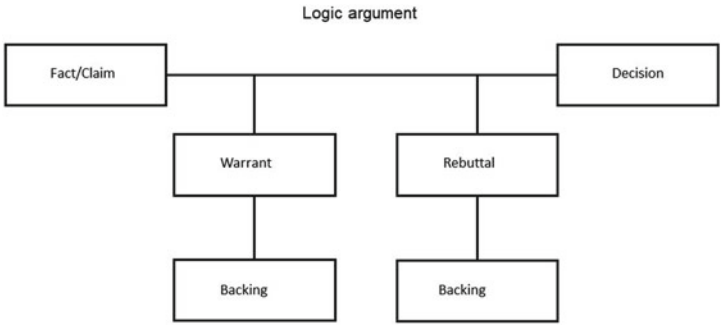


Fig. 15.2 Logic model for evaluating plausibility of influence

Returning to our model (Fig. 15.1), an analyst could conclude that, in spite of alternative explanations for the observed achievement differences between classes taught by teachers with different education levels, TIMSS results influenced policymakers' decision to enact a policy requiring all science teachers to have a master's degree. As noted, this conclusion could be by degrees and the analyst should include the warrant and rebuttal as the basis for this conclusion. The final, and perhaps most important step in the process is to consider the consequences of attributing influence and enacting a policy change based on ILSA results. This is in the form of a conditional consequential statement (CCS). Again, considering the TIMSS example, the CCS might be as follows: if country X requires science teachers to earn a master's degree, then the educational systems in country X can expect mixed achievement results, given that attributing influence to TIMSS results is not fully supported. Of course, there could be other consequences (i.e., medium-term teacher shortages, overwhelming demand for teacher training programs, and so on). However, it is important to delineate these consequences from those that are directly attributable to the influence of the ILSA in the given setting. To highlight this point, imagine that TIMSS did use an experimental design that randomly assigned teachers to different training levels. Further imagine that the TIMSS results showed that teachers with master's degrees taught classes that consistently outperformed classes taught by teachers with bachelor's degrees. Then, without going through the full exercise, assume that the decision from the logic argument is that the ILSA influence is fully supported. Then, a different CCS could be that if country X requires science teachers to have master's degrees, country X can expect higher average achievement in science on future TIMSS cycles.

15.4.1 Case 1: Worked Example

Norway had lower than expected PISA 2000 results and one resultant policy change was to implement a national quality assessment system (Baird et al. 2011). In fact, the OECD's report *Reviews of Evaluation and Assessment in Education: Norway* explained that poor PISA results spurred Norwegian policymakers to "focus attention on the monitoring of quality in education" (Nusche et al. 2011, p. 18). Using our model, a supporting analysis might proceed as follows. The first step is a matching analysis. That is, can PISA reasonably provide the necessary evidence to drive an expansion of a national evaluation system? Norwegian policymakers used PISA data, which showed that Norwegian students were lower performing than other peer industrialized nations, even though spending per child was one of the highest in the world. Initially, this claim might be regarded as consistent. This, then, triggers an analysis through the logic model. PISA is regarded by the OECD as a yield study (OECD 2019), measuring literacy and skills accumulated over the lifespan. It takes place at one point in time, when sampled students are 15 years old. Assuming this OECD claim is reasonable, PISA outcomes are attributable to a lifetime of learning. Then, the warrant for implementing a national assessment system could be that PISA performance was low relative to industrialized peers, and a national assessment

system will help Norwegian policymakers understand why. The backing is that PISA measures the accumulated learning through to the age of 15 and underachievement can be linked to learning deficiencies at some point between birth and age 15. A rebuttal to this argument, however, is that PISA does not explicitly measure schooling or curriculum, but rather, the totality of learning, both inside and outside of school. A national assessment that is (and should be) linked to the national curriculum will not fully align to PISA and risks missing the source of the learning deficiencies that lead to underperformance. The original argument also relies on the assumption that lower than desirable performance in the 2000 PISA cohort will be stable in future cohorts.

A CCS in this case might be: given that PISA showed Norway's achievement was lower than its industrialized peers and that PISA is a yield study, implementing a comprehensive national assessment system could reasonably be attributed to PISA results. But the fact that PISA does not measure curriculum imposes challenges in assessing the educational system and improving PISA outcomes. Thus, Norway can expect mixed results in future PISA cycles from enacting a policy that dictates a national assessment system. Here, a fairly simple but systematic analysis indicates that PISA is a mediocre evidentiary basis from which to enact such a policy and, although speculative, Norway might have used PISA as justification for a policy that they already wanted to initiate. This claim is substantiated to some degree by the fact that Norway's performance in TIMSS in 1995 was also relatively low; however, no similar policy reforms were enacted.

15.4.2 Case 2: Worked Example

The "Top 5 by 2025" case can be used to illustrate another worked example. PISA results clearly influenced Australian policymakers' desire to climb the ranks in the PISA league tables. Here, then, is a clear consistency; Australia's ranking in PISA drove a desire to improve on that position. The logic model becomes a somewhat trivial exercise where the warrant is that PISA rankings show the relative ordering of Australia's 15-year-olds in mathematics, science, and reading. The backing, again somewhat trivial, is the evidence that higher scores map onto better achievement in these domains. A plausible rebuttal is that simple rank ordering changes are somewhat meaningless without considering measures of uncertainty. For example, if Australia moves up two or three places in the league table, this change might not be statistically significant. Nevertheless, the decision is relatively straightforward: PISA results can reasonably be attributed to influencing the decision to seek a top five position in the PISA league tables. However, a desire for improvement, as understood by a position on a ranking (top five) within a timeframe (by 2025), does little to demonstrate improved decision making or better understanding of policy settings and their impact. A desire for improved rankings is the opposite of influencing policy; in fact it may act as a barrier for making policy changes necessary for that improved ranking.

Then, a conditional consequential statement (CCS) might be: if Australia uses PISA results to influence a decision to move up the league table, then Australian schools can expect initiatives intended to drive improvement in mathematics, science, and reading. Certainly, as with any CCS, there is no guarantee that these consequences will happen. Perhaps policymakers will take no concrete action to realize the gains necessary to move into the top five. Further, downstream consequences also become important in this example. If initiatives to improve mathematics, science, and reading come at the cost of other content areas (e.g., art, civics, or history), second order consequences might involve narrowing of the curriculum or teaching to the test. Depending on the incentives that policymakers use to achieve the top five goal, there might be undue pressure to succeed, raising the risk of cheating or otherwise gaming the system (e.g., manipulating exclusion rates or urging low performers to stay home on test day). Clearly, this is not a full analysis of the potential consequences of such a policy; however, this and the previous examples demonstrate one means of using the model to systematically evaluate whether ILSA results can reasonably influence a policy decision and what sort of consequences can be expected.

15.5 Discussion and Conclusions

Ensuring that ILSAs do not have undue influence on national systems requires active engagement from the policy community to include an examination of the intended and unintended consequences of participation. Thus, while ILSAs can be an important piece of evidence for evaluating an educational system, resultant claims should be limited to and commensurate with what the assessment and resulting data can support. It is imperative to recognize that ILSAs are tasked to evaluate specific agreed upon aspects of educational systems by testing a representative sample of students. For example, PISA generally assesses what the OECD and its member countries agree that 15-year-olds enrolled in school should know and do in order to operate in a free market economy. To do this, they measure the target population in mathematics, science, and reading. Importantly, PISA does not measure curriculum, nor does it measure constructs such as history, civics, philosophy, or art. Other assessments such as TIMSS have a closer connection to national curricula. Nevertheless, even TIMSS is at best only a snapshot of an educational system taken every four years. As such, inferences can be made but only provide a cross-sectional perspective of a narrowly defined population regarding its performance on a narrowly defined set of topics. Although the majority of ILSA data is collected based on rigorous technical standards and is generally of good quality, it is not perfect and includes error, some of which is reported and some of which is not.

Given the high stakes of ILSA results in many participating countries, it is not surprising that there are both promoters and detractors of the assessments. For example, in the academic literature there exists a strong critical arm arguing that some of the most prominent ILSAs do more harm than good to educational systems (Berliner 2011; Pons 2017; Sjøberg 2015). Questions around the value of ILSAs have

also been posed by major teacher unions (Alberta Teachers' Association 2016) and in the popular press (Guardian 2014) and, with a specific focus on PISA, by the director of the US Institute of Education Sciences (Schneider 2019). Importantly, the USA is one of the largest state funders of the most popular ILSAs (Engel and Rutkowski 2018). In the face of these and other criticisms, promoters of ILSAs contend that the tests have important information to offer and have had a positive influence on educational systems over time (Mullis et al. 2016; Schleicher 2013). Yet, as we have argued in this chapter, demonstrating the specific influence that ILSAs have had on educational systems is not always straightforward given the differing definitions of influence in the literature, along with the inherent complexity of isolating influence in large, complex national educational systems.

Once a definition of influence is established, as we have done in this chapter, it is possible to demonstrate instances when ILSAs clearly have an influence. Our two examples are both problematic for a number of reasons. First, both examples misuse ILSA results in order to encourage and implement policy change. In the case of Norway, poor results on PISA changed how their entire educational system is evaluated. In the case of Australia, policymakers set unrealistic goals and failed to explain how a norm referenced moving goal is justifiable as the ultimate benchmark of educational success, superseding the more common goals a citizenry and its leaders have for its education system. We contend that both cases demonstrate how, without a clear purpose and active management, ILSAs can influence policy in ways that were never intended by the designers.

Although admittedly not foolproof, we argue that one way to properly manage ILSA influence on educational systems is for participating systems to purposefully nominate their reasons for participation and forecast possible intended and unintended consequences of their own participation. Our proposed model depends on an empirical exercise with an assumption that it is possible to establish direct relationships in a complex, multifaceted policy world. We accept this criticism, but note that this is, in many ways, what ILSAs are attempting to do by collecting empirical data on large educational systems. Just like ILSAs, we do not contend that results from our empirical exercise will fully represent the ILSA/policy interaction. However, results from the model should provide more information than is currently available to countries and provide them with: (1) a better understanding of how participation may influence or be influencing their educational systems; and (2) what valid interpretations and uses of ILSA data can and should be made.

We realize that this is a serious endeavor fraught with difficultly but, without a clear purpose and plan for participation in ILSAs, those who understand what claims can and cannot be supported by the data are often sidelined once the mass hysteria of ILSA results enter the public sphere. As such, we developed our model as tool for those who fund participation in ILSAs to be more purposeful concerning the process. We realize the suggested model will require most systems to engage in additional work, but we maintain that systematically evaluating the degree to which ILSA results can serve as the basis for implementing policy changes will help prevent misuse of results. Further, documenting the process provides transparency surrounding what national governments expect from the data and enables testing

organizations to better explain to their clients what valid information the ILSAs can provide. Outside of education, similar forecasting models are well established in the policy literature and common practice in many governmental projects around the world (Dunn 2011, p. 118). Given the high stakes of ILSA results, anticipating or forecasting consequences from the perspective of what ILSAs can and cannot do is a step toward better informing policymakers in their decision making. Our model can also be used to link the influence of ILSAs to any proposed or enacted policy. In other words, the model can work as a tool to understand whether the results from ILSAs are an adequate evidentiary basis to support or inform the policy. We recognize our model will not prevent all misuse or unintended influence of ILSAs, but it invites a more purposeful process.

Educational systems and policies designed to guide and improve them are extremely complex. We are not so naive as to believe that it will ever be possible to document or even understand exactly how ILSAs influence or impact participating educational systems. However, that does not mean that the endeavor is fruitless. We contend that defining terms and participating in an intentional process are two important ways toward understanding how ILSAs influence policy and holding policymakers and testing organizations accountable in how they promote and use the assessments.

References

Alberta Teachers' Association. (2016). Association to push for PISA withdrawal [webpage]. Edmonton/Calgary, Canada: The Alberta Teachers' Association. https://www.teachers.ab.ca/News%20Room/ata%20news/Volume%2050%202015-16/Number-18/Pages/PISA-withdrawal.aspx.

Amrein-Beardsley, A., Berliner, D. C., & Rideau, S. (2010). Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas, 18,* 1–36.

Baird, J., Isaacs, T., Johnson, S., Stobart, G., Yu, G., Sprague, T., & Daugherty, R. (2011). *Policy effects of PISA*. Oxford, UK: Oxford University Centre for Educational Research. http://oucea.education.ox.ac.uk/wordpress/wp-content/uploads/2011/10/Policy-Effects-of-PISA-OUCEA.pdf.

Barber, M., Moffit, A., & Kihn, P. (2010). *Deliverology 101: A field guide for educational leaders.* Thousand Oaks, CA: Corwin.

Beaton, A. E., Mullis, I., Martin, M., Gonzalez, E., Kelly, D., & Smith, T. (1996). *Mathematics achievement in the middle school years. IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: Boston College. https://www.iea.nl/publications/publications/mathematics-achievement-middle-school-years.

Berliner, D. C. (2011). The context for interpreting PISA results in the USA: Negativism, chauvinism, misunderstanding, and the potential to distort the educational systems of nations. In M. A. Pereyra, H. G. Kotthoff, & R. Cowen (Eds.), *Pisa under examination* (pp. 75–96). Comparative Education Society in Europe Association, Vol 11. Rotterdam, the Netherlands: Sense Publishers.

Best, M., Knight, P., Lietz, P., Lockwood, C., Nugroho, D., & Tobin, M. (2013). *The impact of national and international assessment programmes on education policy, particularly policies regarding resource allocation and teaching and learning practices in developing countries.* Final report. London, UK: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London. https://research.acer.edu.au/ar_misc/16.

Betsill, M. M., & Corell, E. (2008). *NGO diplomacy: The influence of nongovernmental organizations in international environmental negotiations*. Cambridge, MA: MIT 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. Paris, France: OECD Publishing. https://doi.org/10.1787/5k9fdfqffr28-en.

Cox, R. W., & Jacobson, H. K. (1973). *The anatomy of influence: Decision making in international organization*. New Haven, CT: Yale University Press.

Dunn, W. N. (2011). *Public policy analysis* (5th ed.). Boston, MA: Pearson.

Engel, L. C., & Rutkowski, D. (2014). Global influences on national definitions of quality education: Examples from Spain and Italy. *Policy Futures in Education*, *12*(6), 769–783. https://doi.org/10.2304/pfie.2014.12.6.769.

Engel, L. C., & Rutkowski, D. (2018). Pay to play: What does PISA participation cost in the US? *Discourse: Studies in the Cultural Politics of Education*, 1–13. https://doi.org/10.1080/01596306.2018.1503591.

Ertl, H. (2006). Educational standards and the changing discourse on education: The reception and consequences of the PISA study in Germany. *Oxford Review of Education, 32*(5), 619–634. https://doi.org/10.1080/03054980600976320.

Grek, S. (2009). Governing by numbers: The PISA "effect" in Europe. *Journal of Education Policy, 24*(1), 23–37.

Gruber, K. H. (2006). The German "PISA-Shock": Some aspects of the extraordinary impact of the OECD's PISA study on the German education system. In H. Ertl (Ed.), *Cross-national attraction in education: Accounts from England and Germany* (pp. 195–208). Oxford, UK: Symposium Books Ltd.

Guardian. (2014, May 6). OECD and Pisa tests are damaging education worldwide. *The Guardian*. https://www.theguardian.com/education/2014/may/06/oecd-pisa-tests-damaging-education-academics.

Gür, B. S., Celik, Z., & Özoğlu, M. (2012). Policy options for Turkey: A critique of the interpretation and utilization of PISA results in Turkey. *Journal of Education Policy, 27*(1), 1–21.

Hoplins, D., Pennock, D., Ritzen, J., Ahtaridou, E., & Zimmer, K. (2008). *External evaluation of the policy impact of PISA*. Report no. EDU/PISA/GB(2008)35/REV1. Paris, France: OECD.

Kamens, D. H., Meyer, H.-D., & Benavot, A. (2013). *PISA, power, and policy: The emergence of global educational governance*. Oxford, UK: Symposium Books Ltd.

Keohane, R. O., & Nye, J. (1998). Power and interdependence in the information age. *Foreign Affairs, 77*(5), 81–94. https://doi.org/10.2307/20049052.

Lingard, B., Martino, W., Rezai-Rashti, G., & Sellar, S. (2015). *Globalizing educational accountabilities*. Abingdon, UK: Routledge.

Loughland, T., & Thompson, G. (2016). The problem of simplification: Think-tanks, recipes, equity and "Turning around low-performing schools". *The Australian Educational Researcher, 43*(1), 111–129. https://doi.org/10.1007/s13384-015-0190-3.

Mullis, I. V., Martin, M. O., & Loveless, T. (2016). *20 years of TIMSS: International trends in mathematics and science achievement, curriculum, and instruction*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Boston College. https://www.iea.nl/publications/study-reports/international-reports-iea-studies/20-years-timss.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.

Ninomiya, A., & Urabe, M. (2011). Impact of PISA on education policy: The case of Japan. *Pacific-Asian Education, 23*(1), 23–30.

Nusche, D., Earl, L., Maxwell, W., & Shewbridge, C. (2011). *OECD reviews of evaluation and assessment in education: Norway*. Paris, France: OECD Publishing.

OECD. (2017). *PISA 2015 technical report*. Paris, France: OECD. http://www.oecd.org/pisa/data/2015-technical-report/.

OECD. (2018). FAQ: PISA [webpage]. Paris, France: OECD. http://www.oecd.org/pisa/pisafaq/.

OECD. (2019). *PISA 2018 Assessment and Analytical Framework*. Paris, France: OECD Publishing. https://doi.org/10.1787/b25efab8-en.

Oliveri, M. E., Rutkowski, D., & Rutkowski, L. (2018). Bridging validity and evaluation to match international large-scale assessment claims and country aims. *ETS Research Report Series, 2018*(1), 1–9. https://doi.org/10.1002/ets2.12214.

Pons, X. (2017). Fifteen years of research on PISA effects on education governance: A critical review. *European Journal of Education, 52*(2), 131–144. https://doi.org/10.1111/ejed.12213.

Purves, A. C. (1987). The evolution of the IEA: A memoir. *Comparative Education Review, 31*(1), 10–28. https://doi.org/10.1086/446653.

Rautalin, M., & Alasuutari, P. (2009). The uses of the national PISA results by Finnish officials in central government. *Journal of Education Policy, 24*(5), 539–556. https://doi.org/10.1080/02680930903131267.

Rizvi, F., & Lingard, B. (2009). *Globalizing education policy*. Abingdon, UK: Routledge.

Schleicher, A. (2013). Lessons from PISA outcomes. *OECD Observer* No. 297 Q4 2013. Paris, France: OECD. http://oecdobserver.org/news/fullstory.php/aid/4239/Lessons_from_PISA_outcomes.html.

Schneider, M. (2019). Mark Schneider: My response to essay rebuttal—I'm concerned about the PISA exam's future and the implications of its sponsor's global ambitions. *The 74 million news site* [webpage]. https://www.the74million.org/article/mark-schneider-my-response-to-essay-rebuttal-im-concerned-about-the-pisa-exams-future-and-the-implications-of-its-sponsors-global-ambitions/.

Schwippert, K., & Lenkeit, J. (Eds.). (2012). *Progress in reading literacy in national and international context: The impact of PIRLS 2006 in 12 countries*. Münster, Germany: Waxmann Verlag GmbH.

Sellar, S., & Lingard, B. (2013). The OECD and global governance in education. *Journal of Education Policy, 28*(5), 710–725. https://doi.org/10.1080/02680939.2013.779791.

Sjøberg, S. (2015). OECD, PISA, and globalization: The influence of the international assessment regime. In C. H. Tienken & C. A. Mullen (Eds.), *Education Policy Perils* (pp. 114–145). Abingdon, UK: Routledge.

Takayama, K. (2008). The politics of international league tables: PISA in Japan's achievement crisis debate. *Comparative Education, 44*(4), 387–407. https://doi.org/10.1080/03050060802481413.

The Parliament of the Commonwealth of Australia. (2012). *Australian Education Act 2012*. Pub. L. No. 223, C2012B00223 (2012). Canberra, Australia: Australian Government. https://www.legislation.gov.au/Details/C2012B00223.

Toulmin, S. E. (2003). *The uses of argument* (Updated ed.). Cambridge, UK: Cambridge University Press.

Waldow, F. (2009). What PISA did and did not do: Germany after the "PISA-shock". *European Educational Research Journal, 8*(3), 476–483.

David Rutkowski is an Associate Professor with a joint appointment in Educational Policy and Educational Inquiry at Indiana University (IU). Prior to IU, David was a Professor of Education at the Center for Educational Measurement (CEMO) at the University of Oslo, Norway. He also worked as a researcher for the International Association for the Evaluation of Educational Achievement (IEA) in Hamburg Germany. David's research is focused in the area of educational policy, evaluation, and educational measurement with specific emphasis on international large-scale assessment. David has collaborated with or consulted for national and international organizations including the US State Department, USAID, UNESCO, IEA and the OECD. David has worked on and lead evaluations and projects in over 20 countries to include Afghanistan, South Sudan, Trinidad and Tobago and the US. He currently is the editor of the IEA policy brief series, co-editor of the journal *Discourse*, serves on the IEA publication editorial committee (PEC), and is a board member of several academic journals.

Greg Thompson is Associate Professor of Education Research at Queensland University of Technology (QUT). Prior to entering academia, he spent 13 years as a high school teacher in Western Australia. Thompson's research focuses on educational theory, education policy, and the philosophy/sociology of education assessment, accountability and measurement with a particular emphasis on large-scale testing. Recent books include *The Global Education Race: Taking the Measure of PISA and International Testing* (Brush Education), *National Testing in Schools: An Australian Assessment* (Routledge) and *The Education Assemblage* (Routledge).

Leslie Rutkowski is Associate Professor of Inquiry Methodology at Indiana University. She earned her Ph.D. in Educational Psychology, specializing in Statistics and Measurement, from the University of Illinois at Urbana-Champaign. Leslie's research is in the area of international large-scale assessment. Her interests include latent variable modeling and examining methods for comparing heterogeneous populations in international surveys. In addition to a recently funded *Norwegian Research Council* grant on developing international measurement methods, Leslie published the edited volume *Handbook of International Large-Scale Assessment* (Rutkowski, von Davier, and Rutkowski, 2014) with Chapman & Hall. She is also currently writing a text book on large-scale assessment under the Guilford stamp. Leslie currently serves as associate editor for the Springer journal, *Large-scale Assessments in Education,* and as executive editor for the Springer book series, *IEA Research for Education.*