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.

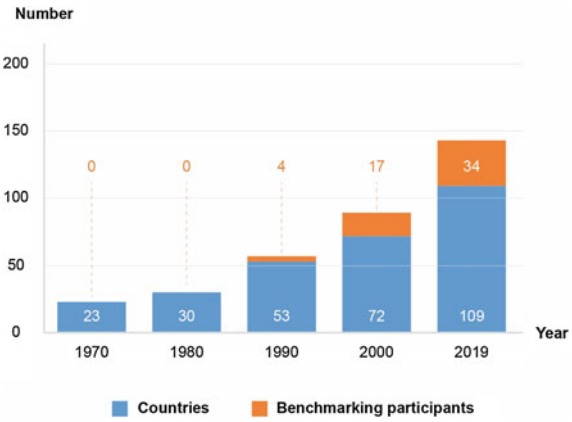

Fig. 2.1 Cumulative growth in unique participants in IEA studies over time. *Note* totals represent unique participants across all studies. The Second International Mathematics Study, for example, had only 20 participants, with outcome data collected from just 15 participants for the cross-sectional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

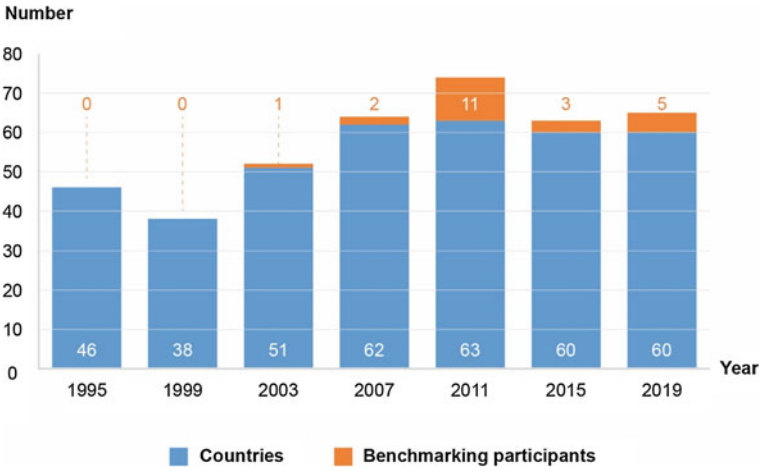

Fig. 2.2 Compositional changes and growth in participation in TIMSS, 1995–2015

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

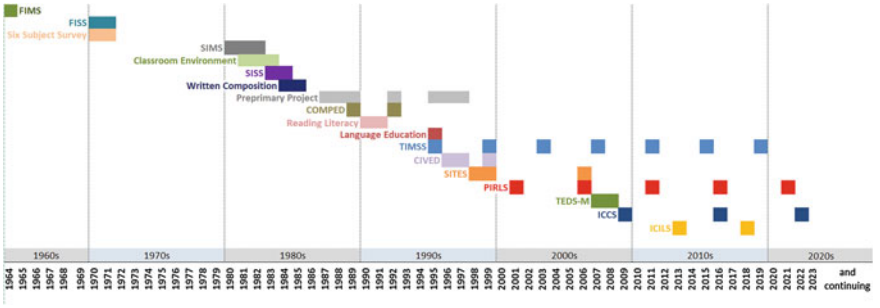

Fig. 2.3 Development of IEA's program of ILSAs from 1995 onwards. *Source* Hastedt (2020) Springer International Publishing, reproduced with permission of Springer Nature Switzerland AG. Disclaimer: this figure is copyright protected and excluded from the open access licence

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.