20

Large-Scale Assessments in International Contexts

Henry I. Braun
Boston College

Irwin Kirsch
ETS (Retired)

Braun, H. I., & Kirsch, I. (2025). Large-scale assessments in international contexts. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 1423–1510). Oxford University Press.
DOI: 10.1093/oso/9780197654965.003.0020

Broadly defined, international large-scale assessments (ILSAs) are survey-based studies designed to assess the knowledge, skills, experiences, behaviors, and attitudes of populations in an international, comparative context. Depending on the population of interest, ILSAs can be characterized as school based or household based. School-based ILSAs select participants (e.g., students, teachers, school principals) based on their school membership and location, whereas household-based ILSAs select participants based on the location of their households.

For school-based populations, ILSAs produce estimates of distributions of learning outcomes in key domains for subpopulations defined by various combinations of individual characteristics and contextual factors. The data collected by school-based ILSA enable estimation of the strength of the relationships between learning outcomes and students' backgrounds and school-related factors. Moreover, administering an ILSA periodically enables tracking over time of learning outcomes, as well as changes in the patterns of relationships of learning outcomes to selected individual characteristics and background factors. Analogous statements pertain to household-based studies, usually referred to as surveys of adult populations, with the modification that the focus on school characteristics is replaced with educational, social, and labor market outcomes. In addition, because adult surveys encompass a broad age range, they enable comparisons across age cohorts, yielding useful information on secular trends.

Today's ILSAs can trace their origins to pilot studies that began in the late 1950s and, by the early 1960s, had led to the creation of the International Association for the Evaluation of Educational Achievement (IEA). From a methodological perspective, ILSAs are also beneficiaries of efforts in the United States during the 1960s that culminated in the National Assessment of Educational Progress (NAEP). Prior to these initiatives, no systematic or standardized comparative effort had been undertaken to determine what different student populations knew or could do in various cognitive domains, including language arts and mathematics. This newfound attention to educational outcomes marked a major enhancement of reporting efforts that heretofore had focused primarily on educational inputs, such as the number of schools, teachers, and students in a particular country and how these numbers changed over time.[1]

The sponsors and developers of IEA assessments and of NAEP believed that the ongoing, systematic collection and analysis of outcome data, as well as patterns of relationships between those outcomes and various factors, would yield valuable insights. The interest in using the information generated by large-scale assessments to track trends in outcomes and to acquire deeper understandings about the relationships between outcomes on the one hand and individual or contextual characteristics on the other hand marked a milestone in an evidence-based approach to educational policy. Over time, the complex dynamics between ILSAs and the ever-changing education policy landscape have had important implications for the methodologies used to develop and deploy these assessments, as well as for data analysis and reporting results, topics we will return to later in this chapter. It bears mentioning that a key feature of ILSAs is that they always report score distributions and patterns of relationships at a group level,

in contrast to traditional testing programs that report individual-level results. This feature also has important implications for the design and analysis of ILSAs and is the reason why such large-scale assessments are often referred to as group-score assessments.

Since their inception, ILSAs have experienced remarkable growth in participation and salience, reflecting, we believe, perceptions of increased utility of comparative information among policy makers and key stakeholders. Perceived utility is fueled in large part by mounting concerns about the levels and distributions of human capital[2] and how they are associated with longer term outcomes for individuals and societies. Of course, the actual utilization of the information gleaned from ILSAs depends on a host of political, economic, and other factors that vary across countries, as well as within countries, over time (Feuer, 2012, 2013; Ritzen, 2013).

At the same time, because of the cross-sectional nature of ILSAs, there are limitations to the inferences one can make regarding the relative efficacy of different educational systems in building human capital as well as the consequences for individuals as they make their way through the educational system and then later transition into adulthood and participate in the labor market (Singer & Braun, 2018). Although ILSA outcomes, especially league tables for school-level results, generate headlines and lead to calls for action, the road from evidence to constructive policy impact is strewn with potholes. In particular, because of their cross-sectional approach to data collection, attempts to employ ILSA data to establish causal linkages are at best speculative and at worst misleading. Unfortunately, these limitations are not always understood or respected in policy discussions. The increased visibility of ILSA results along with occasional, ILSA-based, overly broad prescriptions for education reforms have given rise to criticisms of the influence of ILSAs on education policy—particularly with regard to the "homogenization" of national education systems (e.g., Benavot & Smith, 2020; Meyer & Benavot, 2013). A related criticism concerns the applicability of education policies in one nation to the systems of another—especially if the nations differ markedly with respect to history, culture, and governance (Carnoy et al., 2015). These concerns are addressed in this chapter's section on "Policy and Political Challenges."

In the case of adult surveys, interest in the results is distributed across many governmental departments, each being concerned with one or another aspect of the study. We can expect that as a consequence of pandemic-related disruptions, there will be increased attention to the need for upskilling or reskilling. Concerns about the use of data for policy in this realm appear to be milder than those in the realm of K–12 education policy.

In the early 21st century, ILSA results enter policy landscapes characterized by complex interactions among multiple stakeholders with access to a variety of sources of information. Accordingly, it is not a simple matter to evaluate whether changes in ILSA instrumentation and procedures are enhancing the value of the assessment enterprise as a source of policy-relevant information. To this task, we advance the notion that a deeper understanding of the utility of ILSAs can be obtained through consideration of a framework introduced by Messick (1987). Messick argued that large-scale assessments

were a form of policy research and, accordingly, should be judged by their contributions to policy analysis and policymaking. His framework was intended to decompose policy utility into its constituent components to provide clearer guidance to assessment sponsors and assessment designers. In this chapter, Messick's framework is refined and extended to take account of more than 30 years of further development and, especially, the transition to digitally-based technologies. Employing this framework, we argue that digital technologies have the potential to further increase the relevance and utility of ILSAs, contributing to their continued growth.

Increased participation in ILSAs, along with the greater prominence of their results, has led to a number of challenges. The task of accommodating increased heterogeneity in the distributions of proficiency and in the languages of the assessments, within rigid constraints of time and budget, has heretofore been accomplished through significant process innovation and methodological advances (as will be detailed in the section "Transitioning to Digitally-Based Assessments"). For example, countries can choose in which language or languages they want to assess and report results. For school-based surveys, these results are typically related to the language(s) of instruction, whereas for adult assessments, they are related to the language(s) of the society. Nonetheless, there are surely limits that will soon be reached and alternative strategies will be required.

A second challenge is that increasing heterogeneity among participating countries has given rise to concerns that cross-national comparability has been compromised to some degree. More specifically, in the case of school-based ILSAs, comparative analyses are complicated by the fact that, across countries, substantially different proportions of a birth cohort are enrolled in school and, of those enrolled, not all have been exposed to the same educational experiences, varying both in quantity and in quality. In the case of adult surveys, different proportions of the target populations are accessible through household surveys. Helping stakeholders understand the implications for policy analysis of these challenges to comparability remains a work in progress.

Following this introduction, we provide a brief description of several key large-scale student and adult assessments. Then we place these assessments in their historical context noting two key inflection points—the first marked by significant innovations in measurement and psychometrics and the second by significant innovations in administration and data collection resulting from the introduction of digital platforms. With the emergence of these assessments as significant sources of credible evidence for policy makers and other key stakeholders, we next introduce and extend a framework proposed by Messick (1987) that provides both developers and users of ILSAs with a set of design criteria to evaluate the potential utility of these surveys.

With the transition to digital platforms, we then focus on the key role of digital platforms in the management, development, and implementation of ILSAs and how new technologies are impacting their workflow and related processes. Next we address methodological advances impacting the scaling and analysis of the data, and also the production and dissemination of different data products. Then we focus on a range of issues concerning ways to evaluate the impact of ILSAs and some key technical

challenges that must be addressed in future cycles, as well as a number of the political challenges their growth and influence have raised. We conclude with a summary and final conclusions.

OVERVIEW OF KEY INTERNATIONAL ASSESSMENTS

A major focus of this chapter is on the three major school-based ILSAs: Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), and Programme for International Student Assessment (PISA). The first two are sponsored by IEA; the last is sponsored by the Organisation for Economic Co-operation and Development (OECD). However, it also describes a number of other ILSAs covering a range of domains and populations.

These assessments each include two sets of instruments: one for measuring cognitive outcomes and one for eliciting information on a range of background characteristics and contextual factors. Cognitive outcomes include the core domains of reading, literacy, mathematics, and science knowledge and skills, as well as other domains that are assessed with less frequency or are introduced as innovative domains (e.g., problem-solving, financial literacy). Instruments for eliciting background characteristics are targeted to the participating students, their teachers, school principals, and their parents. It bears mentioning that the IEA assessments employ a curriculum-based model for defining their assessment frameworks. In contrast, PISA employs more of a literacy-based model. Although these studies may appear to be very similar, deep conceptual differences manifest in many ways, including sample selection, item development, and instrument design. Table 20.1 contains information on each of these assessments, providing a brief description of content area, numbers of participants, and years conducted.

TIMSS is administered on a 4-year cycle to students in Grades 4 and 8.[3] Students are assessed in both mathematics and science. Paying participants include countries, subnational jurisdictions (e.g., states and provinces), and even school districts. In 2019, TIMSS began a transition to computer-based assessment delivery. About half the countries used digital devices for the assessment. TIMSS 2023 completed the transition to computer-based assessment, facilitating the introduction of innovative item types for more comprehensive coverage of the TIMSS mathematics and science frameworks. Scores in each TIMSS cycle are placed on a scale originally established in 1995.

PIRLS is a literacy assessment administered to fourth graders on a 5-year cycle.[4,5] In 2016, the IEA introduced ePIRLS, a computer-based assessment of online reading. PIRLS participants had the option of also participating in a pilot test of ePIRLS. For participating countries, students typically took PIRLS one day and ePIRLS on the following day.[6] For the 2016 administration, the IEA also introduced PIRLS Literacy, a reading assessment intended for countries with lower levels of reading proficiency. Through a linking strategy based on common items, scores on this assessment are placed on the main PIRLS scale.

Table 20.1 School-Based International Tests of Educational Achievement: Scope and Timing

Sponsor[a]	Description	Countries	Year(s) Conducted
IEA	First International Mathematics Study (FIMS)	12 countries	1964
IEA	Six Subject Study		1970–1971
	Science	19 systems	
	Reading	15 countries	
	Literature	10 countries	
	French as a foreign language	8 countries	
	English as a foreign language	10 countries	
	Civic education	10 countries	
IEA	First International Science Study (FISS; part of Six Subject Study)	19 countries	1970–1971
IEA	Second International Mathematics Study (SIMS)	10 countries	1982
IEA	Second International Science Study (SISS)	19 systems	1983–1984
ETS	First International Assessment of Educational Progress (IAEP-I, Mathematics Study and Science)	6 countries (12 systems)	1988
ETS	Second International Assessment of Educational Progress (IAEP-II, Mathematics and Science)	20 countries	1991
IEA	Reading Literacy (RL)	32 countries	1990–1991
IEA	Computers in Education	22 countries	1988–1989
		12 countries	1991–1992
IEA	International Computers and Information Literacy Study	35 countries	2013
		13 countries	2018
		30 countries	2023
IEA	Preprimary Project:		
	Phase I	11 countries	1989–1991
	Phase II	15 countries	1991–1993
	Phase III (longitudinal follow-up of Phase II sample)	15 countries	1994–1996
IEA	Third International Mathematics and Science Study (TIMSS)	45 countries	1994–1995
		40 countries	1997–1998

Sponsor[a]	Description	Countries	Year(s) Conducted
IEA	Civics Education Study (ICCS)	28 countries	1999
		35 countries	2009
		24 countries	2016
		25 countries	2022
OECD	Program for International Student Assessment (PISA)[b]	43 countries	2000 (reading)
		41 countries	2003 (math)
		57 countries	2006 (science)
		65 countries	2009 (reading)
		64 countries	2012 (math)
		72 countries	2015 (science)
		79 countries	2018 (reading)
		About 88 countries	2022 (math)
IEA	Progress in International Reading Literacy Study (PIRLS)	34 countries	2001
		41 countries	2006
		48 countries	2011
		50 countries	2016
		About 70 countries	2021
IEA	Trends in International Mathematics and Science Study (TIMSS)	45 countries	2003
		48 countries	2007
		63 countries	2011
		57 countries	2015
		64 countries	2019
		About 65 countries	2023

Note. Adapted from "Sampling Issues in Design, Conduct, and Interpretation of International Comparative Studies of School Achievement" by J. R. Chromy, in A. C. Porter and A. Gamoran (Eds.), *Methodological Advances in Cross-National Surveys of Educational Achievement* (pp. 80–117, Table 4.1), 2002, National Academies Press.
[a] IEA = International Association for the Evaluation of Educational Achievement; OECD = Organisation for Economic Co-operation and Development.
[b] Subject in parentheses is the focal domain for that administration.

PISA assesses 15-year-old students in the core domains of reading, mathematics, and science literacy on a 3-year cycle. In each country, eligible students are those enrolled in Grade 7 or above. During each cycle, one of the three core domains is the focal domain and therefore more testing time (i.e., items) is devoted to that domain than to the others. During the 2015 cycle, PISA completed the transition from a paper-based assessment to a computer-based assessment while maintaining a limited

paper-based option for countries not selecting the computer-based option. During the 2018 cycle, PISA also introduced multistage, adaptive testing as part of the computer-based assessment. In PISA 2018, each student was tested in two of the three core domains. As options for countries participating in the computer-based assessment, PISA currently assesses an innovative domain beyond the three core domains as well as the domain of financial literacy. Beginning with the 2025 cycle of PISA there will be an optional foreign language assessment that will alternate cycles with the optional financial literacy assessment.

In addition to these school-based surveys, this chapter also discusses one major assessment of adult skills, the Programme for the International Assessment of Adult Competencies (PIAAC), sponsored by the OECD. The assessment of adult literacy began with a number of national assessments in the United States.[7] The design of the International Adult Literacy Survey (IALS), conducted in three rounds between 1994 and 1999, drew on the frameworks developed and refined by these earlier national surveys. Information on IALS and subsequent international surveys of adult literacy can be found in Figure 20.1.

PIAAC, first fielded in 2012, is a computer-based assessment for adults ages 16–65 in literacy and numeracy and a measure of problem-solving in technology-rich environments. This household survey was administered in three rounds over a period of 5 years (2012, 2015, and 2017). These rounds enabled more countries to participate within

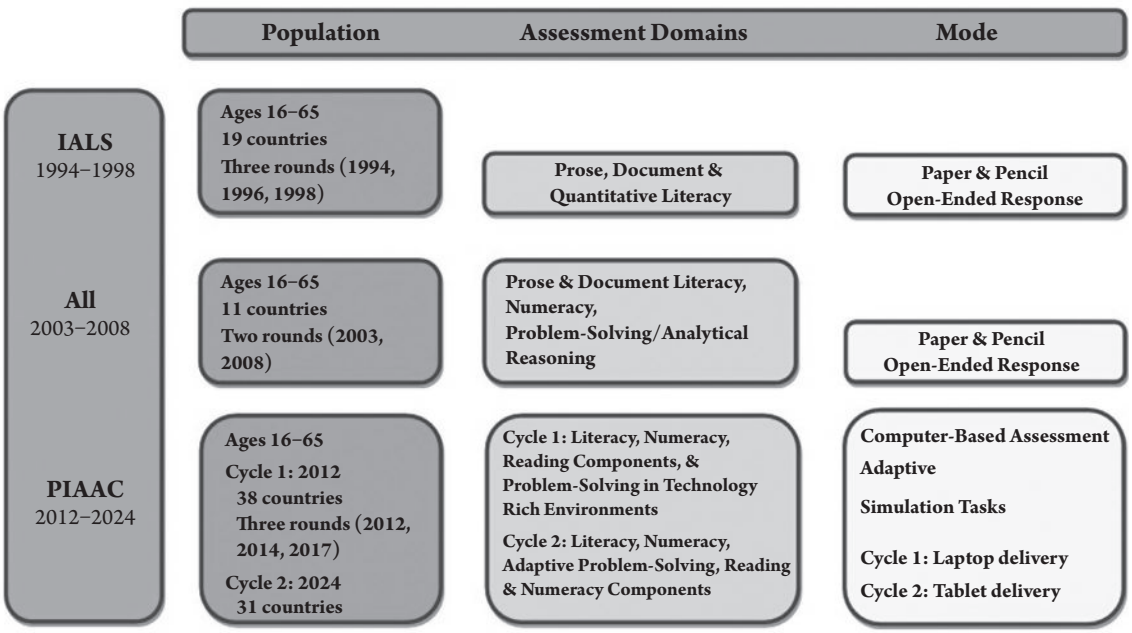

FIGURE 20.1
International Adult Assessments

a cycle and allowed countries (e.g., the United States) to do further national studies between cycles.

Laptops were used for data collection in all three rounds. For those respondents who were unwilling or unable to take the assessment on a laptop computer, a paper-and-pencil option was available for the assessment of literacy and numeracy only. In addition to the cognitive instruments, PIAAC incorporates an extensive background questionnaire, capturing information on a broad range of issues including demographics, education, labor force status, work experience, occupation, use of cognitive skills at home and at work, and incomes.

PIAAC is developed and implemented on a 10-year cycle. At the time of writing, PIAAC is conducting Cycle 2 with 31 participating countries, all of which are employing tablets for survey delivery.

It should be noted that in addition to the well-known global assessments that are discussed in this chapter, a number of other international (but regional) assessments play important roles in regional conversations related to education by providing information for credible cross-national comparisons. Information about those assessments can be found in a recent World Bank report (Clarke & Luna-Bazaldua, 2021).

Finally, the IEA also sponsors two periodic, specialized assessments: the International Computer and Information Literacy Study (ICILS) and International Civic and Citizenship Education Study (ICCS). The ICILS instrument assesses students in the eighth grade (or its equivalent) with regard to their computer literacy and information literacy. An optional module assesses computational thinking. Through the collection of additional background information, ICILS reports describe how these components of digital competence relate not only to one other, but also to relevant school and out-of-school contexts.[8]

The ICCS assesses students in the eighth or ninth grades. It reports on 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. In addition, ICCS collects rich contextual data on the organization and content of civic and citizenship education in the curriculum, teacher qualifications and experiences, teaching practices, school environment and climate, and home and community support.[9]

HISTORY OF INTERNATIONAL ASSESSMENTS

Comparative international assessments had their genesis in a pilot study involving a mathematics assessment in six countries in the late 1950s, followed by a 12-country study, comprising five subjects, conducted by the IEA in 1964 (Husén, 1967; Suter, 2019). These early studies were motivated by the recognition that the education systems and policies of different countries constituted a natural laboratory, and by using a common assessment, international comparisons could yield insights on what strategies and

policies deserved further study. By the same token, the structure of the accompanying background questionnaire was motivated more by models of learning than by considerations of education policy.

Subsequently, the IEA sponsored a series of studies through the 1970s and 1980s, focusing primarily on mathematics and science (assessed separately), culminating in the Third International Mathematics and Science Study (TIMSS)[10] in 1995 with 45 participating countries. Building on its Six Subjects Study (1970–1971) and its Reading Literacy Study (1990–1991), the IEA introduced PIRLS in 2001.

It is remarkable that in the late 1980s, in anticipation of the growing importance of international assessments, a number of U.S. federal agencies established, under the auspices of the National Academy of Sciences, the Board on International and Comparative Studies in Education (BICSE).[11] The main focus of BICSE was to evaluate the technical quality of these studies and to assess their policy relevance—all, presumably, to provide high-level guidance on how to enhance both (Bradburn & Gilford, 1990). As described by Heyneman and Lee (2014, p. 41), BICSE made a strong case for the utility of international studies. Echoing the original impetus behind such studies, it argued that variations in policies and practices among countries provided a natural laboratory in which to study the consequences of such differences. Further, BICSE noted the value to the United States of cross-national comparisons that could bring to light new concepts or an empirical basis for challenging long-held assumptions. These considerations have retained their cogency over the ensuing decades.[12]

The OECD introduced PISA in 2000 and PIAAC in 2012.[13] PIAAC built on two earlier assessments of adults ages 16–65: IALS, which was administered in multiple rounds from 1994 to 1999, and the survey of Adult Literacy and Lifeskills, administered between 2003 and 2008 (see Kirsch et al., 2017, for a history of adult assessments). As will be elaborated, PIAAC is distinguished by the fact that it was the first ILSA to be designed from the outset to be delivered by computer. To fully appreciate the evolution of the systems that support ILSA efforts, however, it is informative to review the history of NAEP.

The history of NAEP has been addressed at length elsewhere (Beaton & Barone, 2017; Jones & Olkin, 2004). Building on substantial planning and development work, the first NAEP assessment was conducted in 1969. Bowing to political constraints related to the primacy of states and local districts in matters of education, the NAEP team had to ensure that the assessment design and sampling plan were such that no summative test scores could be developed, no individual-level results could be reported, and no population groups could be described at the state or local level. This mandate was further reinforced by the decision to report results only at the item level, with each item connected to a particular learning objective so that it could be judged for relevance by both professional educators and individual citizens.

Over the next decade, the context for education and education policy evolved considerably—resulting in a greater federal role in education alongside a growing concern that many schools across the country were neither serving well the needs

of particular subgroups nor accommodating changes in society and labor markets. Policy makers and other key stakeholders began to argue for the need for the kinds of information that the original NAEP design was unable to provide. This led to the establishment of a national commission with the goal of evaluating NAEP and identifying its strengths and weaknesses. Its report (Wirtz & Lapointe, 1982) emphasized the limitations of the existing NAEP design with respect to the interpretability and utility of the results. Drawing on this report, researchers from ETS developed an approach that would enable a new NAEP to address a set of important questions that included the following:

- Are all students learning the skills and flexibility they will need for their own and society's well-being in the 1980s and beyond?
- Are all students being well prepared, and do they have similar opportunities to develop needed knowledge and skills?
- What are the relationships between school factors and student outcomes, especially those related to test-based achievement?

The ETS team proposed a novel design for the NAEP instruments and introduced a number of methodological innovations that changed the face of large-scale assessments. The proposed design was properly titled "A New Design for a New Era" (Messick et al., 1983). The innovations and the subsequent refinements of this design have been reviewed by Mazzeo et al. (2006). With respect to instrument design, the assessment frameworks became more directly tied to the underlying constructs (Kirsch, 2003), enhancing construct validity.[14] Among the methodological innovations were the use of balanced incomplete block (BIB) designs for the booklets administered to students (Mazzeo et al., 2006) and item response theory (IRT) in order to be able to report the results from multiple forms onto a common scale. Subsequently, ETS introduced latent regression models and Bayesian inference procedures to generate multiple imputations of the results in the form of "plausible values" that are now used to obtain population estimates and their corresponding measurement errors (Braun & von Davier, 2018; von Davier & Sinharay, 2014).

The novel NAEP design and methodology marked a key inflection point in the history of large-scale assessments. It set NAEP on a new trajectory that continues to this day. Moreover, these innovations were adapted by the different emerging ILSAs and, to this day, are at the core of the ongoing evolution of designs, methodologies, and platforms (see Mullis & Martin, 2019, for an extended case study). The coordination of these innovations and extensions led to an ever-richer body of information that ILSAs provided to policy makers, researchers, and other key stakeholders whose information needs have been evolving in tandem with changes in the social and economic landscape.

Indeed, starting in the 1990s, there was a growing appreciation of the significant connections between human capital and important outcomes both for individuals

and for the societies in which they live. In the United States, for example, the report by the Secretary's Commission on Achieving Necessary Skills (Kane et al., 1990) focused on identifying the skills needed for the country to enjoy a high-skills, high-wage, high-productivity economy. At the international level, the OECD (1992) noted that low literacy levels were a serious threat to economic performance and social cohesion. And the Definition and Selection of Competencies project provided a theoretical and conceptual foundation for a broad range of competencies that individuals would need to meet the changing demands of modern societies (Rychen & Salganik, 2001, 2003).

As a result, policy makers began to ask new questions focused on adult populations: How are educational attainment and cognitive skills related? How are literacy and numeracy skills related to health and well-being, as well as to participation and success in the labor force? What factors may contribute to the acquisition and decline of skills across age cohorts? How are literacy skills related to voting, trust in institutions, and other indices of social participation? (Kane et al.,1990; OECD, 1992). More recently, there have been organized efforts to develop taxonomies of 21st-century skills in an international context (e.g., Rychen & Salganik, 2001). Binkley et al. (2012) provided an extensive review. As noted earlier, this growing interest led to a series of international assessments focusing on adults, culminating in PIAAC.

Much like the 1983 design for NAEP that put large-scale assessments on a new trajectory, PIAAC marked the beginning of a significant cycle of innovation. As the first ILSA to be designed from the outset as a fully digitally-based assessment, PIAAC expanded what could be measured. It included, for example, technology-based tasks that more directly reflect the changing nature of how people access, use, and communicate information (Leu et al., 2021). In the workplace and in everyday life, it is increasingly important for adults to be able to navigate, critically analyze, and problem-solve in complex, data-intensive digital environments—and the PIAAC platform has made it possible to measure such skills. In addition, PIAAC introduced methodological innovations such as multistage, adaptive testing and more flexible routing for the background questionnaires that have improved the design and delivery of the survey, laying the foundations for future assessments (Kirsch et al., 2017). Other PIAAC innovations include the implementation of interactive stimuli and automated scoring of tasks administered in more than 30 languages, including character-based languages. Consequently, we regard PIAAC as marking another inflection point in the history of large-scale assessment surveys.

Countries now participating in international student and adult assessments represent the majority of the world's gross domestic product, with participation from low- and middle-income countries continuing to grow. That growth is fueled in large part by sponsoring organizations such as the World Bank, UNESCO, and the Inter-American Development Bank (Lockheed, 2013). These organizations regard international assessments as a cost-effective way to monitor the efficacy of their

educational investments in countries where national assessment systems, if they exist, are of variable and at times questionable quality. Their investments represent a growing recognition of the importance of monitoring learning and skill acquisition in support of economic and social development. This shift is also reflected in the United Nations' Sustainable Development Goals (specifically SDG 4 on quality education), which include learning targets and not just time spent in school as evidence of having achieved this goal.

There is a concomitant interest in enhancing the policy utility of ILSA data: Policy makers and other key stakeholders, including researchers representing a broad range of disciplines, are calling for these assessments to measure new and important cognitive domains to provide richer background and contextual information. The intent is to provide deeper understandings of how skills develop and how they relate to educational, social, and economic outcomes. These policy-driven questions provide the impetus for what we refer to as a virtuous spiral (Figure 20.2). Such questions lead to the formulation of new assessment frameworks that guide the development of new instruments. The desire to measure novel constructs such as collaborative problem-solving, as well as to expand the measurement of existing constructs (e.g., incorporating electronic texts in the literacy domain), drives further advances in assessment designs and statistical

FIGURE 20.2
The Virtuous Spiral

models that facilitate richer analyses and deeper interpretations of the data. These, in turn, elicit increased interest among a wider group of stakeholders, leading to further questions. The result is that ILSAs evolve along this spiral of increased relevance and policy utility.

THE MESSICK FRAMEWORK, EXTENDED

Inasmuch as the primary purpose of an ILSA is to provide credible evidence to inform policy makers and other stakeholders, Messick (1987) proposed a framework consisting of several design criteria that could offer guidance for the design, development, and implementation of large-scale assessments. These include comparability, interpretability, and relevance. An overall judgment of the potential for policy utility is then largely dependent on evaluating the strength of these criteria with due regard to purpose, context, and the population(s) of concern (see Kirsch & Braun, 2020, for an exposition of the original Messick framework).

Now, 35 years later, paper-based instruments have given way to digitally-based assessments that build on the past but draw on technological innovations in various domains. These include the use of digital platforms and new electronic tools, along with innovative workflows and processes. They also comprise advances in measurement science and improved methodologies to analyze complex data originating from multiple cultural and linguistic sources (more so than ever before). This point in time is an opportune moment to extend and refine Messick's three design criteria and to examine how they are impacted by technology, both directly and indirectly. This analysis leads to suggestions for innovations that could substantially enhance the policy utility of ILSAs.

Comparability refers to the degree to which the results obtained from a range of reporting groups, representing different cultural and linguistic contexts, have the same meaning in relation to the underlying constructs. In the context of an international assessment, achieving comparability is essential to policy utility. However, the demands are significantly greater than in a single national context because of the much greater heterogeneity that must be addressed. For example, sponsors and developers must make every effort to ensure that the samples of respondents in the different countries are approximately equivalent in the statistical quality of their representations of the corresponding target populations. The degree of success depends on such factors as the nature of the auxiliary information available for the survey design, the degree of cooperation among sampled units, and fidelity of implementation (Rust, 2014). These factors vary across countries, and weakness in any one of them undermines sample quality and, therefore, comparability. If approximate equivalence fails to hold, the country's results must be either reported separately or not at all.[15]

Further, the data generated by the cognitive instruments and background questionnaire must have equivalent meanings (measurement invariance) across cultures and languages (Dept et al., 2025; van de Vijver, 2018; von Davier & Sinharay, 2014). Accordingly, the survey instruments, including the background questionnaire, undergo an

iterative process of development, translation/adaptation, verification, and review (Ebbs & Wry, 2016; OECD, 2016, 2020). Following data collection, psychometric analyses reveal the extent to which the goal of measurement invariance has been achieved. If certain items appear to function differently in a particular country (or set of countries), appropriate adjustments are made to the psychometric models that are used to estimate item parameters and generate the reported scale scores for that country (Fishbein et al., 2020; OECD, 2016, 2020).[16]

Comparability also has a chronological dimension, namely, that cognitive scale scores, as well as composite scale scores derived from the background questionnaire, can be meaningfully related to the corresponding scores from earlier administrations. In the cognitive domain, this generally entails conducting scale linkage procedures that place new items from the current administration on the established scale. On occasion, a new scale is defined and the items from previous administrations are rescaled and placed on this new scale. In either case, the relevant procedures are well known, as are the checks on the validity of the linkage (Mazzeo & von Davier, 2014). With regard to the background questionnaire, the chief requirement is to retain the item set contributing to the composite scale or—if changes are absolutely necessary—to make only minimal alterations. Significant changes may disrupt the ability to continue an existing/earlier scale and require that a new scale be established. This problem arises because the number of items contributing to a scale is typically very small; hence, the loss of one or two items can have a substantial impact.

Interpretability depends, in large part, on the extent to which the instruments administered have been developed through a fully coherent process so that the reported scores can be given substantive or normative meanings that are credible, defensible, and accessible to a range of stakeholders.

The term *full coherence* signifies that each of the key operations of design, development, scoring, scaling, and reporting are not only appropriately linked to the intended measurement goals, but also functionally integrated through continuing collaboration among teams specializing in each of those operations. Thus, the term implies strong, evidence-based support for the desired interpretation(s) or, in other words, demonstrable construct validity (Messick, 1989).

Indeed, validation of the desired interpretations, whether of the cognitive scores or of the background scales, requires a thorough explication of the assumptions underlying the interpretations and an evaluation of the evidence supporting those assumptions (Kane, 2013). In this regard, Pepper (2020) argued that validity efforts in the ILSA context fall short of the guidelines in the 2014 testing standards (American Educational Research Association et al., 2014) and the requirements of the validity argument as articulated by Kane (2013). Pepper took as a case study the mathematics self-efficacy scales in PISA 2003 and PISA 2012, arguing, for example, that essential validity evidence that would be generated by (cognitive) response process analyses is entirely absent. Clearly, further work in this direction should become more of a priority for future assessments.

Relevance is the extent to which (a) the evidence elicited by the cognitive instruments and the background questionnaire is germane to current policy questions and decisions and (b) the assessment design yields results that can be analyzed in such a way as to address current priorities. Foundational to relevance is the designers' reliance on the assessment and questionnaire frameworks developed by international experts and reviewed by participating countries—with an eye on the most construct-relevant and policy-appropriate elements of each domain and construct that are measured. Of course, relevance is strongly dependent on both comparability and interpretability; that is, a deficiency in either one directly undermines the utility of the data in addressing the questions of interest. Judgments of relevance are made by the various stakeholders in each country, as well as by secondary analysts, and are contingent on the particular purposes at hand. The voluminous technical reports accompanying an ILSA provide useful descriptions of instruments and methods, as well as data displays to inform those judgments.

In addition to ensuring comparability and interpretability, ILSA designers adopt different strategies to enhance relevance. One example is modifying the operational definition of *legacy constructs* and extending the corresponding assessment frameworks. Often, new item types are introduced to target new or neglected facets of the construct. Recent reading literacy frameworks in PISA and PIAAC are good examples: The implementation of digitally-based assessments facilitates the introduction of electronic texts into the assessment. Such texts are an increasingly important source of obtaining and communicating information and, hence, deserve the emphasis they receive in the most recent assessment frameworks. For another example, TIMSS (which is now fully digital) has introduced problem-solving and inquiry tasks to simulate problems arising in laboratories and in the real world that require students to apply a combination of procedural skills and content knowledge. In the case of the background questionnaire, examples include the development of measures of certain aspects of learning contexts (or other background factors) that research indicates may be associated with the development of cognitive skills.

We believe that the improvements to be described here contribute to greater comparability, interpretability, and relevance, resulting in increased utility for policy makers and key stakeholders. Increased utility will continue to drive the salience and growth of these surveys that, in turn, will raise new questions, requiring further innovations on the part of developers and contractors. It is this productive dynamic that we have termed the *virtuous spiral*, depicted in Figure 20.2.

From an operational perspective, digital technologies provide the field with a powerful new infrastructure (platform) that both inspires and facilitates the development and refinement of new tools, processes, and workflows. Accordingly, the transition can impact all major phases of ILSAs: management, design, development and delivery, data handling, analysis, and data product generation. Although each of these phases exists with paper-based assessments, their development and implementation in

technology-based platforms now requires higher levels of coordination and integration to achieve the anticipated gains in efficiency and data quality.

TRANSITIONING TO DIGITALLY-BASED ASSESSMENTS

In a digitally-based, large-scale assessment, a well-designed platform makes it possible to introduce efficiencies that positively impact the development and conduct of the survey, including delivery, data capture, and processing operations. For example, the platform facilitates innovations in assessment design and delivery, allowing for embedded routing and branching capabilities that control how respondents move through the various components of the assessment. In particular, the implementation of multistage, adaptive testing models makes it possible to carry out more efficient measurement.

Platforms may also include a portal that functions as a centralized location for monitoring the workflows associated with key tasks related to the conduct of the survey, along with the interactions between contractors and each participating country. These enhancements serve to improve standardization of each stage of the work and allow for prompt interventions and adjustments, should they be necessary. They also can provide (a) item writers with capabilities for new item types and response modes to improve the measurement of legacy constructs or the introduction of new constructs and (b) translators with an improved infrastructure for translating and adapting assessment materials as well as receiving feedback from the verification process.

From the perspective of participating countries, the digital platform can not only support but also improve operational activities. The possibility of automatic data capture and scoring introduces efficiencies in terms of a reduced data entry and scoring burden along with more consistent scoring across countries and languages. Digital platforms also introduce new capabilities such as the capture of a full range of process data associated with respondent actions when interacting with assessment tasks, as well as accommodations that lead to greater accessibility for students and adults participating in the surveys. A full description of the potential efficiencies of a digital platform is beyond the scope of this chapter. This section, however, discusses the platform's role in a subset of key activities. Collectively, these enhancements can have significant impacts on the relevance, comparability, and interpretability of an ILSA, thereby increasing its overall utility to policy makers and key stakeholders.

Survey Instrument Development

The platform must possess the functionality to support item development and authoring for both the cognitive instruments and the background questionnaire in multiple languages and orthographic systems. Consequently, platform developers engage at the outset with domain experts and instrument developers to understand

the technical implications of (a) measuring new constructs, (b) revised and extended frameworks that broaden what can be measured, (c) delivering new item types, and (d) the approaches required to address a greater range of respondents' proficiencies due to the expansion in participating countries. At the same time, domain experts and instrument developers take into account the range of displays, item types, and interaction modes that the platform supports, resulting in some modifications of item types and/or response modalities. This collaboration often results in timely platform development and enhancements that contribute to the relevance and interpretability of what is being assessed.

Assessment designs for paper-based surveys are limited by operational constraints related to time and costs associated with assembling, checking, printing, shipping, and handling multiple paper forms. By contrast, digitally-based assessments, especially when combined with some form of adaptive testing (see the section "Survey Administration"), offer many advantages. First is the potential to eliminate errors associated with the assembly and distribution of multiple paper booklets both within and across languages. Second, the digital platform can accommodate designs involving complex routing algorithms that entail a very large number of virtual forms. In particular, it can assign an appropriate, predesigned form to each participant, precisely following the assessment design (a practice that is particularly germane to adaptive testing). Moreover, the platform can support both instrument administration in different formats and data capture for a range of response modalities. Finally, greater efficiency in the estimation of proficiency distributions, in combination with new item types, makes possible both broader construct coverage of existing domains and the assessment of new constructs, topics we address in the sections "Developing Coherent Cognitive Assessments" and "Background Questionnaires."

Developing Coherent Cognitive Assessments

Following a coherent process for assessment development enhances the interpretability of the findings. In this regard, a construct-centered or evidence-centered approach to assessment development is most helpful (Messick, 1994; Mislevy et al., 1999). Successful implementation yields a consensus on an operational definition of the target construct; thus, it provides assessment developers with a road map for the design and development of the tasks, as well as for the collection of evidence that can be used to represent performance in relation to the construct. This process yields a reporting scale that can be more appropriately interpreted.

The assessment framework for each domain contains specifications for item development: (a) the identification of key task characteristics to be varied singly and jointly, (b) the numbers and types of items (stimulus materials, response formats, and levels of difficulty) required to populate the item pool, (c) guidelines for instrument assembly, and (d) considerations related to scoring open-ended items (see Lennon & Kirsch, 2025, for an example). In advance of field testing, item quality is evaluated in terms of the items' links to the assessment framework and a judgment that

their formats, layouts, and content follow established design principles. Subsequent to the main assessment, as well as the scaling and analysis of the cognitive data, the assessment framework facilitates the creation of item maps and the validation of task characteristics. These are often used to characterize proficiencies at different points along each cognitive scale.

In an international context, development of cognitive assessments that yield scores that are valid and comparable requires that every effort be made to establish the items' cross-cultural, cross-linguistic, and cross-national validity. In that regard, the digital platform should facilitate (a) flexible content management, so that development files can be shared as needed; (b) item previewing, so that items can be examined in both source and target languages, along with the corresponding item layouts; and (c) consistency in item translation/adaptation across the item pool. Thus, the platform must support the active involvement of participating countries, with continuous coordination among geographically dispersed teams, comprising domain experts, test developers, psychometricians, platform developers, and graphic designers.

Background Questionnaires

ILSAs also yield information on a set of constructs that are measured through the background questionnaires. In addition to standard demographic information, the constructs targeted by the background questionnaires are based on frameworks developed through collaborative efforts among representatives of participating countries, international content experts, and questionnaire developers. One goal is to assess important characteristics or factors associated with students, teachers, and schools for surveys focused on in-school populations and important social, educational, and labor market factors for adult assessments. In addition to these context-related constructs, school-based surveys also target students' attitudes and behaviors along with family-related characteristics. Although these constructs are of interest in their own right, they are particularly valuable because of the insights that can be gleaned from their relationships with the cognitive outcomes, as well as the differences in those relationships across populations and through time.

Since the mid-1990s, background questionnaires have become more comprehensive in scope and their development more systematic. For example, rather than focusing on authoring individual items, the trend has been to develop sets of items whose responses can be combined into a scale that relates directly to a target construct, placing a new burden on item developers and the committees that advise them. Scale construction by means of IRT modeling is now accepted practice, with due attention to model fit, reliability, and validity (Martin et al., 2014).

As stakeholders have come to appreciate the value of the background questionnaire, it has gained in importance, thereby putting more pressure on improving the assessment of existing constructs while accommodating new indicators—all within strict time constraints for administration. The result is an ongoing tension between introducing these new indicators and maintaining trend for (at least some) existing indicators. This

tension provides the impetus for methodological developments, such as incomplete designs analogous to those employed for the cognitive instruments. PISA 2022 implemented a within-construct rotation design for the background questionnaire. However, implementing incomplete designs for the background questionnaire makes strong assumptions about conditional independence of context variables and has implications for the scaling of the cognitive domains that have not yet been fully resolved (von Davier, 2014).

Nonetheless, this dynamic provides yet another instance of the virtuous spiral introduced in Figure 20.2. As with the cognitive instruments, the platform plays an important role in supporting more complex designs and routing patterns often associated with current background questionnaires. These improvements support the relevance of the information that is gathered, as well as the quality and consistency of the information that is captured.

Translation/Adaptation/Verification of the Instruments

An essential step in establishing cross-national comparability is the translation, adaptation, and verification of the items in the cognitive instruments and the background questionnaire. This process is difficult, costly, and time-consuming. By supporting new workflows, the digital platform contributes to making the process both more accurate and efficient. As noted earlier, the platform accommodates the full range of languages used by participating countries (including right-to-left and ideographic languages) in addition to supporting the coordination of the work of test developers, linguists, and cross-cultural survey methodologists. This team produces informative translation and adaptation notes that both explain the underlying constructs and offer guidelines for use during the test translation and adaptation process. These improvements in this work process help to improve quality control, thus facilitating standardization and resulting in improved comparability.

In the case of PIAAC, for example, the item-by-item translation and adaptation guidelines (a) explain what the item is intended to measure; (b) specify which adaptations are mandatory, desirable, acceptable, or ruled out; (c) draw the translators' attention to terminology problems, translation traps, and patterns in response options; and (d), in the case of recurring elements or elements already present in trend materials, indicate how to access previous translations of these elements.[17]

For the cognitive items, the guidelines provide information on certain crucial assessment-related features such as literal matches (e.g., between stimuli and questions) that need to be maintained in the translated national versions, level of language difficulty, distractors, and so on.

A specially developed set of integrated tools in the platform makes it possible for these guidelines to appear in the translation tool when a translator on the national team (or reconciler, or verifier) processes a text segment. This is a technical innovation offering significant added value by streamlining processes, reducing the number of documents and tools required, and providing a translation environment that unites

all relevant information—thus enabling translators to better attend to key elements of the translation task at no additional cost for countries. We expect that forthcoming artificial intelligence–based tools will offer further enhancements and efficiencies to the translation/adaptation process.

Population Sampling Operations

The primary goal of an ILSA is to obtain credible and comparable estimates of population distributions of proficiencies in cognitive domains. The essential first step in sampling is a clear definition of the target population. Then a probability (random) sample of units is selected from the population, because randomization justifies using the machinery of probability theory to make inferences from sample characteristics to population characteristics. An added benefit is that with a well-conducted survey it is possible not only to obtain approximately unbiased estimates of population parameters (e.g., means, variances, percentiles), but also to quantify the uncertainty attached to those estimates (e.g., estimates of the standard errors of the parameter estimates).

Because of considerations related to cost and logistics, as well as reporting requirements, large-scale surveys almost never draw simple random samples of the units of interest (students or adults); rather, they employ complex sampling designs such as multistage cluster designs.[18] An important design consideration is the trade-off between the amount of information derived from a sampling unit and the cost of obtaining that information. The key is that, at least in principle, there is a known probability used for the selection of each unit. These probabilities are used in the calculations leading to the estimates of population parameters and their variances.

The principles and techniques of population sampling are treated in many texts (e.g., Lohr, 2010). Rust (2014) presented a comprehensive treatment of considerations in ILSA sampling. TIMSS and PIRLS sample schools and students in such a way as to be able to estimate proficiency distributions for particular grades. Consequently, they typically employ multistage cluster samples where schools are first grouped into relatively homogeneous categories (termed *strata*) characterized by such factors as geographic region, political status, and school type, as well as characteristics related to mean school achievement. At the country level, the choice of characteristics to classify schools for the purpose of selection also depends on the information available, the interest in the reporting accuracy of the results at various subnational levels, and considerations of efficiency. Within strata, schools are selected with probability proportional to size.[19] The clusters are randomly chosen classrooms within the schools providing instruction in one or both of the target grades. The assessment instrument is administered to all students within the selected classrooms. This approach is efficient, is minimally disruptive to the school, and facilitates estimating relationships between student achievement and their teacher or classroom characteristics. The final sampling plan represents the ideal plan adapted to the real-world constraints particular to each country.[20] ILSAs recognize that achieving a sampling plan with 100% coverage is not realistic. Typically, countries are not annotated if they have no more than 5% exclusions. At the school level, reasons

for exclusions include schools located in difficult-to-reach regions or particularly small schools; at the student level, reasons may be nonofficial language speakers and students with disabilities.

By contrast, PISA estimates proficiency distributions for a particular age cohort (i.e., 15-year-olds). Accordingly, its design must take into account not only that students of the same age may be enrolled in different grades, but also that those grades may be located in different system levels (i.e., lower secondary and upper secondary). Consequently, there may be curricular differences that, along with other factors (e.g., differential participation and coverage), can impact cross-national interpretability of the results.

The sample design for PISA 2018 also utilized stratification for school selection. As with the IEA studies, the choice of stratification factors depended on the country. Within each selected school, a random sample of students is obtained from the student population in the targeted age cohort. The number of students selected within each school varies from 35 to 42 depending on the assessment options chosen by the country (see OECD, 2020, chap. 4, for further information).[21]

PIAAC is a household survey. For some countries it employs a type of multistage, clustered area sampling with households as the ultimate sampling unit. Within a selected household, the instrument is administered to a randomly chosen adult whose age falls within the target range of 16 to 65. In other countries with registries of either households or individuals, sampling is conducted directly from the registry. Consequently, the actual implementation of the sampling procedure varies by country, depending on the structure and completeness of the sampling frame. The type of procedure employed also affects variance calculations (see OECD, 2016, for more detailed information).

As is the case with all fieldwork, the obtained sample in a school-based survey differs from the ideal sample in a number of ways. An especially challenging problem is the existence of schools that are not included in the sampling frame. For those schools in the sampling frame, selected schools may refuse to cooperate and selected students within schools may be absent, refuse to cooperate, or not fully respond to the cognitive instrument and/or the background questionnaire.[22] In the case of household surveys, the sampling frame or the registry may be inaccurate or incomplete. Once households are selected and visited, there may be no one at home or the selected respondent may refuse to participate or only partially respond to the survey.

Inasmuch as the utility of cross-national surveys is critically dependent on the quality of the samples drawn, each ILSA has a set of procedures to address sample quality issues. These procedures are documented in the technical reports referenced above (see Rust, 2014, for general considerations about sample designs and procedures). Procedures may include replacement sampling for noncooperating units and technical adjustments for various types of nonresponse. Notwithstanding the use of such procedures, substantial differences across countries in response rates at the various stages of design implementation reduce the credibility of comparisons of proficiency distribution estimates as well as other targets of inference. ILSAs typically set thresholds for response rates; results for countries not meeting the threshold are reported separately

or not at all. Nonresponse bias analysis is also conducted to estimate the effects of these issues on the overall estimates.

When information about the population of interest is lacking or incomplete, ILSA teams face significant challenges in ensuring sample quality within each participating country.[23] This may be a particular concern with household surveys. Because many of these countries have limited experience and capacity in conducting large-scale surveys, they require both extra support and more careful monitoring. In some countries, information to develop adequate sampling plans for national samples may be insufficient. At the same time, ILSA participation offers country representatives intensive training and the opportunity to garner field experience in the selection of probability samples that can be employed for later in-country surveys as well as future ILSA administrations.

ILSA teams are always seeking ways to reduce survey error. A model for improving data quality is the total survey error (TSE) framework, originally developed by Hansen and explicated by Hansen et al. (1953). The TSE framework covers all types of errors that may arise in survey design, sample selection, data collection and processing, scaling and analysis, and creation of data products. The TSE framework makes a distinction between sampling and nonsampling errors. Sampling error results from variability in the estimates stemming from the selection of a random fraction of the target population. Nonsampling errors may be introduced at any phase of the survey process. Consequently, they must be taken into consideration throughout survey operations and management (including development and implementation) as well as data collection, handling, and analysis. Operationalizing the TSE framework requires striking a balance between enhancing data quality and operating within survey constraints (Biemer et al., 2017).

The transition from paper-based to digitally-based interviews and assessments has enabled many advancements in error control. One strategy is to establish processes that detect various sources of nonsampling error during data collection and to remedy them when possible. In this regard, the digital platform plays a key role through its capacity to collect and present what is referred to as *para-data* in a timely manner (Mohadjer & Edwards, 2018).

Para-data are the survey process data that are generated during data collection. For example, in carrying out the PIAAC household survey, the para-data may include the record of contact information, instrument timings, voice recording of interviews, geolocation of the interview, and interviewer work activities (hours and travel routes). They contribute to indicators of data quality, costs, and interviewer effectiveness. Because para-data can be quite voluminous, they are collected and summarized in a performance dashboard comprising a set of survey control charts and data graphs that monitor sample yield by interviewer and overall response rates at different levels of aggregation, highlighting unusual outcomes. With real-time transmission of information, the performance dashboard allows survey managers to react to operational challenges in a timely manner, facilitating what is sometimes termed an adaptive data collection strategy.

Survey Administration

In this section we review two modes of administration—paper and pencil and digitally based.

Paper-and-Pencil Administration

In preparation for this mode of administration, the item pool for a cognitive domain is organized into disjoint sets of items, termed *blocks* or *clusters*. For each ILSA, blocks are created according to specific criteria related to the number of items or item sets, content coverage, distribution of item difficulties, and timing. In a matrix sampling design, students are presented with a booklet that comprises two or more blocks, along with a background questionnaire. For example, in the case of booklets containing two cognitive blocks, the blocks are systematically paired into booklets. They are usually organized according to a BIB or partially balanced incomplete block (pBIB) design (Mazzeo et al., 2006; L. Rutkowski et al., 2014). In a full BIB design, each block is paired once with every other block.[24] This ensures that it is possible to compute covariances for all pairs of items, thereby facilitating placing estimated item parameters on a common scale (see the section "Scaling, Population Modeling, and Proficiency Estimation") and using statistical analysis techniques that require the use of a complete covariance matrix, such as factor analytic techniques. Because of logistical constraints, it is not always possible to employ a full BIB design and designers resort to using pBIB designs, in which not every pair of blocks appears together in a booklet. Nonetheless, with modern psychometric theory, it is still possible to place all estimated item parameters on a common scale.

As one might expect, there are many variations on a theme. In a focused BIB design, each student receives blocks assessing a single subject (e.g., reading in PIRLS). In an unfocused BIB, students are assessed in two or more subjects (e.g., mathematics and science in TIMSS; reading, mathematics, and science in PISA). In the latter case, it is possible to compute correlations between cognitive domains. Of course, the intended interpretations of the results of the psychometric analyses depend on the assumption that the responses to each item (block) have been produced by randomly selected samples of students.[25] This assumption is met by various spiraling strategies for booklet distribution (Mazzeo et al., 2006). Detailed descriptions for each ILSA can be found in the corresponding technical reports.

To accommodate the increasing heterogeneity in country-level distributions of proficiency, ILSAs have adopted different strategies. For example, in PIRLS 2021, the proportional distribution of booklets by difficulty varies by country. Booklets are categorized as more or less difficult, depending on the distributions of item difficulties in the component blocks. Countries with (assumed) higher levels of proficiency receive a greater proportion of the more difficult booklets, while countries with (assumed) lower levels of proficiency receive a greater proportion of the less difficult booklets. By orchestrating a better overall match between the assessment and the country's students, it is possible to obtain more accurate estimates of the overall proficiency distribution within the constraints imposed by paper-based administration.

Digitally-Based Administration

As noted earlier, the transition to digitally-based administration facilitates the introduction of new item types as well as new strategies for administration. With regard to the latter, it is certainly possible to organize the sets of items for administration in predetermined sequences of approximately equal difficulty (as is done in paper-and-pencil administration) and assign each set to randomly equivalent samples of respondents. Of greater interest is the introduction of adaptive testing. Adaptation can be done at the level of the individual item or of sets of items.[26] The latter is usually referred to as multistage, adaptive testing.

In multistage, adaptive testing, after a set of items has been administered and responses have been evaluated, an algorithm implements a decision rule that, for the next stage, routes the respondent to one of a number of possible item sets. The algorithm accounts for the student's performance on the previous item sets as well as content and other constraints. The intent is to improve the final estimation accuracy by better matching item difficulties to the current estimate of the respondent's proficiency, which is especially important when dealing with substantial heterogeneity in proficiencies within and among countries. In particular, adaptive testing reduces the chances that students will be exposed to items that they find extremely easy or extremely difficult.

Happily, it appears that adaptive testing induces greater respondent engagement, as indicated by lower rates of nonresponse or random responding. In conjunction with timing data that enable making distinctions between omitted items and items not reached, greater accuracy in proficiency estimates is achieved. At present, one difficulty is that in some ILSAs many item sets include one or more items that require human scoring. In that case, the decision rule must rely solely on responses to the machine-scored items. As a consequence, the selection of the next set of items is based only in the performance on these machine-scored items. This results in somewhat less improvement in efficiency than would be the case if the responses to all items were used.

With the large item pools typical in ILSAs, the number of item sets (or testlets) is correspondingly large, and consequently for multistage, adaptive testing, the number of different item paths (sometimes referred to as *virtual forms*) can run into the many thousands. Clearly, having such a large number of forms in a paper-and-pencil administration would be both infeasible and unaffordable. At the same time, the key innovation in matrix sampling has been retained, namely, that each respondent is only administered a small proportion of the full item pool.

Adaptive testing in an international context was first carried out in an individual-level test of adult literacy, setting the stage for later, more advanced implementations—first in PIAAC and subsequently in PISA. Yamamoto et al. (2018) discussed some advantages of multistage, adaptive testing over fixed form tests, especially in the ILSA context. They also note a number of challenges including (a) avoiding (to the extent possible) items that require human scoring in the decision-making process,[27] (b) ensuring that each virtual form meets construct representation requirements, (c) maintaining desired levels of item exposure control, and (d) employing appropriate analytic procedures. Indeed, implementing multistage, adaptive testing adds considerably to the complexity of the

assessment system. Yamamoto et al. provided a comprehensive discussion of these complexities and their resolution in the context of PIAAC and PISA.

Accommodations for Testing

Comparability of national samples is an essential element of ILSA utility. One potential threat to comparability is related to the assessment of individuals with disabilities, inasmuch as countries' practices differ in a number of ways: how they identify students with different types and degrees of disability, the nature of the accommodations typically employed during testing, and their protocols for exclusion from an ILSA. Attempting to harmonize these practices across participating countries is evidently infeasible.

On the one hand, the baseline strategy adopted by TIMSS and PIRLS is to review the accommodations proposed by the countries and to approve them, unless there is a clear threat to the validity of the assessment. For example, reading questions to a blind student may be acceptable in TIMSS but not in PIRLS. Clearly, allowing variation in accommodations represents a different level of standardization of assessment administration. On the other hand, the purpose of test accommodations is to allow students to demonstrate what they know and can do, but with only minimal changes to the target construct (Thurlow, 2014; see also Zwick and Rodriguez & Thurlow, both in this volume). To the extent that this is the case, and if students are afforded accommodations with which they are familiar, then a certain level of comparability is achieved—in the sense that each student has a fair opportunity to meet the challenges presented by the assessment. From an international perspective, this protocol prioritizes obtaining better estimates at the national level (for a country that includes these individuals in their target population) at the expense of some loss in cross-country comparability in estimates of proficiency distributions.

PISA does not employ accommodations in the main assessment, with two exceptions: (a) allowing for extra time and (b) administering a shortened (1-hour) version of the full assessment with a restricted set of response formats. Schools designate which of their selected students are eligible for one or the other of these accommodations, which are available for math and reading only. For PISA 2025, all items in the shortened assessment will be fully accessible, though not all item types will be represented in the assessment. The data from these items do not contribute to the estimation of the international IRT parameters, but they do so in the estimation of the plausible values (PVs) for the national distributions.

With the advent of digitally-based assessments, new policies and protocols are under development. For example, the OECD sponsored a small pilot study to investigate the feasibility of offering accommodations for students with a range of disabilities (Laitusis et al., 2018). The results were mixed with respect to practicality and efficacy, especially for more complex item types such as simulations. In any case, the report's findings and recommendations will inform subsequent phases of implementation research for PISA 2025, as well as for TIMSS and PIRLS as they transition to digitally-based assessments. It is expected that future administrations will explore the feasibility of implementing affordances such as varying font size and text to speech.

Overall, accessibility will be enhanced as new items are developed according to the principles of universal design and added to the item pools. At the same time, this approach is more feasible for some item types than for others. For the latter, special modifications will be required to enhance accessibility. Because of cost constraints, the number of such modified items will be limited; hence, they may be located in a small number of forms that can be administered when particular accommodations are requested.

In this regard, one other aspect of ILSA administration in school settings bears mentioning. The number of different devices that students are using day to day is growing. Current planning contemplates allowing students to bring their own devices, as long as they meet certain interoperability standards, as well as building a platform that will host those devices, provided they conform to a number of requirements. Carrying out this plan will be quite challenging. It remains to be seen to what extent the profusion of devices with different capabilities introduces construct-irrelevant variance and impacts score trends.

SCALING, POPULATION MODELING, AND PROFICIENCY ESTIMATION

IRT scaling, population modeling, and proficiency estimation are essential components of the ILSA workflow. The accuracy of the model parameter estimates, as well as the estimates of the precision of those estimates, constitutes the statistical foundations for generating the appropriate, individual-level proficiency distributions that are ultimately transformed into population-level results. Accurate estimates are essential to achieving comparability across all relevant reporting groups and, hence, valid interpretations of the results.

Introduction

It is customary to provide countries with standard, mandatory software that is used to manage, integrate, and validate their national data. Participating countries employ the software for data processing as well as to run checks to ensure, as much as possible, that the within-country data capture and integration accurately reflect the values given by the sampled persons and/or the interviewers.

In particular, the software is able to produce a series of reports that give an overview of the data quality and field operations progress. Each country is required to generate and review these reports during field operations and again prior to submitting the database to the contractor for international processing. Along with the national database submission, each participating country submits supporting data documentation that provides the ILSA team with detailed information regarding issues or technical difficulties in the administration of the assessment.

For each country, the input to this segment of the workflow is a set of validated data files. The output is a set of files containing, among other things, the estimated parameters

of the score distributions for the overall population and for designated subgroups. By this point in ILSA history, the statistical and psychometric procedures have evolved considerably in complexity from those employed in the early years of NAEP. Paraphrasing von Davier and Sinharay (2014), current methods may be viewed as sophisticated imputation approaches that combine the several advantages of IRT with an explanatory approach based on collateral information to produce accurate subgroup results—despite the relatively short testing time and the sparseness of the data. The use of collateral information is especially important when participants are asked to respond to multiple domains within the same limited time. Further details can be found in von Davier and Sinharay (2014), as well as in the technical reports published by each ILSA. Analysis teams also expend considerable effort in checking the intermediate results at each stage of the process to ensure accuracy and comparability of the results while accounting for possible country-by-item interactions. Moreover, these procedures are continually refined to improve efficiency and to accommodate new demands from sponsors and participating countries.

For simplicity, the discussion that follows assumes a single latent trait for the domain of interest.[28] With the primary goal of producing unbiased estimates of score distributions for populations and subpopulations, ILSA assessment designers are faced with the formidable challenge of meeting the requirement of full construct representation under the constraint of limited testing time. For paper-and-pencil administration, various types of matrix sampling are employed (see the section "Survey Administration"). Such plans are characterized as "missing by design," and the use of matrix sampling has implications for subsequent analyses. In practice, a limited number of forms are developed and the design can be represented by a matrix that displays the fraction of the full item pool that is contained in each form. Forms are distributed in such a way that each form is administered to random sets of respondents of approximately equal size.

For digitally-based assessments and, especially, with the advent of adaptive testing algorithms, the matrix designs have become increasingly complex. As is the case with paper-and-pencil administration, respondents respond to a relatively small fraction of the total item pool. However, on average, the forms administered to a group of respondents will be better matched to the group's proficiency distribution. The desired group-level estimates can be obtained through the use of advanced statistical methods—assuming that each item or item set has been administered to a proper probability sample of respondents, that the sample sizes are sufficiently large, and that there are satisfactory linkages across forms.

The analysis proceeds in four stages:

- Stage 1, item calibration: estimation of the parameters of the item response model
- Stage 2, population modeling, including latent regression analysis
- Stage 3, generation of plausible values
- Stage 4, linking plausible values to the established reporting scale

Item Calibration

All ILSAs make use of IRT to calibrate the items in the item pool. Calibration involves estimating the item parameters that characterize the probabilistic relationship between test-taker proficiency (i.e., their location on the scale corresponding to the latent trait) and performance on the item. The items may be scored dichotomously (two possible scores) or polytomously (three or more possible scores). The psychometric models employed differ across ILSAs: TIMSS and PIRLS use the three-parameter logistic IRT model for multiple-choice items, while PISA uses the two-parameter logistic IRT model. The general partial credit model is used by all ILSAs for calibrating all other items. (For a discussion of the models used in IEA assessments, consult von Davier et al., 2020).

Field trial instruments incorporate designed collections of new items or new item sets, along with a set of previously administered items that will be used to estimate trends. The total pool of items is administered to large convenience samples to estimate the measurement invariance of the trend items, as well as to obtain preliminary item parameter estimates for the new items. These analyses are used to furnish initial estimates of how well the trend items and new items function, both within and across participating countries with regard to comparability and overall quality.

If the items are selected for operational use, then the response data obtained from the main data collection are used to update the item parameter estimates for the new items and to evaluate the degree of comparability across countries. Operationally, if an item appears to have substantially different psychometric properties for a particular country, it is calibrated separately for that country.[29]

With regard to the background questionnaire, item sets addressing constructs such as attitudes, dispositions, and behaviors are used to generate scales for secondary analysis. The process is also carried out by employing IRT models. The models take account of the fact that most of these items use ordered response categories. Both PIRLS 2016 and TIMSS 2019 have used the Rasch partial credit model to conduct item calibration employing the full data sample, with each country contributing equally to the analysis. Subsequently, the reliability of the derived scales is evaluated and the scales are subjected to validation procedures, such as tests of unidimensionality.

For both the cognitive instruments and the background questionnaire, the field tests and main administrations are essential to the maintenance and replenishment of the item pools. The transition to digitally-based assessments, particularly with the introduction of multistage, adaptive testing, necessitates an increase in the sizes of item pools for the cognitive domains, with the demands for piloting and calibrating new items becoming correspondingly greater. Increases in the item pools are also driven by the need to accommodate greater heterogeneity among participants with respect to proficiencies in the focal domains. Further, with the concomitant exponential increase in the number of (virtual) forms, more complex assessment designs are needed to obtain sufficient information for calibration and evaluation of differential item functioning (Yamamoto et al., 2018). Such designs, made feasible by the

digitally-based assessment platform, evidently contribute to improved construct representation.

It bears mentioning that both PIRLS and TIMSS use country-level adaptive designs (and sometimes benchmarking population adaptive). Such designs employ rotations of harder/easier booklets with the proportions based on prior information from a previous cycle. The goal is to improve estimation, particularly at the low end of the proficiency scale.[30]

Population Modeling

The background questionnaire incorporates a set of questions related to constructs believed to be associated with (cognitive) proficiencies. These include both background characteristics and contextual factors. The second stage of analysis makes use of these additional data to obtain more accurate estimates of proficiency distributions. Specifically, this stage involves estimating the parameters of a so-called latent regression model (Mislevy, 1991; Mislevy et al., 1992; von Davier, Khorramdel, et al., 2019; von Davier & Sinharay, 2014) in which an individual's unobserved proficiency is regressed on a suite of variables derived from their responses to the background questionnaire. This model is usually referred to as a *population model*. In practice, the full suite of background and contextual variables is replaced by a smaller set of principal components that account for approximately 70% to 80% of the observed variance.[31] The reduction in the number of explanatory variables yields more stable estimates of the regression coefficients and the variance–covariance matrix of the model, but with negligible reduction in accuracy (Oranje & Ye, 2014; Thomas, 2002; Wetzel et al., 2015). It is also customary to assume that the residuals about the regression plane follow an approximately normal distribution. The entire system is placed in a Bayesian framework with the introduction of a unit normal prior for each individual. The prior is successively updated with each full cycle of the expectation–maximization algorithm that is employed in the estimation procedure.[32]

Concatenating the IRT model and latent regression model yields the marginal probability distribution for the observed item responses, given the transformed set of background variables, the estimated item parameters, and the parameters of the latent regression. Multiplying the marginal distributions across respondents results in a likelihood function. The corresponding log-likelihood function is then maximized with respect to the parameters of the latent regression, while the parameters of the IRT model are held fixed at the parameter values obtained in the first stage of the analysis (to impose comparability constraints across participating countries).[33]

We note in passing that maximization of the log-likelihood function is a very difficult exercise in numerical analysis and considerable effort has been expended in devising different approaches to carrying out the computations (von Davier & Sinharay, 2014). The output is an estimate of the conditional distribution of the maximum likelihood estimate (MLE) of the vector of regression coefficients, as well as the MLE of the variance of the distribution of an individual latent trait. The latter is held fixed at its estimated value in all further computations.

Plausible Values

As noted earlier, the methodology employed by large-scale assessments is not intended to yield point estimates of ability at the individual level but, rather, estimates of group-level distributions of ability. This task is accomplished by producing a collection of parameter estimates that are used to generate a set of imputed values for each individual. In this context, these imputed values are PVs. Collectively, the PVs associated with an individual make manifest the available information regarding their proficiency in the domain. The average of the PVs is an estimate of the expected score were someone with their vector of principal components to be administered the full item pool. The variability among the PV is used to estimate the total measurement uncertainty associated with an estimate of a group-level parameter.

To obtain a single set of PVs (one for each respondent in the sample), it is necessary to approximate the proficiency distribution for each respondent. Since that distribution is assumed to be normal, it is only necessary to estimate its mean and variance. An estimate of the variance was obtained in the previous stage. The mean depends on the covariate values associated with the respondent and the (estimated) vector of regression coefficients of the latent regression model. That estimated vector has a posterior distribution that is approximately multivariate normal, with mean equal to its MLE and a variance–covariance matrix equal to the estimated variance–covariance of the MLE.

A random draw from that posterior distribution produces a vector of regression coefficients, enabling full specification of the proficiency distribution for each respondent. A single independent, random draw from that proficiency distribution yields 1 PV for the respondent. These steps are carried out for each respondent to generate a full set of PVs for the sample. The entire process is then repeated N times to produce N full sets of PVs. The number N of imputations was decided by most ILSAs following recommendations by Little and Rubin (1987) for multiple imputations. Early choices of 5 or 10 were limited by computing power, data storage, and time constraints. As computing power increases, 20 or more sets of imputations will be common.

Linking

An important use of ILSA data is tracking changes in skill distributions over time at the national and subnational levels. Such comparability over time requires linking the proficiency scale of the current assessment to the scale established in prior assessments. Linking is accomplished by embedding so-called trend items in the survey instrument. Trend items are those items employed in the current assessment that were also present in earlier assessments. Roughly speaking, by comparing performance across administrations on these items, appropriate scale transformations can be estimated and applied. However, actually carrying out the procedure involves a range of technical issues related to both design and analysis (Mazzeo & von Davier, 2014). Note that trend items contribute to proficiency estimation just as nontrend items do. In this regard, a PISA study (von Davier, Yamamoto, et al., 2019) deserves mention. In this study, PISA data from 2000 to 2012 (comprising more than 2 million students) were reanalyzed in a single

linking model to better understand aspects of linking, as well as issues related to model fit. One consequence was improved procedures that maximized the use of trend information over more than two cycles.

Although every ILSA adopts a somewhat different approach to scale maintenance, each faces similar challenges. In particular, designers must consider such factors as the degree of construct representation exhibited by the trend items, the number of trend items, their colocation in blocks, and block placement. Further information can be found in Mazzeo and von Davier (2014) and in the technical reports associated with each ILSA.

Ideally, successive assessments (and the corresponding item pools) are constructed according to the specifications of a common assessment framework. The set of trend items should be both fully representative of that framework and sufficiently large to support stable estimates of the linear regression parameters that transform the current scale into the earlier one. Consequently, it is not unusual for the trend items in paper-and-pencil administration to constitute approximately 50% of the item pool.

For trend items to function as intended, and to minimize the (unwanted) contributions of sources of variation unrelated to the focal domain, the context for each item is held constant to the extent possible. In the case of individual items, their location within a block, as well as the psychometric characteristics and speededness of the other items in the block, should be controlled. An alternative strategy is to employ intact item sets or item blocks from earlier assessments. Although this controls the local context, it may make full construct representation more difficult to achieve. In paper-and-pencil administration, block position within a booklet must also be considered. Further complications arise when considering the different designs employed for administering a single content domain to a student (e.g., PIRLS) or for administering two or more content domains (e.g., TIMSS, PISA). For example, designs for the latter situation possess some advantages from a measurement perspective because they make it possible to take advantage of the correlation of proficiencies between domains. However, they make keeping the within-instrument context for trend items constant more difficult because it is now necessary to also consider the paired domain as part of the context.

Finally, the human scoring of constructed response items must be strictly comparable across administrations. Accordingly, the training of scorers for the current administration must replicate the training conducted in the previous administration. Close monitoring of scorer behavior is essential. One approach is to seed responses with known scores from previous administrations into the current workflow and to compare the scores assigned at the two time points. Systematic discrepancies trigger retraining and rescoring and, occasionally, eliminating the item for the discrepant group.

As is always the case with IRT, there is a fundamental indeterminacy in fixing the latent trait scale after calibration. Either by comparing the item statistics on the trend items in adjacent administrations or by concurrently recalibrating the items in the two administrations, it is possible to resolve that indeterminacy in the newer administration

so that item parameters are placed on the scale established in the earlier administration. A linear transformation then places the performances in the current administration on the reporting scale established at the outset of the ILSA. This process enables valid comparisons of score distributions across administrations.

Notably, there is an underlying tension between scale maintenance and the desire to expand the assessment framework and/or introduce new item types. Similar issues of scale maintenance arise when there is interest in making comparisons across cycles for those scales developed from the background questionnaire. For further details, as well as descriptions of the linking strategies employed by the paper-and-pencil administration associated with the different ILSAs, see Mazzeo and von Davier (2014). It is also typical that each ILSA administration poses unique challenges. For example, TIMSS 2019 saw the advent of eTIMSS, which was administered in about half of participating countries. This new administration necessitated a "bridge study" to link eTIMSS to the standard paper version of TIMSS. Further, a number of less difficult blocks of fourth-grade math items were also introduced with the goal of improving measurement accuracy for lower performing countries. Finally, eTIMSS included items (problem-solving and inquiry tasks) that took advantage of the affordances of digital administration but had no counterparts in the paper version. Carrying out the procedures of scaling and linking in this setting was considerably more complicated than in past administrations. For details, consult Martin et al. (2020, chaps. 11 and 12).

Although all ILSAs are transitioning from paper-and-pencil administration to digitally-based assessments, paper-and-pencil administration versions are likely to be required for some time to come: For some countries, the technical infrastructure is not sufficient to support digitally-based assessments, while in other countries with the appropriate infrastructure, some numbers of respondents will not have sufficient familiarity with technology. Consequently, as noted previously in the case of TIMSS 2019, estimating mode effects (i.e., the impact of administration mode on item parameters and the implications for secondary analyses) is an essential step in effecting the transition (Fishbein et al., 2018). Estimation of mode effects and making appropriate adjustments to place scores from the two modes on the same scale can be technically challenging (von Davier, Khorramdel, et al., 2019; von Davier, Yamamoto, et al., 2019).[34]

As the operational aspects of digitally-based assessments have become more routine, capacity (and incentive) to introduce novel item types to enhance construct representation has grown. These have included items with new response formats (e.g., ranking, multiple response, drag and drop), as well as various types of interactive items such as simulations that are now used in the assessment of science. In each case, suitable psychometric models must be proposed and the item parameters must be properly calibrated. As the divergence between the paper-and-pencil administration and digitally-based assessment instruments increases, the challenge of placing the scores on the same scale will become more formidable and the results more dependent on model assumptions (see the section "Threats to Relevance" for further discussion).

Inasmuch as digitally-based assessments, especially with adaptive testing, can accommodate larger item pools, trend items may constitute only about 40% of the item pool. In the case of an adaptive assessment, it is critical that all the virtual forms administered contain sufficient numbers of trend items.[35] Thus, the challenge of successfully implementing a linking strategy is somewhat greater with this mode of assessment. Down the road, as the digitally-based assessment is administered with a greater range of delivery devices, estimating "device effects" will assume greater importance. Depending on the empirical findings, some restrictions on acceptable devices may have to be instituted.

Measurement Invariance

Issues related to comparability have already been mentioned in the sections "The Messick Framework, Extended" and "Population Sampling Operations." The latter section addressed the procedures developed to obtain statistically equivalent probability samples across participating countries. Equally important is ensuring that the cognitive scores and the responses to the background questionnaire have comparable meaning across participating countries as well. The terms *measurement invariance* or *cross-cultural equivalence* refer to this notion of comparability of meaning.

Measurement invariance is a term that refers to the degree to which the underlying construct retains the same meaning across different settings. For ILSA, the settings are the different participating countries. In the ILSA context, the requirements for strict measurement invariance are very difficult to satisfy. Indeed, a persistent technical problem is the lack of agreement on how to quantify deviations from measurement invariance at the different levels of stringency. One approach in common use is multigroup confirmatory factor analysis. Although a number of fit statistics are available for such models, as well as for recent refinements, their distributional properties in the context of ILSA data are a subject of ongoing research (Byrne & van de Vijver, 2017; L. Rutkowski & Svetina, 2017; van de Vijver et al., 2019). Van de Vijver (2018) reviewed the recent literature and noted that Bayesian approaches (e.g., Carlin & Louis, 2000) may prove useful in this context. In the interim, psychometricians will continue to employ "rules of thumb" with an admixture of intensive data scrutiny. In the cognitive domain, current procedures, such as freeing the globally estimated IRT parameters for those items judged to have strong interactions with countries/languages, may not be sufficient.[36]

Van de Vijver (2018) proposed a tripartite framework for evaluating and addressing threats to cross-cultural equivalence: construct bias, method bias, and item bias. In the context of ILSAs, he asserted that with regard to the cognitive domains and, particularly, the scales derived from responses to the background questionnaire, method bias is perhaps the greatest threat to comparability. He cited such issues as variations in response styles, familiarity with the types of questions, and even poor reading skills in school-based ILSAs. In the case of PIAAC, because of the nature of the administration, individual differences in listening skills and aural memory are also potential sources of method bias.

In the case of sample surveys, a number of indicators can quantify the degree of departure from the ideal. Similarly, with regard to measurement invariance, there is a tripartite framework for evaluating the level of measurement invariance achieved (van de Vijver et al., 2019). In order of increasing stringency, the levels are configural, metric, and scalar. There are statistical procedures that can be used to distinguish among the levels and, hence, suggest limitations on the interpretability of the comparisons among countries. (Consult van de Vijver et al., 2019, and the references therein for further details.)

Notably, van de Vijver et al. (2019) also discussed steps that are taken to enhance measurement invariance, particularly with regard to the background questionnaire, where linguistic and cultural differences play an important role. They emphasized the importance of developing consensus-based frameworks (at different levels of generality) to guide the overall design and item development. There is considerable overlap with the presentation in the section "Transitioning to Digitally-Based Assessments." Item-by-country interactions are addressed by Glas and Jehangir (2013) and von Davier and Bezirhan (2022). Further discussion of threats to comparability is contained in the section "Threats to Relevance.".[37]

Generating Results

With the reporting scale established, it is possible to obtain estimates of the parameters (e.g., means, variances, and percentiles) of the proficiency distributions of the reporting groups of interest (e.g., gender, socioeconomic status, immigration status). Parameter estimates are obtained by aggregating the PVs of the individuals comprising the reporting group. Typically, an interim (weighted) estimate of the parameter is calculated for each set of PVs, and the final reported estimate is a simple average of the interim estimates.

Variance estimation is not as straightforward inasmuch as the estimated variance of a group-level estimate has two components: one due to sampling and the other due to measurement error. The former reflects the fact that the respondents are a probability sample from the population (group) of interest and that, in a full replication, a different sample would have been obtained. The estimate of this variance component is derived using standard methods for sample surveys with proper attention to the nature of the sampling frame and the corresponding sampling weights. Again, different ILSAs use different estimation strategies (Rust, 2014; von Davier & Sinharay, 2014). The second component reflects the measurement error in the estimation of proficiencies. For a specific group-level parameter, this component is proportional to the average of the squared differences between the interim estimates and the final reported estimate.

Building on the process for instrument design and development, in conjunction with patterns of item responses, it is possible to describe, in substantive terms, the differences among performances at various locations along the scale. Such descriptions contribute to greater interpretability, thereby enhancing the utility of the assessment results to policy makers and other stakeholders (Kirsch, 2001; Mosenthal &

Kirsch, 1991). These methods have been applied to the cognitive scales constructed for PISA and PIAAC.

A somewhat different approach, based on the behavioral anchoring methods first developed for NAEP, is employed by PIRLS and TIMSS.[38] With this approach, specific benchmarks along the reporting scale are selected and psychometric analyses are conducted to select items that discriminate between adjacent benchmarks (Martin et al., 2020, chap. 15). Identification of commonalities among the items at each benchmark are then used to provide substantively grounded descriptions of performance at the benchmarks.

Design Issues

In ILSAs, it is commonly the case that multiple proficiencies are estimated or that a single proficiency has been decomposed into some number of facets. In that case, multivariate versions of the latent regression model can be implemented. The overall plan is the same, although the mathematical notation becomes more unwieldy and, more important, the calculations become substantially more burdensome (von Davier & Sinharay, 2014). The benefit, however, is that the analysis can take advantage of the correlations among proficiencies (or facets) to yield more accurate estimates overall.

However, this situation can be a double-edged sword. For example, a difficulty arises if multiple skill domains are assessed but are not evenly represented in the assessments administered to individuals. Suppose that each individual is only tested on a subset of the full set of skill domains. In that case, the appropriateness of the PV obtained for the skill that was not tested is strongly dependent on the correctness of the population (statistical) model. PISA is a case in point. For example, the PISA database contains reading PVs for students who received a test form with mathematics and science tasks—but no reading tasks. For these students, generating the PVs for reading depends on the correlation of reading proficiency with mathematics proficiency, science proficiency, and background variables based on data from those students who received a test form that combined reading either with mathematics or with science.[39]

Consideration of this issue highlights the difference between tests that provide individual-level scores and tests that provide only group-level scores. In the case of the former, fully model-based scores would generally be unacceptable. In the case of the latter, because of the random allocation of domain question sets across the population sample, most of the individuals comprising the group would have been exposed to items from each domain. For example, in PISA 2018, reading was the main domain. Accordingly, all students took some reading items, approximately 54% took some math items, and approximately 54% took some science items. Consequently, the estimated group mean score will largely reflect the contributions of those data. However, the estimated standard error associated with that estimated group mean score will reflect the increased measurement error due to the cognitive data missing from one third of the group.

Supplementary Considerations

As noted earlier, an important use of ILSA data is the estimation of relationships between proficiencies and background data or contextual factors. By including the corresponding variables in the latent regression (or at least their proxies embedded in the principal components), approximately unbiased estimates of those relationships are obtainable if the model is correctly specified (Mislevy, 1991). Achieving this goal is an important aspect of the rationale for including all available information in the latent regression model.

At the same time, there is continuing pressure to include a larger number of scales in the background questionnaire. One suggestion is to adopt rotated (incomplete) designs analogous to those used in the cognitive instruments (i.e., between-construct rotation). However, unlike the situation for the cognitive scales where the assumption of an underlying latent trait induces a set of useful conditional independence relationships, such an assumption is not feasible for the background questionnaire. Moreover, it is not clear how to modify the latent regression model to accommodate such designs without making strong assumptions of conditional independence, whose failure would seriously compromise the quality of the estimates reported (von Davier, 2014). Nonetheless, efforts are underway to use within-construct rotations for some background questionnaire indices with a large enough set of constituent items.

An additional consideration arises from the fact that, in the context of a digitally-based assessment, it is possible to collect process data, and the log files that are generated can record every keystroke. The sheer amount of such data is overwhelming; research is being conducted on how such data might be compressed into a small number of indicators that could be incorporated into expanded latent regression models to improve estimation accuracy (Bergner & von Davier, 2019; Kroehne et al., 2016; von Davier, Khorramdel, et al., 2019). Another approach would be to use such indicators to assess the quality of the data before item calibration. For example, individuals whose log files suggest aberrant response processes or patterns could be removed. Models for capturing the speed/accuracy trade-off may prove useful here, as well as in secondary analyses (van der Linden, 2007; Yamamoto & Lennon, 2018). See the sections "Impact" and "Technical Challenges" for further discussion.

DATA PREPARATION, DATA PRODUCTS, DATA ANALYSIS TOOLS, AND RELATED ACTIVITIES

At the conclusion of the analysis stage, each participating country receives a set of databases specific to that country. Ultimately, an international database is constructed, combining certain country-specific databases. The sponsoring organizations, countries, researchers, and other key stakeholders employ the international database for a variety of purposes. ILSA contractors, in conjunction with the sponsoring organization and the participating countries, carry out a number of procedures leading to the dissemination

of the data that have been produced. They also engage in activities that promote the use of the data for a variety of purposes. This section briefly describes five key areas of activity: data preparation, data analysis and reporting tools, data products, data security and confidentiality, and data quality.

Data Preparation

Internal Data Processing

The main objective of this stage is to ensure that the data files adhere to international formats, that the data accurately reflect the information collected by each country, and that the different survey files can be linked appropriately. Upon receipt of the national databases, the contractor archives the data to a preliminary international database. The initial review of the national databases includes the evaluation of the supporting documentation and the required consistency among data sets. Each country is consulted during this step so as to address any issues or concerns.

Once the data files have been made consistent with the international database structure, as specified in the international codebooks, a set of programs is applied to conduct "data cleaning." This process consists of three activities:

- checks related to the identification codes associated with the variables;
- checks related to linkage of variables within a data set and records between data sets, with particular attention to any discrepancies between the observed and expected data patterns; and
- checks related to cognitive and background variables.

Further quality reviews of the submitted data sets are conducted according to the technical standards and guidelines that are associated with each ILSA. These occur at multiple stages of the data preparation workflow. Typically, the contractor supports countries in issue resolution and uses an automated reporting mechanism to communicate with countries regarding the status of data submissions, quality review, and any data quality issues encountered. Once data cleaning and restructuring are completed, the files that include sampling information and the sampling weights are merged into each national database.

Creation of Derived Variables

After the international variables have been harmonized and all data validated, a set of derived variables is computed and added to the data set. A derived variable is typically constructed by combining two or more variables, resulting in a new variable or a scale score. These variables are fully documented, as required for further analysis and reporting activities. Derived variables undergo checks using state-of-the-art quality control methods.

Simultaneous with the creation of the derived variables, the cases in the data file are inspected to ensure that they are appropriate for inclusion in the data file. As a part of this evaluation, case sampling weights are computed and linked to the cases.[40] This

augmented data file is merged into the preliminary international database to yield a fully integrated international database (master database). Extracts from this database are then sent back to each country for its own use and forwarded to those responsible for data analysis.

Analysis and Reporting

Preparation of Databases

To support the export of data products meeting the needs of various users, different data versions can be generated. These usually comprise three files:

1. The national file submitted by the country and processed by the sponsoring organization. This file has all the data collected within a country, including additional subpopulations and variables that were part of the national data collection activities and samples. These files are typically only available from the national center and only to the sponsoring organization and the contractors.
2. A restricted user file with all the variables and cases intended for use in international comparisons. This file is typically used for secondary analyses. It is published by the sponsoring organization and usually available provided assurances of confidentiality are made by the user.
3. A public user file (PUF) contains all the variables and cases intended for use in international comparisons but might have some variables or values for them suppressed or masked. The intention of suppressing or masking values is to prevent the identification of participating individual or subgroups (i.e., schools, neighborhoods). These PUFs are posted on the Internet and can be accessed and used with no restrictions.

Once these databases are completed and have undergone preliminary quality control, standard programs are applied to generate multiple tables that will ultimately appear in technical reports and other documents associated with the ILSA.

Preparation of Public Use Files

Following the initial data-cleaning process, an iterative process of data review and correction takes place—initially among the contractors and later involving the participating countries as well as the sponsoring organization. By default, the PUFs maintain all international variables approved for release. They only include those records that fulfill criteria set for cases to be used in weighting and analysis.

Some procedures are specific to the sponsoring organization. For example, for OECD ILSAs, any and all national variables identified by a country for deletion are dropped. In addition, all international variables earmarked for suppression by a country are blanked (i.e., set to the appropriate missing value for all cases). By contrast, for TIMSS and PIRLS (sponsored by IEA), national variables are typically not included in the international database, but are made available to the nation asking these additional questions so the data can be merged. However, the international source versions

of context questionnaires can include items specifically created for that purpose. As an example, a small number of international variables can be provided for countries to include additional home possessions variables that could contribute to an indicator of socioeconomic status. These variables are labeled "country specific" in the database, and their English back translations are listed in a National Adaptations Database included with the user guide. Any context variables earmarked for suppression by a country are blanked. However, cognitive items may be first reviewed in consultation with the country. All such cases are documented in the user guide or in the technical report.

Each country's database is included in the following types of electronic files:

- Data files in SAS and SPSS formats, with separate files for each respondent (student, teacher, principal, parent). Other formats are available for different ILSAs. This file (or multiple data files containing appropriate linking information and instructions) will contain all the variables approved for release to the public.
- For PISA and PIAAC, a data codebook document file in either PDF or EXCEL format. This two-part document contains a brief listing of the name, position, format, and description of each data variable, plus an expanded listing for the discrete variables that have coded information, including their code values, descriptors, and (unweighted) frequencies within each data set.
- For TIMSS and PIRLS, a codebook that provides a list of variables with the corresponding labels and formats. Additional information about variables is provided in the user guide or in the technical report. Finally, data almanacs contain frequency distributions and summary statistics for all variables.
- For TIMSS and PIRLS, a software package (IDB analyzer) is available for R, SPSS, and SAS versions. It is updated to include special routines to load and analyze data from IEA assessments.[41]
- Documentation that describes the contents and structures of the data files, explains how to employ the other resource files in the product, and provides the technical information needed to support users in conducting analyses of the data. This includes the names and uses of the key variables in each data set, as well as indicators of sample quality (e.g., response rates).

Preparation of Data Tables, Compendia, and Reporting Tables

SUMMARY TABLES FOR COGNITIVE ITEMS AND BACKGROUND VARIABLES One set of tables contains weighted summary statistics for each participating country on each cognitive item and on each variable in the background questionnaires. For the latter, summary tables display the international averages for each variable, with each country weighted equally. For each variable, the summary tables display the question that was asked, the location in the corresponding questionnaire, and the variable name in the data file. Note that the table format depends on whether the data are categorical or continuous.

The cognitive data summary tables (sometimes referred to as item analysis tables) contain summary information about the responses to the cognitive items. For each country, they contain weighted summary statistics, including variable identification, sample size, number of valid cases, weighted percentages of individuals corresponding to each valid response, weighted percentages of individuals who did not select any of the valid response options, and the average score on each of the cognitive domains. Standard errors are also included, where applicable. The item analysis tables are used by both the team and countries for further quality control, verifying data structure accuracy, and validation purposes.

Unweighted item analysis results are also generated because they are particularly useful in verifying the accuracy of the data structure. When necessary, for internal quality control, the team may produce separate analyses to compare item statistics for respondent groups taking different modes of assessment (for example, paper-and-pencil administration and digitally-based assessments) or for respondent groups determined by selected background variable categories.

COMPENDIA Using the international restricted user file and PUFs as the source data, compendia are sets of tables that provide the distribution of students or adults according to the variables collected by the questionnaires. The public version is essentially a redacted version of the summary data tables. The purpose of the public compendia is to support PUF users so they can better appreciate the contents of the PUF and, importantly, use the compendia results to verify that they are performing PUF analyses correctly. For confidentiality reasons, some countries may decide to alter their data or remove respondent records from their PUF files. In such cases, individual country summary statistics reported in the compendia may differ slightly from statistics used for international reporting.

INTERNATIONAL REPORT TABLES These report tables are publication-ready tables that support the international report. Typically, approximately 350–400 international report tables of varying length and complexity are generated. These tables can be used directly by the sponsoring agency or by countries as a means of quality control if they choose to conduct their own analyses and table production.

Preparation of the Technical Report

The technical report summarizes and describes all aspects of the study, including development of the frameworks for the background and context questionnaires, as well as for the cognitive domains; design and development of the survey instruments; development of the computer platform to manage, develop, and deliver the survey; translation and verification processes and procedures; survey operations and quality control procedures; sampling and weighting, data collection, and data processing; scaling, analysis, and preparation of data products; and reporting the results.

The technical report also includes any additional information related to the implementation of the psychometric and statistical methodologies at a level of detail that allows researchers to understand and replicate the analyses. To the extent possible, the report also addresses a range of common questions that both more and less advanced users of these complex databases might pose.

Tools and Training

A number of publicly available tools have been developed for use by countries, sponsoring organizations, and secondary researchers. For ILSAs sponsored by the OECD, two tools are available: One is a Data Explorer that allows users to navigate a secure, hosted database that includes all assessment cycles. The second is the IEA's International Database Analyzer (IDB Analyzer) that was adapted to work with OECD ILSAs. It enables local computer access and analysis of public-use and restricted-use databases.[42]

Each tool addresses a slightly different set of needs and audiences. The Data Explorer is a web-based application that is usually hosted either by the sponsoring organization or by a government entity. It protects the underlying database while allowing relatively easy point-and-click navigation of the database to generate publication-ready summary cross-tabular tables and graphics.

The IDB Analyzer can be used in conjunction with restricted-use or public-use versions of the international database and uses microlevel data stored in a local computer (IEA, 2021). The IDB Analyzer satisfies the three main technical requirements that an analysis of these data must meet. They are (a) the use of sampling weights, (b) accommodating in variance estimation the complex multistage cluster sample design that was implemented, and (c) the use of multiple imputed proficiency estimates (i.e., the PVs).

Most commonly, the analyzer is used to generate data tables and graphs, as well as to execute either normal theory regressions or logistic regressions. The primary benefit of employing this tool is that it efficiently performs the required analytical computations, thus relieving the user of the programming burden. This process is especially critical in the computation of variances of estimates. In addition, the tool is accompanied by sufficient technical information and references to support both more- and less-advanced users in performing statistically correct analyses of the complex survey sample data.

To enhance data utility, sponsors and contractors organize and conduct workshops in various venues with different cost structures. The workshops focus on understanding the structure of the international database, the use of the IDB Analyzer and the Data Explorer, or other tools including standard statistical software. The participants in these workshops include staff from each participating country and interested representatives from the sponsoring agency. The sessions are arranged and coordinated by the contractor using staff members who are involved in or knowledgeable about the project design, the database preparation, and the appropriate analysis of the international data. The goals of the workshops are to help the representatives from each participating

country to understand the strengths and limitations of the data and to ensure that they are comfortable using the analytic tools that will be available and will report the results of the analyses with the relevant context and appropriate cautions.

In addition, the contractor or other organizations offer countries or interested researchers a fee-based option to organize and conduct regional workshops based on country-specified topics dealing with the technical aspects of using and analyzing a particular database. Additional workshops, presented under the auspices of the IEA–ETS Research Institute, the Institute for Education Sciences (U.S. Department of Education), and international donor organizations, are intended for interested researchers to expand the pool of knowledgeable users.[43] For example, the Institute for Education Sciences funded a multiyear set of training workshops on the use of the PIAAC data that resulted in dozens of research projects and publications. In addition, the IEA has constructed a website containing information about many ILSAs.[44]

Data Security and Confidentiality

Data protection and confidentiality are essential to the conduct of all major ILSAs. Organizations wishing to conduct these types of assessments must comply with the national requirements of participating countries, along with international requirements associated with entities such as the European Union. Typically, all staff who work with these databases are required to sign a statement that they have read the information protection policy and that they understand and agree to abide by its provisions.

A growing number of organizations have adopted the International Organization for Standardization (ISO) 27000 series as their information security framework. ISO 27000 is an international series of standards that drives all aspects of the information security program and the way services are delivered. In the future, any organization wishing to develop or implement ILSAs will have to demonstrate that it is following applicable legal and regulatory obligations, with particular attention to the protection of personally identifiable information.

Typically, ILSA contractors have adopted stringent security policies, standards, and practices to maintain the confidentiality, integrity, and availability of data. These controls are continually reviewed to ensure they align with industry best practices. In addition, contracting organizations need to support secure transfer protocols, including file-based technologies to securely exchange data over the Internet. This enables them to effectively manage and protect data transmissions with secure file transfer protocol.

Data Quality

International large-scale assessments are complex surveys that require the development of a set of procedures for all major phases to help ensure that the sources of bias and variability in survey results are kept to a minimum so that reported results are relevant, comparable, and interpretable. Here, we distinguish two ways in which large-scale assessments address issues of data quality: the technical standards and guidelines that are developed and/or refined at the beginning of any assessment and the data

adjudication process that is conducted at the completion of the data collection and analysis process.

Technical Standards and Guidelines

At the beginning of each ILSA cycle, the ILSA team collaborates with the sponsoring organization and participating countries to develop and/or refine a set of technical standards and guidelines to ensure that all aspects of the survey satisfy accepted quality assurance guidelines. These standards and guidelines represent policies or best practices that must be adhered to in the development and conduct of the survey. Where compliance is not possible, countries may apply for derogations from the standards, as long as they are judged not to compromise the survey objectives. The overarching goal is to minimize total survey error (see the section "Population Sampling Operations"). The standards and guidelines cover a wide range of topics, including ethics, sampling, instrumentation, translation and adaptation, the use of hardware and software (if it is a technology-based assessment), data collection and training, data capture and processing, and data confidentiality and security.[45]

Data Adjudication

Compliance with the technical standards and guidelines is an important component of ensuring the quality of the national data through the various stages of the workflow. At the conclusion of the workflow, yielding the databases described previously, data adjudication is implemented. Its objective is to render a judgment regarding the overall quality and fitness of the data of each country and to impose, if necessary, any limitations that should apply to the public dissemination and use of these data. In other words, the intent is to go beyond compliance with the standards and guidelines and determine whether the national data are of a sufficient quality to support their intended interpretations and uses. Adjudication typically involves three key areas: sampling and coverage of the target population, data collection and instrumentation, and quality of the translation/adaptation procedures.

For the OECD programs, the contractors develop an adjudication report that is reviewed and discussed at the final meeting of the Technical Advisory Group. If questions arise during the meeting, a sampling referee is available to offer advice as needed. The Technical Advisory Group then prepares a final set of recommendations to the sponsoring organization regarding the fitness of the data for publication. For IEA programs, the adjudication is carried out by a team comprising the different contractors and IEA staff.

IMPACT

Messick (1987) argued that large-scale assessments are a type of policy research and should be judged as such. In that light, if we regard policy utility as having the potential to inform, influence, or impact policy, then it is reasonable to ask to what extent—and

in what ways—that potential has been realized by ILSAs. As noted in the section "The Messick Framework, Extended," the overall judgment of the potential for policy utility is dependent on evaluating the strength of an ILSA with respect to the design criteria of comparability, interpretability, and relevance—along with due regard to purpose, context, and the population(s) of concern. For the most part, users of ILSA results assume that the level of comparability is at least adequate for their purposes and that they (or their staff) have reached appropriate and defensible interpretations of the data. Ultimately, the judgment of relevance provides at least part of the motivation to cite the data in discussions and policy debates. However, the contexts in which ILSA data are employed and their actual impact on education policies and practices are the result of complex dynamics that vary widely over time and space. Thus, the evaluation of impact requires considering a broad range of cases from which a few generalizations can be gleaned.

Background

To best address the question of impact, it is helpful to both distinguish different types of impact and then systematically evaluate the various kinds of evidence that are available. In this regard, D. Rutkowski et al. (2020) presented a framework for evaluating ILSA influence on policymaking. The framework comprises a causal model for tracing the impact of ILSAs on a particular policy and a logic model for evaluating the evidence for such impact.

Drawing on D. Rutkowski et al. (2020), we define impact as occurring when it is possible to connect (in some way) ILSA data, technical reports, or secondary analyses to changes in policy and/or practice in a participating country or even shifts in the content of conversations about policies in education, workforce development, or labor market regulation at the national or cross-national levels. However, as Wagemaker (2014) pointed out in discussing school-based ILSAs, it can be problematic to judge policy impact because

> major policy initiatives or reforms are more likely to result from a wide variety of inputs and influences rather than from a single piece of research. Research is also more likely to provide a heuristic for policy intervention or development rather than being directly linked, in a simple linear fashion, to a particular policy intervention. (p. 12)

Two further challenges are (a) a frequent time lag between ILSA reports and policy decisions and a further lag between policy initiation and its effect (if any) on measured achievement or skills and (b) variations in policy implementation across the target units within a country that can reduce estimates of policy impact (Braun et al., 2006; Burdett & O'Donnell, 2016). Thus, correctly judging impact involves triangulating different kinds of evidence, including statements of key actors, stakeholders, and others as to the role of ILSAs in policy decisions, formulations, and implementation (Fischman et al., 2019). Accordingly, we discuss two types of impact: direct and indirect.

Direct Impact

Direct impact occurs when results released to the public (appear to) have a relatively clear, and even immediate, influence on policy makers' discussions, priorities, and actions. Some actions may address governance and funding; others relate to curriculum, teacher training, pedagogy, and assessment; and yet others involve achieving greater harmonization across different units within the education system.

The literature on school-based ILSAs is replete with examples of how ILSA results precipitated intense discussions that led to substantial policy changes. Perhaps the most celebrated example is the impact of PISA 2000 results on Germany (Ertl, 2006). The relatively poor performance of German students (following their poor performance on TIMSS 1995) resulted in unprecedented cooperation among the federal government and all the German states, leading to changes in curriculum and assessment. The need for such changes was supported by results from PIRLS 2001, which indicated that children from immigrant families on average performed substantially more poorly than their peers. In response, the national research agenda shifted toward more empirically based studies with direct relevance to practice (Schwippert, 2007).

However, Pons (2017) suggested that this type of analysis may be too simple. He argued that before reaching a judgment on the "PISA effect" (or its equivalent), it is necessary to consider multiple factors, including the prior policy context and education-related political debates. On this point, he asserts, "PISA uses in the policy process greatly depend on the dynamics of the domestic policy debate and on the preexistence of specific structuring controversies that PISA results illuminate in a new way" (Pons, 2017, p. 139). This is consistent with the perspective offered by Ritzen (2013) and Braun (2013). Indeed, there are many types of "policy borrowing," ranging from fairly direct adoption to adapting certain policy features as part of a broader policy formulation (Burdett & O'Donnell, 2016).

Heyneman and Lee (2014) cited additional examples of direct impact in countries as diverse as Denmark, Macedonia, Kuwait, and Japan. They also provided an extensive list of direct impacts on policies in other countries, as well as on subject-specific policies and practices. Less noticed, but perhaps equally important, ILSA results have also been used to provide support for existing policies. Heyneman and Lee (2014) cited the example of PIRLS 2001 in England. The relatively strong performance of English students was used as evidence of the effectiveness of the National Literacy Project that was introduced in 1996. Similarly, the performance of students in Australia and New Zealand on PISA has been taken as support for existing educational policies. Breakspear (2012) called out countries such as Singapore and Spain that use PISA to complement (and to validate) the results of their national assessments. Canada participates in most ILSAs, often with oversampling, and makes systematic use of the results to guide policy at the national and provincial levels (Volante, 2013).

Singapore represents an interesting case study (Ng et al., 2020). Singapore has participated in IEA studies since the 1990s and employs ILSA results to inform curriculum and pedagogy, as well as to evaluate the efficacy of various reforms. The authors stress,

however, that these results are only one of many sources of evidence that policy makers and education leaders consider in deciding on policy strategies and education initiatives. They also note that to discourage score corruption, the Ministry of Education uses the scores only to inform system-level decisions and not for any form of accountability.

Baird et al. (2016) conducted an intensive study of the impact of PISA on six countries.[46] Based on both a comprehensive document review and multiple interviews, they concluded that in five of the countries (excluding England), some of the policy changes could reasonably be attributed to the impact of PISA results. For example, the two federal systems (Canada and Switzerland) initiated a process of greater harmonization of achievement standards and assessments across provinces and cantons, respectively. In the case of France, PISA prompted substantial curricular changes, as well as initiation of a sample-based national assessment modeled on the ILSA design.

With regard to PIAAC, there is little evidence of direct impact. On the one hand, this may be because different ministries focus on one or another aspect of the results and then evaluate them in light of other relevant information, as well as political considerations. On the other hand, the assessment of adults' cognitive skills, as well as their distribution at different levels of educational attainment, has proven revelatory.

Indirect Impact

Indirect impact occurs in a variety of ways. First, ILSA findings may be used to justify, legitimate, or build public support for policy prescriptions that have already been formulated and even initiated (Fischman et al., 2019). Second, the results of one or more ILSAs over multiple cycles can change the nature of the discourse regarding important aspects of education, skills, and skill development. In addition to the achievement results themselves, subsequent to the release of the PUFs, analysts delve into the ILSA databases to investigate patterns of relationships and how they differ across countries and over time.

The comparative findings can influence policy discussions regarding the need to remedy apparent deficits or inequities. In other cases, researchers attempt to draw inferences regarding the efficacy (or lack thereof) of particular policies or constellations of policies with regard to certain outcomes. Yet others seek to draw attention to possible unintended consequences of employing ILSA data in policy discussions. Such secondary analyses, whatever their nature, appear years after the initial release of the data. Nonetheless, the findings and interpretations may also influence the discourse related to particular policies and, in some instances, even provide the impetus for changes in policy or practice.

As noted earlier, ILSAs have experienced increased participation by medium- and low-income countries. Undeniably, there is some element of coercion by funders. However, in addition to providing critical policy insights and guidance, there are many other salutary benefits. These include constructive changes in regulatory policies (e.g., curriculum content, performance standards) and in so-called behavioral policies (e.g., pedagogy, educator professional development; Lockheed, 2013; Wagemaker, 2014).

A particularly important benefit associated with participation in school-based or adult assessments is the contribution they make to capacity building. By participating in an ILSA, country representatives and technical experts receive extensive training and support in all aspects of managing and conducting large-scale assessments (Meinck et al., 2020). These experiences and the expertise gained can then be used not only in future ILSAs, but also in improving (or initiating) national assessments that are directly targeted at informing education policies of greatest priority. Lockheed (2013) cited numerous self-reports of such impact.

Since 2003, participation in the ILSAs that focus on adults has grown in tandem with a marked shift in focus among policy makers who have moved from viewing educational attainment as an appropriate proxy for skills to recognizing the added value of comparable and valid/relevant measures of skills of interest.[47] One reason is the accumulating evidence that there are substantial economic returns to skills beyond those accounted for by educational attainment or years of education (Fogg et al., 2018; Hanushek et al., 2015; Kirsch et al., 2007). Such findings not only enhance policy interest in tracking skills, but also highlight the societal perils in tolerating extreme inequalities in skill distributions. Consequently, in many countries, finding ways to accelerate skill growth while reducing inequities have become policy imperatives—fueling greater interest in studying the policy initiatives of "high-flyers" (i.e., those countries at the top of the league tables or those who have experienced large gains over two or more cycles).

Indeed, greater attention in the media and the exponential increase in Internet searches related to ILSAs are testament to their growing role around the world in discussions of education policies and practices. As Ritzen (2013) pointed out, the comparative outcomes derived from ILSAs make these assessments powerful instruments of transparency. Their findings can jumpstart (or accelerate) national conversations regarding quality and equity in education or in labor market dynamics. Over time, such conversations, particularly at the policy level, can lead to shifts in perspectives, new understandings, and the realization that far-reaching system changes are not only possible but also much needed (Conaway, 2020).

Secondary Analyses: Exemplars

There has been an increase in the number and range of researchers using ILSA results for studies addressing important policy issues, with a concomitant increase in the utilization of the IEA's IDB Analyzer. Pons (2017) provided a useful review and analysis of articles employing PISA data. With respect to adult assessments, Maehler et al. (2020) compiled a bibliography of published papers between 2008 and 2019 covering a range of topics including research results associated with PIAAC assessment. Since 2013 when the PIAAC data were first released, the number of publications has increased annually. The current bibliography contains more than 600 publications and 21 technical reports. More generally, the IEA gateway (https://ilsa-gateway.org/studies/papers) contains a comprehensive list of articles employing both IEA- and OECD-sponsored ILSAs. For a more recent integrative review, see Hernández-Torrano and Courtney (2021).

Among the issues considered in these analyses are skill gaps within and among participating countries. Of particular import are studies comparing countries with respect to the gradient of performance against a measure of socioeconomic status.[48] Although a gradient exists in every country among school-age children (i.e., mean performance being positively correlated with a measure of socioeconomic status), there is a wide range of gradient values, some strikingly large (OECD, 2019, chap. 2). As it happens, the gradient in the United States is typically the largest among OECD countries.

There are similar findings for adult populations. For example, Goodman et al. (2015) examined the skills distributions and inequality among millennials in the United States with respect to their overall performance, as well as in comparison to other countries participating in PIAAC. For the United States, the authors considered differences in skill distributions by factors such as educational attainment and race/ethnicity. With regard to gender differences, Braun (2018), employing PIAAC data, demonstrated that in every participating OECD country the earnings of women employed full-time trailed those of men employed full-time, even after adjusting for age, family background, cognitive skills, educational attainment, and occupational category. Again, the degree of disadvantage varied widely across countries. In an empirical study using PIAAC data, Vera-Toscano et al. (2017) demonstrated that, over and above educational attainment, measured cognitive skills have incremental predictive validity for social outcomes such as participation in volunteering activities.

Strietholt and Scherer (2018) argued that combining different ILSAs or linking ILSAs to other data sets increases the range of research questions that can be addressed. They cite, for example, the study by Martin et al. (2013) that took advantage of the conjunction of PIRLS and TIMSS in 2011. Through careful planning, 34 countries and 3 benchmarking countries administered PIRLS and TIMSS to the same fourth-grade students. Thus, each student generated data on skills in reading, mathematics, and science and responded to background, attitudinal, behavioral, and school context questions. Among other things, the authors found substantial heterogeneity across countries in the amount of variation in the cognitive outcomes accounted for by various background variables. By contrast, within countries they found typically small differences in the amount of variance explained across cognitive domains.

Strietholt and Rosén (2016) offered an example of combining different assessments over time. Specifically, they linked the Reading Comprehension Study 1971, the Reading Literacy Study 2001/1991, and PIRLS 2001, 2006, and 2011 to study reading trends over a span of 40 years.[49] Hanushek et al.'s (2015) study cited previously involved combining data from PIAAC with income data obtained from various national and international sources. In summary, the results of cross-national, comparative assessments, as well as the studies that employ those results, can provoke useful discussions among researchers and stakeholders that likely would not otherwise take place.

Technical Issues in the Study of ILSA Impact

As this section and the references therein have amply documented, ILSAs have had considerable indirect and even direct impact on education and skill development policies around the globe. However, the process of policy formulation, adoption, and implementation has a strong political component that can dominate the influence of ILSA data and related secondary analyses. Indeed, as already noted, impact depends on the readiness of the relevant national actors to take account of the information provided, to decide on which action(s) to take, to articulate policies and plans, and to commit both the necessary funding and the political capital to implement these policies. ILSA information is just one source of evidence (or pressure) among many that can motivate or drive educational change.

Notwithstanding these political realities, technical considerations should (and sometimes do) urge caution on ILSA interpretations and impact. In the early 2000s, critics of ILSAs typically focused on technical deficiencies related to sampling and translation/adaptation that, in their view, substantially reduced ILSA utility (Bracey, 2008; Goldstein, 2004). Later, focusing on PISA 2012, L. Rutkowski and Rutkowski (2016) undertook an analysis of these technical issues with regard to their implications for defensible interpretations of the results and the use of those results in policy discussions. Their principal recommendation was that sponsors should be fully transparent regarding the bias and uncertainty (variance) associated with published results and their implications for the interpretation of such results. Despite the significant improvements in PISA technical quality achieved since then, as well as those realized with other ILSAs, the cautions and recommendations offered by the authors should be borne in mind by both sponsors and users of all ILSA results.

For example, the assumption of cross-national comparability of domain scores and responses to the background questionnaire is only an approximation (see the section "Measurement Invariance"). Consequently, conclusions should be tempered by an acknowledgment that for a particular analysis the assumption may not hold for some countries to an extent that renders the conclusions suspect, at least for those countries. Such cautions may limit the impact or even the influence of the study. On this point, it is beneficial (if not essential) that participating countries have the "in-house" expertise to analyze their data, to explain the findings to various stakeholders, and to contribute constructively to policy conversations. Participation in ILSAs has stimulated many countries to invest in developing such expertise, along with the infrastructure required to support high-level methodological research.

More broadly, success in policy borrowing is difficult to achieve. Education systems are complex: They comprise many component systems with complicated and shifting dynamics. Identifying a particular policy or practice as a key driver of academic achievement may neglect concomitant factors or conditions that are essential to its effectiveness. Even if the identification is approximately correct, adapting the policy or practice to a setting that differs in many relevant ways (e.g., power dynamics, traditions and culture of schools, resources) is

challenging. In the course of technical adaptation and responding to political pressures, the essence of the innovation may be "lost in translation" (Burdett & O'Donnell, 2016).

Furthermore, the information provided by an ILSA is typically more narrowly focused than the general questions posed by policy makers (D. Rutkowski & Delandshere, 2016). For example, policy makers would like an answer to the question of whether observed differences in achievement among countries can be attributed to differences in education policies. However, as pointed out by Braun and Singer (2019),

> Although it is the kind of question policy-makers often view as most relevant, most methodologists agree that ILSAs are ill-equipped to provide unambiguous answers. Cross-sectional data based on samples without random assignment of subjects into experimental and control groups do not lend themselves to valid causal inferences. Attempting to infer causation from correlation, even when correlations are high, can lead to false, incomplete, or misleading conclusions. (pp. 79–80)

Because the successful implementation of educational policies across so many different contexts depends on a broad range of relevant factors, both educational and otherwise, it is exceedingly difficult to disentangle the influence of these factors with cross-sectional, observational data. Table 20.2 displays six possible uses of ILSA results with judgments on the general suitability of ILSA data for such uses. Evidently, suitability declines as the uses shift from description to inference.

Table 20.2 Purposes of School-Based International Assessments

Item	Purpose	Capacity of ILSAs to Achieve This Purpose
1	To disturb complacency about a nation's education system and *spur education reforms*.	Outstanding
2	To *describe* and *compare* student achievement and contextual factors (e.g., policies, student characteristics) across nations.	Excellent
3	To *track changes over time* in student achievement, contextual factors, and their mutual relationships, within and across nations.	Excellent
4	To *create de facto international benchmarking*, by identifying top-performing nations and countries, or those making unusually large gains, and learning from their practices.	With caveats
5	To *evaluate* the effectiveness of curricula, instructional strategies, and education policies.	With extreme caution
6	To *explore causal relationships* between contextual factors (demographic, social, economic, and educational variables) and student achievement.	Dangerously difficult

Note. ILSA = international large-scale assessment.

D. Rutkowski and Delandshere (2016) argued that causal claims based on ILSA data should be critically evaluated through the lens of a validity framework comprising four facets: construct validity, internal validity, external validity, and statistical conclusion validity. The findings of such an evaluation can be quite sobering, as the examples cited by the authors attest. In general, causal claims can be buttressed both by an a priori theory of action and by an argument of the comprehensiveness of the observed variables included in the analysis. Nonetheless, such claims must remain tentative until other supporting evidence comes to light (Singer et al., 2018).

That said, even though ILSAs rarely permit causal inferences, their findings can reveal striking patterns, raising issues that can be carefully examined through further empirical studies. This caution is particularly relevant in moving from tentative causal descriptions to credible causal explanations, with the latter being more relevant to policy decisions and actions (D. Rutkowski & Delandshere, 2016). In this regard, small-scale experimental studies that demonstrate the importance of certain interventions or strategies can then be more broadly validated within and across national samples.

Some Extensions to Discerning Impact

It is sometimes possible to take advantage of auxiliary information to circumvent the limitations of cross-sectional studies. One strategy employs an instrumental variable to obtain an estimate of a causal effect. Unfortunately, it is usually difficult to find a suitable instrument, especially in an international context. Pokropek (2016b) presented one example using data from Poland. Another example was offered by Bedard and Dhuey (2006). They employed school entry cutoff dates as an instrument to estimate the impact of delayed entry into schooling (see also Marchionni & Vazquez, 2019). In other settings, regression discontinuity designs may also prove useful (Robinson, 2014).

A different approach is to conduct true longitudinal studies in which individuals participate in related assessments on two or more occasions. Differences in outcomes can be linked to differences in treatments (or other factors) to yield causal estimates with some degree of credibility. Carrying out such studies always poses substantial technical, logistical, and sometimes ethical challenges, especially in an international context. One strategy is to construct a pseudo-longitudinal study by concatenating different assessments suitably spaced in time. Kaplan and McCarty (2013), employing data for PISA and Teaching and Learning International Survey (TALIS) from Iceland, examined a number of methods for creating a synthetic data file. At the national level, there are a few instances of such studies. In Denmark, for example, a subsample of the PISA 2000 sample was interviewed and tested as part of PIAAC 2012. By linking the assessment data to data from administrative registers, the author was able to account for some of the variation in individual rankings between assessments in terms of both fixed and varying characteristics of the individuals (Rosdahl, 2014). However, the possibility of construct shift suggests caution in interpretation.

An alternative is to combine data from successive cycles of the same assessment (e.g., PIRLS 2016 with PIRLS 2021) and relate changes in the assumed causal agent with observed changes in achievement. This feature has been exploited by Gustafsson (2013), Gustafsson and Nilsen (2016,) and Hooper (2017). The basic idea, attributable to Gustafsson, is to employ the difference-in-differences methodology at the national or subnational level. Needless to say, the validity of the causal claims depends on building the counterfactual case that changes in achievement can only be due to the changes in the hypothesized causal agent. In the differences-in-differences methodology (Cunningham, 2021), other plausible causes are discounted by means of the parallel trends assumption, where trends between similar populations (e.g., countries) are expected to be parallel if an effective causal agent is not present. Parallel trends between similar countries (e.g., Sweden, Norway) occur frequently in economic data where higher level economic trends (e.g., global economy) have a strong influence on lower level economies (e.g., national economies). Parallel trends are less common in international assessment data, where cycle-to-cycle fluctuations in achievement results do not tend to follow a similar pattern across countries, making it more difficult to assume that deviations from parallelism are due solely to the presumed cause. Given the demand for credible policy evaluation, these and other methods are the subject of continuing research.

The study of correlates of student achievement is of long-standing interest. Because TIMSS and PIRLS sample students by class, collecting data on students and their teachers, they provide an opportunity to link teacher characteristics and teacher practices to student achievement at an international level. However, because student achievement is cumulative, its association with teacher characteristics in a particular year may be quite weak. Comparisons among countries with regard to both teacher practices and student achievement can be suggestive of directions for further investigation. O'Dwyer and Paolucci (2019) discussed what has been learned, along with a careful analysis of the obstacles to making inferences about causal relationships.

The challenge of addressing questions regarding the efficacy of reform initiatives with ILSA data has engendered considerable ingenuity among methodologists, leading to new analytic strategies. Similarly, as the section "Transitioning to Digitally-Based Assessments" amply demonstrates, ILSA teams have harnessed emerging technologies to construct innovative platforms to carry out the processes that undergird the ILSA programs. These innovations have enabled the ILSA teams to meet the ever-increasing demands of sponsors and countries while meeting constraints of time and cost. Both developments exemplify the virtuous spiral (Figure 20.2). In the section "History of International Assessments" we noted that the ETS proposal to reinvent NAEP (Messick et al., 1983) represented a creative response to policy makers' demands for NAEP to provide data that were both more relevant and more interpretable. Certainly, as the global landscape continues to evolve and as ILSAs yield more useful insights, new demands will arise necessitating further innovation. Accordingly, the following two sections discuss some of the technical and political challenges that ILSA teams will likely face in the coming years.

TECHNICAL CHALLENGES

Notwithstanding the remarkable growth in ILSA participation since the 1990s, as well as the bright prospects for ILSAs with a future based on digital platforms, program sponsors and contractors will be confronting a number of technical challenges to policy utility. Many of these challenges arise from the need to maintain (or even enhance) relevance as the context of use and the intended applications continue to evolve. With the corresponding changes in the assessment frameworks, instrument development, and mode of administration comes the need for additional evidence to support validity claims.

Indeed, an underlying theme of this chapter has been the assurance of validity of interpretation and use within a policy setting. A judgment of validity must be made with due regard to the context of use and the particular purpose(s) for which the results will be employed (Kane, 2013, 2016). Much of the relevant evidence is contained in the technical reports that accompany each ILSA and has been discussed in various sections of this chapter. The section "Transitioning to Digitally-Based Assessments," for example, describes the key phases of the workflow processes, including the specification of the frameworks that guide the development of the instruments and the accompanying quality control procedures that are applied during assessment development, translation/adaptation, and implementation. These methodologies, along with the data generated through monitoring their implementation, constitute evidence for "procedural validity," forming a foundation for a validity argument.

Similarly, the section "Scaling, Population Modeling, and Proficiency Estimation" provides an overview of the statistical and psychometric models and procedures that transform the raw data into the published results. The mathematical and verbal representations of these models and procedures make explicit the strategies employed to take account of ILSA design features (e.g., planned missingness, group-level reporting) in generating the results. Thus, these strategies can be—and have been—subjected to critical review and ongoing refinement to support the validity claims.

The section "Policy and Political Challenges" addresses some of the political challenges ILSA sponsors and contractors face with the greater salience of ILSA results in national and global educational policy debates. Much of that discussion can be framed in terms of the consequential validity of ILSAs, which is essential to the role of ILSAs as policy research and can sometimes become politicized and highly charged. In contrast, this section focuses on issues having more to do with construct validity (Messick, 1989).

We now review some of the technical challenges confronting ILSAs, using the extended Messick framework to organize the discussion. At the same time, it is important to recognize that challenges represent not only problems, but also opportunities. By appropriately meeting these challenges, ILSAs can continue to traverse the virtuous spiral (Figure 20.2), developing in ways that enhance their utility and impact. We begin by considering direct threats to comparability and interpretability.

Threats to Comparability and Interpretability

One threat, as noted previously, is related to the likelihood that ILSAs will have to maintain dual systems (paper-and-pencil administration and an ever-evolving digitally-based assessment) for some time—both because of insufficient technical capacity in many newly participating countries and because segments of some target populations (e.g., older adults) may not be capable of responding with tablets or other digital devices. To this point, through statistical adjustments based on estimated mode effects, ILSAs have maintained a single reporting scale for results obtained through either paper-and-pencil administration or digitally-based assessments. The construction of a single scale implies that the two sets of scores are linked to substantially the same construct and support similar interpretations.

Over time, however, ILSAs employing a digital platform will continue to innovate by extending the number and scope of construct facets to reflect the broadening of legacy constructs or the introduction of new constructs that capture the ways in which students and adults interact with new technologies. To some extent, these changes will involve new item types and new response formats that leverage the affordances of the technology platform. This will lead to a growing divergence between the paper-and-pencil administration and digitally-based assessment instruments, thereby making the assumption of the comparability of the scores obtained in the two modes progressively less tenable. Consequently, interpreting the scores (and score differences between countries) without regard to mode will become very problematic—especially if the proportions of respondents in the two modes differ substantially across countries.

We can expect that, eventually, separate scales will be required, with the attendant complications for reporting and interpretation. Moreover, the additional cost of maintaining two different platforms is not trivial. For example, one continuing challenge concerns transitioning paper-and-pencil administration trend items to digitally-based assessments.[50] This requires careful adaptation to keep the information provided by the item in the two formats comparable (Lennon & Kirsch, 2025; von Davier, Khorramdel, et al., 2019).[51] Finally, within the digitally-based assessment orbit, the proliferation of different delivery devices over time will also pose a challenge to comparability both within and across countries.

A second threat concerns the shift to adaptive multistage testing. This move necessitates significant increases in the sizes of the item pools, particularly with the enhancement of construct representation. In conjunction with the growth in the number of participating countries, these increases place greater demands on the capacity of ILSA teams, not only to develop sufficient numbers of items with desired characteristics, but also to carry out the operations of translation and adaptation with the high degree of quality essential for comparability. Furthermore, greater numbers of items to be calibrated necessitate increases in the sample sizes required and/or the number of items administered to each respondent during field trials.

As documented in the section "Transitioning to Digitally-Based Assessments," ongoing innovations in tools and processes are required to meet these demands, all the while respecting severe constraints on cost and time. Furthermore, adaptive multistage testing substantially increases the complexity of assessment design and, consequently, heightens the need to develop new quality control procedures to ensure that the overall design requirements are properly satisfied. In this regard, automated test assembly software that can be implemented for adaptive testing should prove very useful (Luo, 2020).

Third, the introduction of innovative item types may not only drive a divergence between modes of administration, but also introduce nonnegligible construct-irrelevant variance due to differential familiarity with response formats within and especially across countries.[52] Whether a nonnegligible fraction of the observed variance is deemed construct relevant or construct irrelevant bears on how to address the issue of measurement invariance (see the section "Measurement Invariance").

Finally, an ongoing concern will be ensuring an adequate level of measurement precision and comparability of the scores in the cognitive domains in the face of increasing heterogeneity among participating countries with respect to both cognitive skills and relevant background knowledge.[53] L. Rutkowski and Rutkowski (2018) discussed a relevant paper-and-pencil administration strategy, termed *design based*, that involves creating booklets with different average levels of difficulty and adjusting the proportions of relatively "easy" and relatively "hard" booklets administered in a country based on an a priori estimate of the proficiency distribution in that country. Indeed, such a group-level, adaptive strategy has been applied on a limited basis in PISA 2009 and PIRLS 2021.

Questions of comparability also pertain to the scales developed from responses to the background questionnaire. The challenges are greater in this setting than with the cognitive scales inasmuch as the number of items contributing to each background questionnaire–derived scale is very small. Consequently, measurement error is substantially greater. Furthermore, in some countries, scores on the reading/literacy assessment indicate that many (if not most) of the students may have some difficulty understanding the questions and responding appropriately.

L. Rutkowski and Rutkowski (2018) also addressed issues of comparability of the scales derived from the background questionnaire. They described a strategy, termed *model based*, that involves somewhat relaxing the measurement invariance requirements (to partial invariance) and encouraging countries to take advantage of the so-called national option to augment the core items for select background questionnaire constructs with items of specific interest to the country (or to a group of countries). The intent is to enhance interpretability and utility. This strategy was demonstrated using the WEALTH scale of PISA 2012. The exercise illustrated the challenge in achieving cross-national comparability even with a group of similar countries.[54] At the same time, there is a trade-off with comparability across all participating countries. For further empirical analysis, see, for example, He et al. (2019).

Van de Vijver (2018) considered the variation in response styles across countries and cultures particularly problematic with the Likert scale response formats used in many background questionnaire items. In this regard, Kyllonen and Bertling (2013) described some of the problems that arise with items that employ typical rating scale response formats and investigate some alternative response formats. Some formats tend to yield results that exhibit the aptitude–achievement paradox.[55] This is an area of ongoing research (von Davier et al., 2018). A comprehensive treatment of response format issues in an international context can be found in Kuger et al. (2016). A more general discussion of the validity issues related to items requiring self-report can be found in Karabenick et al. (2007).

As noted in the discussion of translation and adaptation procedures (see the section "Translation/Adaptation/Verification of the Instruments"), numerous efforts exist to enhance cross-cultural equivalence in the measurement of the target constructs of the background questionnaire in the face of greater heterogeneity among participating countries. Nonetheless, credible concerns remain (L. Rutkowski & Rutkowski, 2019). In addition to bias, another consequence is that some countries experience higher levels of measurement error in the scales derived from the background questionnaire that, in turn, attenuate the estimates of the relationships between cognitive skills and the constructs underlying those scales. These issues, singly and in combination, can substantially reduce the amount of useful information generated for some countries.

Notably, cross-national differences in measurement error in background questionnaire scales can result in spurious differences in correlations between such scales and cognitive domain scores. An important question, then, is to what extent observed heterogeneity across countries is due to such statistical artifacts. Gurkan (2021) developed an approach to correcting the bias in correlational estimates that takes account of both the multilevel structure of ILSA data and differential measurement error. The method further posits that the tested population comprises two latent classes with different levels of domain proficiency. When applied to PISA 2015 data to examine the relationships between mathematics proficiency and mathematics self-efficacy, the method resulted in substantially reduced correlational heterogeneity across countries, especially for lower performing countries.

A fifth threat, primarily to school-based surveys, is due to the variation among countries in the ways they identify students with different types and degrees of disability, their rules for exclusion from the survey, and the nature of the accommodations employed during survey administration (see the section "Accommodations for Testing"). On the one hand, some of the variation arises from different national customs and regulations and is difficult to mitigate. Certain subpopulations, such as recent immigrants, may also be treated differently across countries, again resulting in some lessening of comparability. On the other hand, some sources of between-country variation may be susceptible to harmonization, that is, coordinating definitions of factors such as education levels and socioeconomic status. Efforts to that effect in the social science literature

are ongoing, with the goal of increasing comparability of patterns of relationships across countries.

A sixth threat arises from the proliferation of mandated, official school surveys, as well as school-based research studies. This has led to "survey congestion," resulting in lower participation rates. Although replacement schools can be located and convinced to participate, the overall trend is worrisome because, over time, there can be greater divergence in sample quality among countries. In this regard, Durrant and Schnepf (2017) linked PISA sample data for England to two large-scale administrative databases with information on students' socioeconomic backgrounds and performance on national public examinations. They were able to characterize both the schools and the students most likely not to respond and, on that basis, made suggestions on how best to select substitute schools. This approach shows some promise if it could be implemented in countries where such linkages are feasible.

In the case of household surveys, such as PIAAC, reduced participation rates may be due to both survey congestion and a general decrease in trust in central authorities. Here again, increasing divergence in sample quality threatens country-level comparability. Lowered participation rates may necessitate offering some incentives on a broader scale.[56] The COVID-19 pandemic and its aftermath may well exacerbate these challenges. It is possible that in the future some forms of remote assessment will offer a partial solution.

Even among those who agree to participate, there are worrisome trends in motivation. Here the concern is greater with school-based surveys than with household surveys. In low- or no-stakes situations, many students do not engage fully with the instrument—or only do so for part of the assessment. The consequence is less accurate measurement at the population and subpopulation levels, as well as reduced comparability, because diminished motivation will vary in prevalence across subpopulations within and across countries. For an estimate of the impact of differential motivation on NAEP results, see Braun et al. (2011). All these threats to comparability serve to reduce the policy utility of international large-scale assessments. Recent approaches to analyzing item-level data for engagement may prove useful in this context (Ulitzsch et al., 2019).

In the future, log file data should prove to be of use. For example, timing data are already helping in the effort to classify items at the respondent level as either omitted or not reached. Researchers have begun the development of indicators based on the raw log file data that can be used to signal possibly aberrant response processes or patterns suggesting lack of effort (e.g., Goldhammer et al., 2016; Pokropek, 2016a; Pools & Monseur, 2021; Wise, 2020). One strategy would be to develop norms for an indicator (possibly at a national level) and to identify individuals with indicator values that fall in the extremes of the distribution. If that is the case, say, for two or more indicators, then that individual would be removed before scaling and analysis. Of course, empirical work would be necessary to validate a decision rule (Soland et al., 2021). In any case, such indicators could be used to monitor system performance and enhance sample quality.

Another possibility is to employ records of participant response processes to improve item design and to contribute evidence of validity (Ercikan & Pellegrino, 2017; Zumbo & Hubley, 2017). Although these seem to be promising directions, many technical, practical, and even ethical questions remain to be explored (Provasnik, 2021). Technical advances and considerable infrastructure development will be required to make these and other possible applications of log file data practicable. Ethical issues should be considered by broad-based groups of stakeholders, with sufficient international representation, although the different perspectives may well make reaching consensus challenging in its own right.

Finally, comparability and interpretability of ILSA results also depend on whether respondents differentially interpret the tasks that are presented and how those interpretations shape their response behaviors. Assessment practitioners are coming to understand the importance of adopting a sociocultural perspective in the design of assessments, as well as in the interpretation of the data generated by those assessments (Mislevy, 2018). In the ILSA context, such considerations are germane to both the cognitive instruments and the background questionnaire. One implication is that the committees that develop these instruments should have broader participation from the different subpopulations in the various countries to enhance the meaningfulness of the questions to all respondents, thereby reducing construct-irrelevant variance.

Although the logistics of carrying out such a program in an international context are formidable, ILSAs sponsors and contractors do make an effort to include representatives from different countries as subject matter experts, to attend test development workshops, and to contribute items for field testing. Evidently, more needs to be done in this regard.

Relatedly, some critics have suggested that ILSA teams should be open to employing different methods for generating validity evidence. For example, Pepper (2020) pointed out that the validity argument would benefit from conducting cognitive laboratories with both the cognitive instruments and the background questionnaire items.[57] In an international setting it would be impractical to do so in every participating country. Consequently, some experimental design would have to be employed to generate evidence reasonably representative of the ensemble of participating countries. Classical convergent and divergent validity studies could also be conducted in selected settings. In any case, more thought should be devoted to these and other approaches to obtaining additional evidence. Similar considerations apply to studying the impact on the performance of certain subpopulations with the introduction of new digital devices.

Threats to Relevance

ILSAs also face challenges with regard to relevance. For the most part, the principal strategy to enhancing relevance has been to increase the range of constructs that are measured through the cognitive instruments and the background questionnaire. However, there are technical limitations to pursuing this strategy. Moreover, each cycle brings novel issues to the fore.

One aspect of relevance is the degree to which ILSA assessments reflect the increasing use of different technologies by students and adults. For example, there is ongoing interest in incorporating various 21st-century skills in school curricula (Binkley et al., 2012; Darling-Hammond, 2012). The assessment of such skills often requires complex stimulus materials and student-constructed responses that currently require human raters for evaluation. Including such (or similar) assessments in ILSAs poses a number of challenges, as has already been mentioned. Nonetheless, to maintain relevance, expanding the range of assessment formats remains a high priority for ILSA teams.

Fortunately, the advent of digitally-based assessments makes possible new item types with different response formats to enhance construct representation. As an example, simulation tasks are becoming increasingly popular in school-based ILSAs because they allow for a more authentic assessment in the domains of science inquiry, mathematics, collaborative problem-solving, and even financial literacy (PISA). In the case of science, simulations can provide students with an opportunity to design an experiment (by selecting variables and associated values) and then run the experiment to generate data in order to respond to various questions. Because such tasks can consume considerable time, their appearance in an ILSA may be more common in specialized auxiliary studies.

With the introduction of adaptive multistage testing, it is optimal to have decisions on the next module to be administered based on the maximum amount of information. Currently, information is only available from selected response items and a limited number of other item types that can be automatically scored. Going forward, new item types should be developed and evaluated in conjunction with the corresponding scoring algorithms. This shift requires item designers to collaborate with computer scientists and other specialists to determine the feasibility of developing accurate and efficient scoring algorithms. In some cases, original item designs will have to be modified with a commitment to maintaining construct relevance. Evidently, implementing automated scoring in an international context with multiple languages and scoring guides is an order of magnitude more complicated than in national assessments. Nonetheless, the implementation of adaptive testing in response to demands for greater accuracy along the full score scale, together with the necessary concomitant technical developments, is yet another instance of the virtuous spiral (Figure 20.2) that captures the ongoing dynamic between stakeholders on the one hand and ILSA sponsors and teams on the other.

On another front, ILSAs must be responsive to changes in the school, work, or other environments where the easy availability of online tools (e.g., for checking spelling, grammatical construction, argumentation, information search and display) has the potential to modify the focal construct. Maintaining both relevance and scale comparability across administrations may be increasingly at odds in both student and adult surveys. As Mazzeo and von Davier (2014) pointed out,

> The question is whether surface characteristics that change quickly due to technological advances will lead to changes in the requirements of underlying skills and knowledge. If students increasingly use technology in everyday activities, and if

> these technologies become easier to apply to everyday problems over time, then the traditional concept of linking assessments over time by means of tasks that stay the same and look the same becomes unsuitable. (p. 255)

This poses a challenge not only to instrument developers and psychometricians, but also to those charged with renovating the digital platform. Investments in continuous innovation in platforms, processes, and procedures will consume substantial resources. More fundamentally, technology-driven changes in the context of measurement can force shifts in how the construct is defined and operationalized. As a result, what may have once been viewed as a source of construct-irrelevant variance may come to be accepted as contributing construct-relevant variance. An example is provided by the construct problem-solving in technology-rich environments, introduced in PIAAC (2012), where the ability to employ the digital tools provided in simulated web, email, and spreadsheet environments is integral to the construct. We can expect that other examples will emerge as different technologies become increasingly embedded in everyday life.

As the number of measured constructs increases in conjunction with the possible introduction of rotated background questionnaire designs, the stability of the key latent regression model may be reduced. This, in turn, would threaten the accuracy of the population model used to generate PVs—especially in the case of sparse designs for the cognitive instruments. Present concerns regarding the impact on proficiency estimates of measurement error in the explanatory variables in the latent regression model will likely become more pronounced (L. Rutkowski & Zhou, 2015). Presumably, at some point there will have to be some trade-offs between relevance and maintenance of the reputation of ILSAs as trustworthy sources of achievement data. Other trade-offs between local fidelity and international comparability will be informed by solid analytic work but will also necessarily reflect policy and political considerations. Consequently, the decisions reached may place additional burdens on the ILSA teams, which will have to accommodate greater variation among the instruments administered in different countries.

Another critical aspect of relevance concerns the degree of alignment between the information provided by the ILSAs and the questions asked by policy makers and other stakeholders. In point of fact, a number of key issues that command some degree of general interest are frequently raised. To improve timeliness, the relevant data could be culled from the comprehensive almanacs and organized and displayed relatively quickly in tables and figures. Further on, targeted secondary analyses could be funded through sponsored initiatives—perhaps initiated even before the public release of the data.

Relevance could be increased by more systematically linking national ILSA data to national databases that contain other types of information (e.g., background data at the individual level). Such an augmented database would support a broader range of secondary analyses, as noted previously in the case of Denmark. Linkages to data at higher levels of aggregation (e.g., at the school or area level) could also prove useful for some analyses. Going forward, designers could build in some connections as a national option to facilitate such linkages.

POLICY AND POLITICAL CHALLENGES

Alongside the technical concerns that relate primarily to construct validity, we now turn to a discussion of some of the political challenges that can be framed in terms of consequential validity. The increased prominence of ILSAs results in policy discussions, as well as the exhortations on the part of some leaders that countries emulate the "high-flyers," has sparked a backlash. Some researchers and commentators have argued that ILSA results have become too influential in national policy discussions and, in particular, that they are a force for "international homogenization" at the expense of national educational differences that should be preserved (Carnoy, 2015; Grek, 2009; Meyer & Benavot, 2013; Sellar & Lingard, 2013). They cited the broader participation in ILSAs, the increased salience of ILSA results in discussions of education policy, and the utilization of ILSA outcomes for purposes of evaluation and monitoring.

Benavot and Smith (2020) linked increasing participation in ILSAs in part to efforts by the UNESCO Institute for Statistics to develop global learning metrics as a means for quantifying progress toward the United Nations' Sustainable Development Goal 4: Education. By participating in an international comparative assessment, countries can satisfy the SDG 4 reporting requirements. The authors argue that these requirements have also led to pressure to conform to a system of "global educational governance." Teacher associations have also weighed in on ILSA participation (Couture, 2016).

A review of the literature, however, suggests that the situation is rather more ambiguous and that a more nuanced view is in order. Successful policy transfer across national boundaries can be difficult—for both technical and political reasons (Atkin & Black, 1997; Braun, 2008; Burdett & O'Donnell, 2016; Oliveri et al., 2018). Drawing on extensive personal experience, Ritzen (2013) observed that the degree of utilization of ILSA results depends not only on a country's readiness (politically and otherwise) to institute changes, but also on its capacity and commitment to carry out the change process. This observation may be especially true for many low- and middle-income countries, whose participation has been essentially mandated (and supported by) international donor organizations (see also Braun, 2013).

Volante et al. (2017), in discussing the normative impact of the assessments sponsored by the OECD, also noted that there is a wide range of policy responses to ILSA results—from negligible to modest to very substantial. Fischman et al. (2019) conducted surveys of ILSA experts and interested stakeholders. They found considerable diversity of opinion regarding whether ILSAs had an impact on national education policymaking, and if so, whether it was constructive. The observed variation is likely due both to differences in experiences among countries and to differences in the vantage points of the respondents.

These debates can be framed in terms of questions regarding the consequential validity of ILSAs. That is, beyond the attribution of impact (see the section "Impact"), the principal issue lies in what cases the impact leads to long-term positive outcomes and

in what cases it does not—and whether the two can be reasonably well distinguished. This section addresses the issue.

Concerns Regarding Impact

Once a country joins an ILSA collaborative, an obvious benefit of continued participation is the value of tracking performance trends over time. It is impossible to ignore the "league tables" that are published after each ILSA data release. However, a focus on rankings based on aggregate mean scores, or changes in rankings over time, can be misleading and lead to confusion among stakeholders. First, countries with mean scores that are statistically close nonetheless could have quite different rankings. Second, many factors determine a country's rank in a particular cycle (e.g., the set of countries participating in that cycle) that have little or nothing to do with the efficacy of its education system (Singer & Braun, 2018).

Nonetheless, it is true that some ILSA proponents argue that a country's best strategy for educational improvement is to emulate the policies of so-called high flyers (National Center for Education and the Economy, 2020; Schleicher, 2018). Kamens (2013) argued that this sort of cheerleading encourages stakeholders to think that there is a "magic bullet" to achieving success. The problem is sometimes exacerbated by misleading media reports that may focus on changes in a country's rank but ignore the actual change in performance. PISA 2015 offers some examples from East Asia. In Japan, one newspaper highlighted a change in rank from fourth to second on science, even though the mean score declined from 547 to 538. By contrast, in Chinese Taipei (Taiwan), a newspaper bemoaned a decline in ranking on reading, even though the mean score increased.[58] To counter these and other misinterpretations, there have been some innovative attempts to present country results through visuals that de-emphasize league tables.

Some authors have argued that for the United States, state-level comparisons are likely to be more informative than international comparisons (Carnoy et al., 2015).[59] For now, that appears to be a minority view. Indeed, these same authors suggested that cross-country comparisons among recognizable subgroups are more informative than aggregate comparisons. For example, they compared countries' performance in mathematics on PISA 2012 within strata defined by a cross-national measure of family academic resources. They showed that although the United States fares poorly overall, it does relatively well in each stratum. Its comparatively lower aggregate scores are due to the greater proportions of U.S. students in the more disadvantaged strata. This finding leads to consideration of different reform strategies than might be suggested by a singular focus on overall rankings. Similar studies based on data from PIAAC could also yield interesting patterns.

Another widespread concern is related to misinterpretations and misuses that are not directly due to technical deficiencies of the assessment or the failure to provide appropriate guidelines (Kane, 2016). That is, despite the best efforts of sponsors and developers, either policy makers make inappropriate policy pronouncements or decisions or various entities overstate the evidential value of ILSA results with respect to policy and

practice. For example, in some instances, national education goals have been framed in terms of specific improvements in country rankings. D. Rutkowski et al. (2020) cited the example of Australia. The Education Act of 2012, approved by the Parliament of the Commonwealth of Australia, called for Australia to be ranked by PISA 2025 in the "top five" on reading, mathematics, and science. A misguided focus on the rankings competition can lead to suboptimal policy choices.

On the one hand, although they are not directly responsible for such misuses, it is still incumbent on ILSA sponsors and developers to counter specific misuses and misinterpretations of ILSA data. On the other hand, a call to reduce the gaps in scores between high and low performers within a country to a level comparable to the gaps seen in peers with comparable overall performance but greater equity should be seen as a constructive use of ILSA data for purposes of benchmarking.

Responding to Political Challenges

Judgments regarding whether ILSAs exert undue influence on national education policies should take into account the country's rationale for participation. As noted in the section "Impact,", ILSA results are sometimes used to justify or legitimate pending policies. Wiseman (2013) offered the example of France. Citing Dobbins and Martens (2012), he stated that "the OECD's PISA was part of a French political agenda to create an evidence base to support a specific policy position. This is not an isolated tactic or unusual use of international achievement studies" (Wiseman, 2013, p. 311). J. Jennings noted that in the United States, the National Council of State Legislators in 2016 seized on the then–most recent PISA results to build support for a wide range of education reforms.[60] In Ireland, surprisingly weak PISA results provided impetus for the passage of long-stalled education reforms (H. Hislop, personal communication, September 21, 20l7). Asserting that these represent constructive uses of ILSA data depends largely on a judgment of the appropriateness of the reform policies.

Wiseman (2013) also argued that, to the extent that ILSAs exert influence on educational policies, the outcome is not necessarily some form of strict policy convergence. Instead, the result may be an "isomorph" of the "model" system, whereby countries extract some key features of the model (e.g., curriculum) but then modify and/or adapt them to better fit the country's traditions and culture, often taking into account the political strengths of different stakeholders. Thus, despite exhortations in support of emulation, the path from ILSA results to policy responses is neither straightforward nor predictable (Pons, 2017).

Ideally, a country's decision to participate in a particular, school-based ILSA would be informed by careful study of the ILSA's avowed purposes and its assessment frameworks, followed by an evaluative judgment regarding their degree of congruence with the country's goals for its students (Oliveri et al., 2018). Presumably, the reputation of the ILSA for generating data that are accurate, reliable, and valid would also factor into the decision. In this context, careful study of a country's outcomes could very well appropriately result in changes in curriculum and pedagogy. As described in the section "Impact,", this strategy has been adopted by many countries—in some cases

(e.g., Singapore) in a thoughtful, systematic fashion. Despite the caveats raised in that section, the strategy may be particularly helpful to countries with lower capacity in the realms of curricular design, test development, and educational measurement. This sort of "homogenization" is consistent with Wiseman's (2013) model of convergence.

It bears repeating that country differences in ILSA-based indicators cannot be directly linked to differences in the quality of their education systems (Singer et al., 2018). More nuanced contextual analysis is in order. For example, the strong performances of students in South Korea and Japan are due, in part, to the pervasive practice of out-of-school tutoring (shadow education), as noted by Heyneman (2013) and others. Countries also differ in their investments in both the well-being of children and teacher quality. These and other differences can contribute to differences in academic achievement, particularly for children living in less advantaged circumstances. A failure to appreciate this complexity can lead to misinterpretations of ILSA results (particularly changes in overall rankings) and, hence, to misguided policy decisions. It is worth noting that such differences—and their misinterpretations—can also occur when making comparisons across population subgroups or geographical regions within a country.

Despite the appropriate concerns regarding the overemphasis on league tables, rankings on other measures can provide useful policy information. For example, one country-level indicator of equity is the difference in the mean scores between the top decile (quintile) and the bottom decile (quintile), with larger differences signaling greater inequality. A related indicator is the slope of the regression of cognitive scores on a measure of socioeconomic status. Again, a steeper slope signals greater inequality. It would be useful to consider disaggregating cognitive outcomes data by other background variables to provide more informative comparisons (Rowley et al., 2020). In school-based ILSAs, for example, reporting results by levels of opportunity to learn (or available proxies) should be considered. In reporting results to their national policy makers, countries should consider presenting a range of indicators. It is particularly striking when countries with similar overall proficiency means differ substantially on one or the other of these indicators. Such reports must include clear, explanatory text to highlight the utility of these indicators and to reduce the possibility of misinterpretations.

To highlight the benefits of participation in school-based ILSAs, sponsors, in conjunction with country representatives and education authorities, should enhance outreach to teachers, curriculum specialists, and education leaders to explicate the ILSA frameworks and how the assessments reflect those frameworks. They could also organize specially designed workshops or webinars that highlight important issues or findings that might help guide interpretations. With a solid understanding of how to interpret the outcomes, the classroom utility of both the frameworks and the disaggregated results would be more evident. Country representatives who participate in the various expert groups during an ILSA cycle are an underutilized resource. They not only help to define and operationalize the various frameworks, but also contribute to the development of described proficiency scales. Thus, they are well positioned to inform education policy discussions within a country.

Similar considerations are relevant to participation in adult surveys. Outreach to various government ministries and other stakeholders could emphasize the value of the data collected by PIAAC for policy decisions related to a range of domains, including education, training, and labor market initiatives. Moreover, the argument can be strengthened by reference to the many secondary analyses already conducted on available data.[61] The results of these analyses enable useful cross-national comparisons that can directly inform policy discussions.

In general, the "duty of care" incumbent on ILSA sponsors and developers to address concerns regarding consequential validity can be framed in terms of ensuring, to the extent possible, that (a) the technical quality of the results supports their intended interpretations and uses; (b) the presentation of the results, along with the corresponding explanatory materials, encourages appropriate interpretations and uses; (c) the text is transparent about possible biases and uncertainties, as well as their implications for interpretation; and (d) the most common, anticipated misinterpretations and misuses are proactively addressed and discouraged. This aspect of validity is all the more important with the increases in ILSA complexity and heterogeneity of participating countries.

CONCLUSION

The inclusion of this chapter in the present volume reflects the growing importance of ILSAs of student and adult populations since the 1960s. From modest beginnings as a set of exploratory studies devised by a small group of scholars interested in international comparative education, school-based ILSAs have grown into a major strand of applied research within the larger domain of comparative education (Suter et al., 2019). As the scope and reach of ILSAs have expanded, they have generated an unprecedented amount of empirical evidence on which researchers and policy makers can draw to analyze education systems and develop strategies to improve student outcomes. Nonetheless, some worry that by virtue of their dominant role, ILSAs have led to a diminished appreciation for the insights that other strands of comparative education have to offer—not the least of which is providing a broader framework within which ILSA findings can be understood and applied (Carnoy, 2019). Indeed, as is evident in the contributions to Suter et al. (2019), these other strands rest on foundations of both theory and rich, empirical studies.

At the same time, the increased attention paid to ILSA scores and rankings, as well as to the results of secondary analyses of these data, is due in large measure to the widespread recognition of the importance of understanding the relationships of education and skills, not only to social and economic development, but also to general well-being.[62] As globalization and technology have accelerated, the economic interdependencies (and competition) among countries, policy makers, and key stakeholders have seen value in benchmarking performance against peers and near-peers—initially in more developed economies and, more recently, in middle- and lower income countries.

In 1987, following the implementation of the new NAEP design, Messick published a paper in which he asserted that because large-scale assessments constitute a form of policy research, their success should be judged based on their policy utility. Further, he reasoned that an element of uncertainty exists between policy research and policy formation, creating a gap that requires informed judgment. He believed that large-scale assessments could provide information to help bridge this gap through the provision of evidence gathered by careful design and development of instruments, followed by appropriate analysis and reporting of results (Kirsch & Braun, 2020; Lennon & Kirsch, 2025). He did not assert, however, that this empirical information was sufficient to the purpose.

To help understand how large-scale assessments could best fulfill this role, Messick (1987) proposed a framework containing three key design criteria: comparability, interpretability, and relevance. This chapter has presented and extended these three design criteria, arguing that they remain applicable for evaluating the utility of next-generation ILSAs. Indeed, a recurring theme in the chapter has been how changes to ILSAs have contributed to greater utility by strengthening one or more of the design criteria.

For example, to enhance relevance, ILSAs are responding to broad interest in new, more innovative domains such as creative thinking, problem-solving, and others that are most suitable for digitally based assessments. However, as always, there are trade-offs. In this case, countries or individuals who receive the paper-based assessments do not participate in these innovative domains, potentially reducing the utility of these surveys for these countries and groups. Moreover, the evolution of ILSAs in this direction will result in a growing divergence between the two modes in what is being measured and, consequently, weakening the justification for reporting results on a single scale. Unfortunately, both practical and political considerations argue for maintaining two delivery modes, with the attendant complications and additional costs.

In addition to employing Messick's (1987) framework to examine various aspects of ILSA development and subsequent uses, this chapter also has described the ILSA developmental path as a virtuous spiral, that is, a trajectory in which each cycle of an ILSA attempts to meet the evolving interests of policy makers and key stakeholders who continually challenge sponsors and researchers to generate information that is relevant to the changes taking place within and across participating countries. As societies undergo changes, policy makers and other stakeholders pose new questions that lead to both novel and better measures of legacy constructs and the introduction of new domains of assessment.

In this context, we highlighted two key inflection points along this trajectory that have led to marked increases in ILSA utility and salience. The first inflection point occurred in the mid-1980s through the mid-1990s with the introduction of important methodological innovations such as BIB spiraling and IRT scaling. These innovations provided policy makers with interpretable information regarding the distributions of skills within and across countries, along with the ability to examine

the relationships among these skills and a host of demographic and background variables.

Over the next 2 decades, additional methodological and technological innovations ultimately led to the second inflection point, one characterized by the advent of digital platforms that support the design, development, and delivery of next-generation ILSAs. Successfully carrying out digitally-based assessments required the construction of novel tools and the integration of new workflows and processes that benefited from the ongoing collaboration of experts across the different phases. The digital platform facilitated the introduction of new constructs, new item types, and improved methodologies, such as adaptive multistage testing. The result has been an improvement in the overall quality of the data, extensions of what can be measured in both student and adult populations, and greater efficiency in the management, delivery, and processing of the data.

As ILSAs have proven over the years to be a reliable source of credible information, this chapter also described some of the ways in which these assessments have impacted policy. Examples include countries where new policies were implemented based on results from in-school surveys including PIRLS, TIMSS, and PISA. There are other cases where these assessments were used to support existing policies or legitimate proposed policies. Beyond their direct impact, there is also evidence of ILSAs' indirect impact, which is reflected in increased media attention and the documented growth in the number of researchers who are using ILSA databases to inform their research and develop policy papers covering a range of issues relevant to school-age and adult populations. Finally, ILSA participation has jump-started or accelerated capacity building, particularly in many middle- and low-income countries, giving them the capability to conduct national or subnational assessments to inform education policy.

Not surprisingly, the growth in relevance and popularity of ILSAs has also led to a number of challenges and concerns. One challenge arises from the substantial increase in the diversity of participating countries, as well as that of social and economic backgrounds, representing a wider range of skills. Maintaining an adequate level of measurement invariance across countries among both the cognitive domains and the constructs underlying the background questionnaires has become increasingly difficult as participation has expanded.

Compounding the challenges presented by the growing diversity among participating countries is the never-ending pressure to extend what is being assessed. Because technology-based assessments facilitate such extensions, there is continued demand to add to the constructs in the background questionnaire, as well as those in the cognitive domains. Given the constraints on survey administration time, these extensions necessitate the development and deployment of more complex designs, such as within-construct rotation of key variables in the background questionnaire—similar to what is done in the cognitive domains.

Predictably, increases in ILSA scope and coverage lead to rising costs for both contractors and participating countries. Greater operational difficulties due to lower

response rates, as well as to the hiring, training, and retaining of survey staff, also contribute to rising costs. These increases cause concerns among the governments who not only fund the assessments, but also must wait patiently for several years from the initiation of survey development to the final reporting of results.

More recently, the questions that have been raised are of a more political nature and have less to do with technical issues. For example, there are widespread concerns regarding the overemphasis on the league tables of country rankings and their relative changes over time. This attention to league tables is due in part to the (misguided) notion that differences in countries' rankings are a credible indicator of differences in the efficacies of their education systems. A natural conclusion, then, is that a surefire strategy recognized for educational improvement involves trying to identify and implement those policies that seem to be effective in high-performing countries (Tucker, 2011).

Aside from technical issues related to data quality (as was the case with Shanghai),[63] the difficulty with this advice is that it fails to acknowledge the natural limitations of ILSA data; namely, they can suggest interesting hypotheses but cannot establish them, largely, but not solely, because of the constraints on making causal inferences from cross-sectional data. More intensive studies are needed, requiring consideration of the broad range of factors operating at different levels from conception to the time of the assessment that contribute to student achievement. It is precisely here that the findings from "traditional" comparative education come into their own. Understanding the historical, cultural, economic, and political influences on the structure and functioning of current education systems—and how they differ across countries—is essential to devising workable policies and interventions that mitigate some of the likely difficulties that arise in a policy transfer process.

As noted in the last section, consideration of ILSA impact (both positive and negative) falls under the rubric of consequential validity. The responsibilities associated with consequential validity lie with both the sponsors and the developers, as well as with those who use and interpret the data. The sponsors and developers are expected to, and often do, provide information regarding the overall quality of the data, along with the appropriate uses and interpretations of the results. The media, policy makers, and others are expected to assume some responsibility for misinterpretations or overstatements of what the data say or what decisions can be supported by the results.

These challenges and concerns, both technical and political in nature, constitute healthy tensions that can and should be discussed and debated within the extended Messick (1987) framework. Although it is certain that ILSAs will continue to face a number of challenges, we believe that the path along the virtuous spiral will continue to provide innovations that will enable ILSAs to address new and more complex questions. We expect that answers to these questions will lead to continued appreciation for the utility for such surveys.

ACKNOWLEDGMENTS

The authors wish to thank the editors, L. Cook and M. Pitoniak, for inviting them to contribute a chapter to the volume and for their support throughout the process. Reviewers of our first draft, A. Ben Simon, M. von Davier, and H. Wagemaker, offered a number of useful comments and suggestions that we endeavored to implement. Subsequent drafts were reviewed critically by M. L. Lennon and E. Gonzalez. Both devoted substantial time to the reviews and we benefited tremendously from their many suggestions, almost all of which we were able to follow. On specific matters we were assisted by B. Fishbein, P. Foy, M. O. Martin, L. Suter, and K. Yamamoto. Finally, A. K. Gontz was absolutely thorough in conducting the final editing of the manuscript. To all we are very, very grateful.

REFERENCES

Ainley, M., & Ainley, J. (2019). Non-cognitive attributes: Measurement and meaning. In L. E. Suter, E. Smith, & B. D. Denman (Eds.), *The SAGE handbook of comparative education studies* (pp. 103–125). SAGE Publications. https://doi.org/10.4135/9781526470379.N7

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

Atkin, J. M., & Black, P. (1997). Policy perils of international assessments. *Phi Delta Kappan, 79*(1), 22–28.

Avvisati, F. (2020). The measure of socio-economic status in PISA: A review and some suggested improvements. *Large-Scale Assessments in Education, 8*, Article 8. https://doi.org/10.1186/s40536-020-00086-x

Baird, J.-A., Johnson, S., Hopfenbeck, T. N., Isaacs, T., Sprague, T., Stobart, G., & Yu, G. (2016). On the supranational spell of PISA in policy. *Educational Research, 58*(2), 121–138. https://doi.org/10.1080/00131881.2016.1165410

Beaton, A. E., & Barone, J. L. (2017). Large-scale group score assessments. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 233–284). Springer Open. https://link.springer.com/chapter/10.1007/978-3-319-58689-2_8

Bedard, K., & Dhuey, E. (2006). The persistence of early childhood maturity: International evidence of long-run age effects. *The Quarterly Journal of Economics, 121*(4), 1437–1472. https://doi.org/10.1093/qje/121.4.1437

Benavot, A., & Smith, W. (2020). Reshaping quality and equity: Global learning metrics as a ready-made solution to a manufactured crisis. In A. Wulff (Ed.), *Grading goal four: A strategic take on the tensions, threats and opportunities in the sustainable development goal on quality education* (pp. 238–261). Brill. https://www.research.ed.ac.uk/files/154046014/Reshaping_quality_and_equity_GLMs_as_a_ready_made_solution_Benavot_Smith_2020.pdf

Bergner, Y., & von Davier, A. A. (2019). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics, 44*(6), 706–732. https://doi.org/10.3102/1076998618784700

Biemer, P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lysberg, L. E., Tucker, N. C., & West, B. T. (Eds.). (2017). *Total survey error in practice*. John Wiley & Sons. https://doi.org/10.1002/9781119041702

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Springer. https://doi.org/10.1007/978-94-007-2324-5_2

Bracey, G. W. (2008, Fall). Disastrous legacy: Aftermath of a nation at risk. *Dissent Magazine*. https://www.dissentmagazine.org/article/disastrous-legacy-aftermath-of-a-nation-at-risk

Bradburn, N. M., & Gilford, D. M. (Eds.). (1990). *A framework and principles for international comparative studies in education*. National Academies Press. https://doi.org/10.17226/9220

Braun, H. (2008). Review of McKinsey report: How the world's best performing school systems come out on top. *Journal of Educational Change, 9*(3), 317–320. https://doi.org/10.1007/s10833-008-9075-9

Braun, H. (2013). Prospects for the future: A framework and discussion of directions for the next generation of international large-scale assessments. In M. von Davier, I. Kirsch, K. Yamamoto, & E. Gonzalez (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy and educational research* (pp. 149–160). Springer Science and Business Media. https://doi.org/10.1007/978-94-007-4629-9_8

Braun, H. (2018). How long is the shadow? The relationships of family background to selected adult outcomes: Results from PIAAC. *Large-Scale Assessments in Education, 6*, Article 4. https://doi.org/10.1186/s40536-018-0058-x

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th grade NAEP reading assessment. *Teachers College Record, 113*(11), 2309–2344. https://doi.org/10.1177/016146811111301101

Braun, H. I., & Singer, J. D. (2019). Assessment for monitoring educational systems: International comparisons. *Annals of the Academy of Political and Social Sciences, 683*(1), 75–92. https://doi.org/10.1177%2F0002716219843804

Braun, H., & von Davier, M. (2018). The use of test scores from large-scale assessment surveys: Psychometric and statistical considerations. *Large-Scale Assessments in Education, 5*, Article 17. https://doi.org/10.1186/s40536-017-0050-x

Braun, H. I., Wang, A., Jenkins, F., & Weinbaum, E. (2006). The Black–White achievement gap: Do state policies matter? *Education Policy Analysis Archives, 14*, Article 8. https://doi.org/10.14507/epaa.v14n8.2006

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

Burdett, N., & O'Donnell, S. (2016). Lost in translation? The challenges of educational policy borrowing. *Educational Research, 58*(2), 113–120. https://doi.org/10.1080/00131881.2016.1168678

Byrne, B., & van de Vijver, F. J. R. (2017). The maximum likelihood alignment approach to testing for approximate measurement invariance: A paradigmatic cross-cultural application. *Psicothema, 29*(4), 539–551. https://doi.org/10.7334/psicothema2017.178

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). Chapman & Hall/CRC.

Carnoy, M. (2015). *International test score comparisons and educational policy: A review of the critiques*. National Education Policy Center. https://nepc.colorado.edu/publication/international-test-scores

Carnoy, M. (2019). The uneasy relation between international testing and comparative education research. In L. E. Suter, E. Smith, & B. D. Denman (Eds.), *The SAGE handbook of comparative education studies* (pp. 569–595). SAGE Publications. https://doi.org/10.4135/9781526470379.n31

Carnoy, M., Garcia, E., & Kavenson, T. (2015). Bringing it back home: Why state comparisons are more useful than international comparisons for improving U.S. education policy (Briefing Paper No. 410). Economic Policy Institute. https://www.epi.org/publication/bringing-it-back-home-why-state-comparisons-are-more-useful-than-international-comparisons-for-improving-u-s-education-policy/

Clarke, M., & Luna-Bazaldua, D. (2021). *Primer on large-scale assessments of educational achievement*. World Bank. https://doi.org/10.1596/978-1-4648-1659-8

Conaway, C. (2020). Maximizing research use in the world we actually live in: Relationships, organizations, and interpretation. *Education Finance and Policy, 15*(1), 1–10. https://doi.org/10.1162/edfp_a_00299

Couture, J.-C. (2016, May 31). Association to push for PISA withdrawal. *ATA News, 50*(18). The Alberta Teachers' Association. https://legacy.teachers.ab.ca/News%20Room/ata%20news/Volume%2050%202015-16/Number-18/Pages/PISA-withdrawal.aspx

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press. https://doi.org/10.12987/9780300255881

Darling-Hammond, L. (2012). Policy frameworks for new assessments. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 301–339). Springer Open. https://doi.org/10.1007/978-94-007-2324-5_6

Dept, S., Ferrari, A., Souto Pico, M., & Lupsa, D. (2025). Translation in Computer-based international large scale assessment: Not a stand-alone component. In L. Khorramdel, M. von Davier, & K. Yamamoto (Eds.), *Innovative computer-based international large-scale assessments—foundations, methodologies, and quality assurance procedures* (pp. 191–219). Springer Open.

Dobbins, M., & Martens, K. (2012). Towards an education approach à la finlandaise? French education policy after PISA. *Journal of Education Policy, 27*(1), 23–43. https://doi.org/10.1080/02680939.2011.622413

Durrant, G., & Schnepf, S. (2017). Which schools and pupils respond to educational achievement surveys? A focus on the English Programme for International Student

Assessment sample. *Journal of the Royal Statistical Society. Series A: Statistics in Society, 181*(4), 1057–1075. https://doi.org/10.1111/rssa.12337

Ebbs, D., & Wry, E. (2016). 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. 1–15). TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-7.html

Ercikan, K., & Pellegrino, J. W. (Eds.). (2017). *Validation of score meaning using examinee response processes for the next generation of assessments*. Routledge. https://doi.org/10.4324/9781315708591

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://www.jstor.org/stable/4618685

Feuer, M. J. (2012). *No country left behind: Notes on the rhetoric of international comparisons of education* (William Angoff Invited Lecture). ETS.

Feuer, M. J. (2013). Validity of international large-scale assessments: "Truth and consequences." In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability and equity* (pp. 197–216). Emerald Group Publishing.

Fischman, G., Marcetti Topper, A., Silova, I., Goebel, J., & Holloway, J. L. (2019). Examining the influence of international large-scale assessments on national education policies. *Journal of Education Policy, 34*(4), 470–499. http://doi.org/10.1080/02680939.2018.1460493

Fishbein, B., Foy, P., & Tyack, L. (2020). *Chapter 10: Reviewing the TIMSS 2019 achievement item statistics.* In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (pp. 10.1–10.70). TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/timss2019/methods/chapter-10.html

Fishbein, B., Martin, M. O., 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*, Article 11. https://doi.org/10.1186/s40536-018-0064-z

Fogg, N., Harrington, P., & Khatiwada, I. (2018). *Skills and earnings in the full-time labor market*. ETS. https://www.ets.org/s/research/pdf/skills-and-earnings-in-the-part-time-labor-market.pdf

Glas, C., & Jehangir, K. (2013). 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). CRC Press.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (OECD Education Working Papers No. 133). OECD Publishing. https://doi.org/10.1787/5jlzfl6fhxs2-en

Goldstein, H. (2004). International comparative assessment: How far have we really come? *Assessment in Education, 11*(2), 227–234. https://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/intl-comp-assessment-rev-essay.pdf

Goodman, M. J., Sands, A. M., & Coley, R. J. (2015). America's skills challenge: Millennials and the future. ETS. https://www.ets.org/s/research/30079/asc-millennials-and-the-future.pdf

Grek, S. (2009). Governing by numbers: The PISA "effect" in Europe. *Journal of Education Policy, 24*(1), 23–37. https://doi.org/10.1080/02680930802412669

Grund, S., Ludtke, O., & Robitzsch, A. (2021). On the treatment of missing data in background questionnaires in educational large-scale assessments: An evaluation of different procedures. *Journal of Educational and Behavioral Statistics, 46*(4), 430–465. https://doi.org/10.3102/1076998620959058

Gurkan, G. (2021). *From OLS to multilevel multidimensional mixture IRT: A model refinement approach to investigating patterns of relationships in PISA 2012 data* [Doctoral dissertation, Boston College]. http://hdl.handle.net/2345/bc-ir:109191.

Gustafsson, J. E. (2013). Causal inference in educational effectiveness research: A comparison of three methods to investigate effects of homework on student achievement. *School Effectiveness and School Improvement, 24*(3), 275–295. https://doi.org/10.1080/09243453.2013.806334

Gustafsson, J. E., & Nilson, T. (2016). The impact of school climate and teacher quality on mathematics achievement: A difference-in-differences approach. In T. Nilson & J. Gustafsson (Eds.), *Teacher quality, instructional quality and student outcomes: Relationships across countries, cohorts, and time* (Vol. 2, pp. 81–95). Springer. https://link.springer.com/chapter/10.1007/978-3-319-41252-8_4

Haertel, E. H. (2018). Tests, test scores, and constructs. *Educational Psychologist, 53*(3), 203–216. https://doi.org/10.1080/00461520.2018.1476868

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory: Vol. 1. Methods and applications.* John Wiley.

Hanushek, E. A., Schwerdt, G., Wiederhold, S., & Woessmann, L. (2015). Returns to skills around the world: Evidence from PIAAC. *European Economic Review, 73*, 103–130. https://doi.org/10.1016/j.euroecorev.2014.10.006

He, J., Barrera-Pedemonte, F., & Buchholz, J. (2019). Cross-cultural comparability of noncognitive constructs in TIMSS and PISA. *Assessment in Education: Principles, Policy & Practice, 26*(4), 369–385. https://doi.org/10.1080/0969594X.2018.1469467

Hernández-Torrano, D., & Courtney, M. G. R. (2021). Modern international large-scale assessment in education: An integrative review and mapping of the literature. *Large-Scale Assessments in Education, 9*, Article 17. https://doi.org/10.1186/s40536-021-00109-1

Heyneman, S. P. (2013). The international efficiency of American education: The bad and the not-so-bad news. In H. Meyer & A. Benavot (Eds.), *PISA, power, and policy. The emergence of global educational governance* (pp. 279–302). Symposium Books.

Heyneman, S. P., & Lee, B. (2014). The impact of international studies of academic achievement on policy and research. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 37–72). Chapman & Hall/CRC. https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-7/impact-

international-studies-academic-achievement-policy-research-stephen-heyneman-bommi-lee?context=ubx&refId=f06897f6-cdc8-40d9-bfd6-824b5d9b9b34

Hooper, M. (2017). *Applying the pseudo-panel approach to international large-scale assessments: A methodology for analyzing subpopulation trend data* [Doctoral dissertation, Boston College]. https://eric.ed.gov/?id=ED578771

Husén, T. (Ed.). (1967). *International study of achievement in mathematics: A comparison of 12 countries* (Vol. 2). John Wiley.

International Association for the Evaluation of Educational Achievement. (n.d.-a). *ILSA Gateway*. lsa-gateway.org

International Association for the Evaluation of Educational Achievement. (n.d.-b). *Tools*. https://www.iea.nl/data-tools/tools

International Association for the Evaluation of Educational Achievement. (2021). *Help manual for the IEA IDB analyzer* (Version 5.0).

Jones, L. V., & Olkin, I. (Eds.). (2004). *The nation's report card: Evolution and perspectives*. Phi Delta Kappa Educational Foundation.

Kamens, D. H. (2013). Globalization and the emergence of an audit culture: PISA and the search for "best practices" and magic bullets. In H.-D. Meyer & A. Benavot (Eds.), *PISA, power and policy: The emergence of global educational governance* (pp. 117–140). Symposium Books.

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

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23*(2), 198–211. https://doi.org/10.1080/0969594X.2015.1060192

Kane, M. T., Berryman, S., Goslin, D., & Meltzer, A. (1990). *The secretary's commission on achieving necessary skills*: *Identifying and describing the skills required by work*. Pelavan Associates. https://wdr.doleta.gov/scans/idsrw/idsrw.pdf

Kaplan, D., & McCarty, A. T. (2013). Data fusion with international large-scale assessments: A case study using the OECD PISA and TALIS surveys. *Large-Scale Assessments in Education, 1*, Article 6. https://doi.org/10.1186/2196-0739-1-6

Karabenick, S. A., Wooley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C., De Groot, E., Gilbert, M. C., Musu, L., Rogat, T., & Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist, 42*(3), 139–151. https://doi.org/10.1080/00461520701416231

Kirsch, I. S. (2001). *The International Adult Literacy Survey (IALS): Understanding what was measured* (ETS Research Report No. RR-01-25). ETS. https://doi.org/10.1002/j.2333-8504.2001.tb01867.x

Kirsch, I. S. (2003). Measuring literacy in IALS: A construct-centered approach. *International Journal of Educational Research, 39*(3), 181–190. https://doi.org/10.1016/j.ijer.2004.04.002

Kirsch, I. S., & Braun, H. I. (2020). Changing times, changing needs: Enhancing the utility of international large-scale assessments. *Large-Scale Assessments in Education, 8*, Article 10. https://doi.org/10.1186/s40536-020-00088-9

Kirsch, I. S., Braun, H., Lennon, M. L., & Sands, A. (2016). *Choosing our future: A story about opportunity in America*. ETS. https://www.ets.org/s/research/report/opportunity/ets-choosing-our-future.pdf

Kirsch, I. S., Braun, H., Yamamoto, K., & Sum, A. (2007, January). *America's perfect storm* [Policy Information Report]. ETS. https://www.ets.org/Media/Research/pdf/PICSTORM.pdf

Kirsch, I. S., Lennon, M. L., Yamamoto, K., & von Davier, M. (2017). Large-scale assessments of adult literacy. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 285-310). Springer Open. https://doi.org/10.1007/978-3-319-58689-2_9

Kroehne, U., Roelke, H., Kuger, S., Goldhammer, F., & Klieme, E. (2016, April 7–11). *Theoretical framework for log-data in technology-based assessments with empirical applications from PISA* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Washington, DC, United States.

Kuger, S., Kleime, E., Jude, N., & Kaplan, D. (Eds.). (2016). *Assessing contexts of learning: An international perspective*. Springer. https://doi.org/10.1007/978-3-319-45357-6

Kyllonen, P. C., & Bertling, J. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *A handbook of international large-scale assessment data analysis: Background, technical issues, and methods of data analysis* (pp. 277–286). Chapman & Hall/CRC Press. https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-17/innovative-questionnaire-assessment-methods-increase-cross-country-comparability-patrick-kyllonen-jonas-bertling?context=ubx&refId=2eff3ab3-f0ae-4957-abb5-ad70d13271d9

Laitusis, C. C., King, T., Cavalie, C., James, K., & Finnegan, A. (2018). *PISA special education needs feasibility study*. OECD.

Lennon, M. L., & Kirsch, I. (2025). Innovations in item development for computer-based assessments. In L. Khorramdel, M. von Davier, & K. Yamamoto (Eds.), *Innovative computer-based international large-scale assessments—foundations, methodologies and quality assurance procedures* (pp. 21–64). Springer Open.

Leu, D. J., Kiili, C., Forzani, E., Zawilinski, L., O'Byrne, W. I., & McVerry, J. G. (2021). New literacies of online research and comprehension. In C.A. Chapelle (Ed.), *The concise encyclopedia of applied linguistics* (pp. 844–852). Wiley–Blackwell. https://doi.org/10.1002/9781405198431.wbeal0865.pub2

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley & Sons.

Lockheed, M. (2013). Causes and consequences of international assessments in developing countries. In H.-D. Meyer & A. Benavot (Eds.), *PISA, power and policy: The emergence of global educational governance* (pp. 163–183). Symposium Books. https://doi.org/10.15730/books.85

Lohr, S. L. (2010). *Sampling: Design and analysis*. Brooks/Cole.

Loveless, T. (2014). *The 2014 Brown Center report on American education: How well are American students learning?* The Brookings Institution. https://www.brookings.

edu/wp-content/uploads/2016/06/2014-Brown-Center-Report_FINAL-4.pdf

Luo, X. (2020). Automated test assembly with mixed-integer programming: The effects of modeling approaches and solvers. *Journal of Educational Measurement, 57*(4), 547-565. https://doi.org/10.1111/jedm.12262

Maehler, D. B., Jakowatz, S., & Konradt, I. (2020). *PIAAC Bibliography—2008–2019* (GESIS Papers, 2020/04). GESIS, Leibniz-Institut für Sozialwissenschaften. https://doi.org/10.21241/ssoar.67732

Marchionni, M., &Vazquez, E. (2019). The causal effect of an extra year of schooling on skills and knowledge in Latin America: Evidence from PISA. *Assessment in Education: Principles, Policy & Practice, 26*(4), 489–515. https://doi.org/10.1080/0969594X.2018.1454401

Marks, G. N., & O'Connell, M. (2021). Inadequacies in the SES—achievement model: Evidence from PISA and other studies. *Review of Education, 9*(3), e3293. https://doi.org/10.1002/rev3.3293

Martin, M. O., Foy, P., Mullis, I. V., & O'Dwyer, L. M. (2013). Effective schools in reading, mathematics, and science at fourth grade. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade—implications for early learning* (pp. 109–180). TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/timsspirls2011/downloads/TP11_Relationship_Report.pdf

Martin, M. O., Mullis, I. V. S., Arora, A., & Preuschoff, C. (2014). Context questionnaire scales in TIMSS and PIRLS 2011. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 229–258). Chapman & Hall/CRC. https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-19/context-questionnaire-scales-timss-pirls-2011-michael-martin-ina-mullis-alka-arora-corinna-preuschoff?context=ubx&refId=ab5d667f-8fe2-4905-bebf-fb86a500a3b6

Martin, M. O., Mullis, I. V. S. & Hooper, M. (Eds.). (2016). *Methods and procedures in TIMSS 2015*. TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/publications/timss/2015-methods.html

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2017). *Methods and procedures in PIRLS 2016*. TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/publications/pirls/2016-methods.html

Martin, M. O., Rust, K., & Adams, R. (Eds.). (1999). *Technical standards for IEA studies.* International Association for the Evaluation of Educational Achievement. https://www.iea.nl/publications/iea-reference/technical-standards-iea-studies

Martin, M. O., von Davier, M., & Mullis, I. V. S. (2020). *Methods and procedures: TIMSS 2019 t*echnical report. TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/timss2019/methods/index.html

Mazzeo, J., Lazer, S., & Zieky, M. J. (2006). Monitoring educational progress with group-score assessments. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 681-699). American Council on Education/Praeger.

Mazzeo, J., & von Davier, M. (2014). Linking scales in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 229-258). Chapman & Hall/CRC. https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-15/linking-scales-international-large-scale-assessments-john-mazzeo-matthias-von-davier?context=ubx&refId=b641e638-90ba-40b2-8991-29db10e0ed5a

Meinck, S., Gonzalez, E., & Wagemaker, H. (2020). Consequential validity: Data access, data use, analytical support, and training. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessment: Understanding IEA's comparative studies of student achievement* (pp. 231–244). Springer Nature. https://link.springer.com/chapter/10.1007/978-3-030-53081-5_13

Messick, S. J. (1987). Large-scale educational assessment as policy research: Aspirations and limitations. *European Journal of Psychology of Education, 2*, Article 157. https://doi.org/10.1007/BF03172645

Messick, S. J. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan Publishing Company.

Messick, S. J. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23. https://doi.org/10.3102/0013189X023002013

Messick, S. J., Beaton, A. E., & Lord, F. M. (1983). *National Assessment of Educational Progress reconsidered*: *A new design for a new era* (NAEP Report No. 83-1). ETS.

Meyer, H.-D., & Benavot, A. (2013). PISA and the globalization of education governance: Some puzzles and problems. In H.-D. Meyer & A. Benavot (Eds.), *PISA, power, and policy: The emergence of global educational governance* (pp. 9–26). Symposium Books. https://doi.org/10.15730/books.85

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196. https://doi.org/10.1007/BF02294457

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge. https://doi.org/10.4324/9781315871691

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/10.1111/j.1745-3984.1992.tb00371.x

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-centered assessment design* (ETS Research Report No. RR-03-16). ETS.

Mohadjer, L., Berlin, M., Rieger, S., Waksberg, J., Rock, D., Yamamoto, K., Kirsch, I., & Kolstad, A. (1997). The role of incentives in literacy survey research. In A. C. Tuijnman, I. S. Kirsch, & D. A. Wagner (Eds.), *Adult basic skills: Innovations in measurement and policy analysis* (pp. 209–244). Hampton Press.

Mohadjer, L., & Edwards, B. (2018). Paradata and dashboards in PIAAC. *Quality Assurance in Education, 26*(2), 263–277. https://doi.org/10.1108/QAE-06-2017-0031

Mosenthal, P. B., & Kirsch, I. S. (1991). Toward an explanatory model of document literacy. *Discourse Processes, 14*(2), 147–180. https://doi.org/10.1080/01638539109544780

Mullis, I. V. S., & Martin, M. O. (2019). *PIRLS 2021 assessment frameworks*. TIMSS & PIRLS International Study Center, Boston College. http://pirls2021.org/frameworks/

Mullis, I. V. S., Martin, M. O., & von Davier, M. (2021). *TIMSS 2023 assessment frameworks*. TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/timss2023/frameworks/pdf/T23_Frameworks.pdf

National Center for Education and the Economy. (2020, March 19). *Groundbreaking Maryland education reform bill passes in midst of Coronavirus crisis*. https://ncee.org/tucker-writing/groundbreaking-maryland-education-reform-bill-passes-in-midst-of-coronavirus-crisis/

Ng, H. L., Poon, C. L., & Pang, E. (2020). Using IEA studies to inform policymaking and program development: The case of Singapore. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessment: Understanding IEA's comparative studies of student achievement* (pp. 245–260). Springer Nature. https://link.springer.com/chapter/10.1007/978-3-030-53081-5_14

O'Dwyer, L. M., & Paolucci, C. (2019). Challenges in practice: A critical examination of efforts to link teacher practices and student achievement. In L. E. Suter, E. Smith, & B. D. Denman (Eds.), *The SAGE handbook of comparative education studies* (pp. 471–491). SAGE. https://doi.org/10.4135/9781526470379.n26

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 No. RR-18-27). ETS. https://doi.org/10.1002/ets2.12214

Oranje, A., & Ye, L. (2014). Population model size, bias, and variance in educational survey assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 203–228). Chapman & Hall/CRC. https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-14/population-model-size-bias-variance-educational-survey-assessments-andreas-oranje-lei-ye?context=ubx&refId=8bfdb5cf-e50a-41d3-9738-7e583589a86d

Organisation for Economic Co-operation and Development. (1992). *Adult illiteracy and economic performance*.

Organisation for Economic Co-operation and Development. (2016). *Survey of adult skills technical report* (2nd ed.). https://www.oecd.org/skills/piaac/PIAAC_Technical_Report_2nd_Edition_Full_Report.pdf

Organisation for Economic Co-operation and Development. (2017). *PISA 2015 technical report.* https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/pisa/publications/technical-report/PISA2015_TechRep_Final.pdf

Organisation for Economic Co-operation and Development. (2019). *PISA 2018 results (Volume II): Where all students can succeed.* https://doi.org/10.1787/b5fd1b8f-en

Organisation for Economic Co-operation and Development. (2020). *PISA 2018 technical report*. https://www.oecd.org/pisa/data/pisa2018technicalreport/

Pepper, D. (2020). When assessment validation neglects any strand of validity evidence: An instructive example from PISA. *Educational Measurement: Issues and Practice, 39*(4), 8-20. https://doi.org/10.1111/emip.12380

Pokropek, A. (2016a). Grade of membership response time model for detecting guessing behaviors. *Journal of Educational and Behavioral Statistics, 41*(3), 300–325. https://doi.org/10.3102/1076998616636618

Pokropek, A. (2016b). Introduction to instrumental variables and their application to large-scale assessment data. *Large-Scale Assessments in Education, 4*, Article 4. https://doi.org/10.1186/s40536-016-0018-2

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

Pools, E., & Monseur, C. (2021). Student test-taking effort in low-stakes assessments: Evidence from the English version of the PISA 2015 science test. *Large-Scale Assessments in Education, 9*, Article 10. https://doi.org/10.1186/s40536-021-00104-6

Provasnik, S. (2021). Process data, the new frontier for assessment development: Rich new soil or a quixotic quest? *Large-Scale Assessments in Education, 9*, Article 1. https://doi.org/10.1186/s40536-020-00092-z

Ritzen, J. (2013). International large-scale assessments as change agents. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 13–24). Springer. https://doi.org/10.1007/978-94-007-4629-9_2

Robinson, J. P. (2014). Causal inference and comparative analysis with large-scale assessment data. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 521-546). Chapman & Hall/CRC. https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-28/causal-inference-comparative-analysis-large-scale-assessment-data-joseph-robinson?context=ubx&refId=eb755322-c67a-4e25-9237-854b5df556d1

Rosdahl, A. (2014). *Fra 15 til 27 år: PISA 2000-eleverne I 2011/12 [From 15 to 27 years old: The PISA 2000 students in 2011/12]*. SFI, Det Nationale Forskningscenter for Velfærd. https://www.vive.dk/media/pure/5070/276298

Rowley, K. J., Edmunds, C. C., Dufur, M. J., Jarvis, J. A., & Silveira, F. (2020). Contextualising the achievement gap: Assessing educational achievement, inequality, and disadvantage in high-income Countries. *Comparative Education, 56*(4), 459–483. https://doi.org/10.1080/03050068.2020.1769928

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: Background, technical issues, and methods of data analysis* (pp. 117–154). Chapman & Hall/CRC. https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-11/sampling-weighting-variance-estimation-international-large-scale-assessments-keith-rust?context=ubx&refId=6b706f02-cc7d-4143-abaf-40619d1a0448

Rutkowski, D., & Delandshere, G. (2016). Causal inferences with large scale assessment data: Using a validity framework. *Large-Scale Assessments in Education, 4*, Article 6. https://doi.org/10.1186/s40536-016-0019-1

Rutkowski, D., Thompson, G., & Rutkowski, L. (2020). Understanding the policy influence of international large-scale assessments in education. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessment: Understanding IEA's comparative studies of student achievement* (pp. 261–277). Springer Nature. https://doi.org/10.1007/978-3-030-53081-5_15

Rutkowski, L., Gonzalez, E., von Davier, M., & Zhou, Y. (2014). Assessment design for international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 75–95). Chapman & Hall/CRC. https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-9/assessment-design-international-large-scale-assessments-leslie-rutkowski-eugene-gonzalez-matthias-von-davier-yan-zhou?context=ubx&refId=0ea1bb81-169f-4ce1-a73f-b05180363b3c

Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher, 45*(4), 252–257. https://doi.org/10.3102/0013189X16649961

Rutkowski, L., & Rutkowski, D. (2018). Improving the comparability and local usefulness of international assessments: A look back and a way forward. *Scandinavian Journal of Educational Research, 62*(3), 354–367. https://doi.org/10.1080/00313831.2016.1261044

Rutkowski, L., & Rutkowski, D. (2019). Methodological challenges to measuring heterogeneous populations internationally. In L. E. Suter, E. Smith, & B. D. Denman (Eds.), *The SAGE handbook of comparative education studies* (pp. 126–140). SAGE. https://doi.org/10.4135/9781526470379.n8

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: Categorical indicators & fit measure performance. *Applied Measurement in Education, 30*(1), 39–51. https://doi.org/10.1080/08957347.2016.1243540

Rutkowski, L., & Zhou, Y. (2015). The impact of missing and error-prone auxiliary information on sparse-matrix sub-population parameter estimates. *Methodology, 11*(3), 89–99. https://doi.org/10.1027/1614-2241/a000095

Rychen, D. S., & Salganik, L. H. (Eds.). (2001). *Defining and selecting key competencies*. Hogrefe & Huber.

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

Sands, A., Goodman, M., Kirsch, I., & Dreier, K. (2021). *Opportunity across the states.* ETS. https://www.ets.org/s/research/pdf/opportunity-across-the-states.pdf

Schleicher, A. (2018). *World class: How to build a 21st century school system*. OECD. https://www.oecd-ilibrary.org/world-class_5j8v15v201hd.pdf?itemId=%2Fcontent%2Fpublication%2F9789264300002-en&mimeType=pdf

Schwippert, K. (Ed.). (2007). *Progress in reading literacy: The impact of PIRLS 2001 in 13 countries*. Waxmann.

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

Singer, J. D., & Braun, H. I. (2018). Testing international education assessments. *Science, 360*(6384), 38–40. https://doi.org/10.1126/science.aar4952

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

Soland, J., Kuhfeld, M., & Rios, J. (2021). Comparing different response time threshold setting methods to detect low effort on a large-scale assessment. *Large-Scale Assessments in Education, 9*, Article 8. https://doi.org/10.1186/s40536-021-00100-w

Strietholt, R., & Rosén, M. (2016). Linking large-scale reading assessments: Measuring international trends over 40 years. *Measurement: Interdisciplinary Research and Perspectives, 14*(1), 1–26. https://doi.org/10.1080/15366367.2015.1112711

Strietholt, R., & Scherer, R. (2018). The contribution of international large-scale assessments to educational research: Combining individual and institutional data sources. *Scandinavian Journal of Educational Research, 62*(3), 368–385. https://doi.org/10.1080/00313831.2016.1258729

Suter, L. E. (2019). Growth and development of large-scale international comparative studies and their influence on comparative education thinking. In L.E. Suter, E. Smith, & B. D. Denman (Eds.), *The SAGE handbook of comparative education studies* (pp. 197–223). SAGE. https://doi.org/10.4135/9781526470379.n12

Suter, L. E., Smith, E., & Denman, B. D. (Eds.). (2019). *The SAGE handbook of comparative education studies*. SAGE Publishing. https://doi.org/10.4135/9781526470379

Thomas, N. (2002). The role of secondary covariates when estimating latent trait population distributions. *Psychometrika, 67*(1), 33–48. https://link.springer.com/article/10.1007/BF02294708

Thurlow, M. L. (2014). Instructional and assessment accommodations in the 21st century. In L. Florian (Ed.), *The SAGE handbook of special education* (Vol. 2, pp. 597-631). SAGE Publications. https://doi.org/10.4135/9781446282236.n37

Tucker, M. (Ed.). (2011). *Surpassing Shanghai: An agenda for American education built on the world's leading systems*. Harvard Education Press.

Ulitzsch, E., von Davier, M., & Pohl, S. (2019). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research, 55*(3), 425–453. http://doi.org/10.1080/00273171.2019.1643699

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287–308. https://doi.org/10.1007/s11336-006-1478-z

van de Vijver, F. J. R. (2018). Towards an integrated framework of bias in noncognitive assessment in large-scale international studies: Challenges and prospects. *Educational Measurement: Issues and Practice, 37*(4), 49–56. https://doi.org/10.1111/emip.12227

van de Vijver, F. J. R., Jude, N., & Kuger, S. (2019). Challenges in large-scale education surveys. In L. E. Suter, E. Smith, & B. D. Denman (Eds.), *The SAGE handbook of comparative education studies* (pp. 83–102). SAGE Publications. https://doi.org/10.4135/9781526470379.n6

Vera-Toscano, E., Rodrigues, M., & Costa, P. (2017). Beyond educational attainment: The importance of skills and lifelong learning for social outcomes. Evidence for Europe from PIAAC. *European Journal of Education, 52*(2), 217–231. https://doi.org/10.1111/ejed.12211

Volante, L. (2013). Canadian policy responses to international comparison testing. *Interchange, 44*(3-4), 169–178. https://doi.org/10.1007/s10780-014-9205-7

Volante, L., Fabio, X., & Ritzen, J. (2017). The OECD and educational policy reform: International surveys, governance, and policy evidence. *Canadian Journal of Educational Administration and Policy, 184*, 34–48.

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). Chapman & Hall/CRC Press. https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-13/imputing-proficiency-data-planned-missingness-population-models-matthias-von-davier?context=ubx&refId=f5f1b7e1-533e-40e1-95e9-54dc43ef3848

von Davier, M., & Bezirhan, U. (2022). A robust method for detecting item misfit in large-scale assessments. *Educational and Psychological Measurement, 83*(4), 740–765. https://doi.org/10.1177/00131644221105819

von Davier, M., Gonzalez, E., & Schulz, W. (2020). Ensuring validity in international comparisons using state-of-the-art psychometric methodologies. In H. Wagemaker (Ed.), *IEA research for education: Vol. 10. Reliability and validity of international large-scale assessment* (pp. 187–220). Springer. https://doi.org/10.1007/978-3-030-53081-5_11

von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & 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), 671–705. https://doi.org/10.3102/1076998619881789

von Davier, M., Shin, H.-J., Khorramdel, L., & Stankov, L. (2018). The effects of vignette scoring on reliability and validity of self-reports. *Applied Psychological Measurement, 42*(4), 291–306. https://doi.org/10.1177/0146621617730389

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). Chapman & Hall/CRC Press. https://www.taylorfrancis.comchapters/edit/10.1201/b16061-12/analytics-international-large-scale-assessments-item-response-theory-population-models-matthias-von-davier-sandip-sinharay?context=ubx&refId=d996833d-7903-4b7a-ac0b-d0799cb098cd

von Davier, M., Yamamoto, K., Shin, H. J., Chen, H., Khorramdel, L., Weeks, J., Davis, S., Kong, N., & Kandathil, M. (2019). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assessment in Education: Principles, Policy & Practice, 26*(4), 466–488. https://doi.org/10.1080/0969594X.2019.1586642

Wagemaker, H. (2014). International large-scale assessments: From research to policy. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 11–36). Chapman & Hall/CRC. https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-6/international-large-scale-assessments-research-policy-hans-wagemaker?context=ubx&refId=51ae00d5-8ff4-4252-9d46-ab7232d19820

Wetzel, E., Xu, X., & von Davier, M. (2015). An alternative way to model population ability distributions in large-scale educational surveys. *Educational and Psychological Measurement, 75*(5), 739–763. https://doi.org/10.1177/0013164414558843

Wirtz, W., & Lapointe, A. (1982). Measuring the quality of education: A report on assessing educational progress. *Educational Measurement: Issues and Practice, 1*, 17–19, 23. https://doi.org/10.1111/j.1745-3992.1982.tb00673.x

Wise, S. L. (2020). An intelligent CAT that can deal with disengaged test taking. In H. Jiao & R. W. Lissitz (Eds.), *Applications of artificial intelligence (AI) to assessment* (pp. 161–174). Information Age Publishing.

Wiseman, A. (2013). Policy responses to PISA in comparative perspective. In H. D. Meyer & A. Benavot (Eds.), *PISA, power, and policy the emergence of global educational governance* (pp. 303–322). Symposium Books.

Yamamoto, K., & Lennon, M. L. (2018). Understanding and detecting data fabrication in large-scale assessments. *Quality Assurance in Education, 26*(2), 196–212. https://doi.org/10.1108/QAE-07-2017-0038

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice, 37*(4), 16-27. https://doi.org/10.1111/emip.12226

Zumbo, B. D., & Hubley, A. M. (Eds.). (2017). *Understanding and investigating response processes in validation research*. Springer. https://doi.org/10.1007/978-3-319-56129-5

NOTES

1. In this chapter, we use the term *country* to signify a participating jurisdiction for which results are reported. This term includes "countries" as commonly understood but may also refer to other types of political entities. In ILSA reports, the latter are variously described as *economies* or *benchmarking participants*.
2. Human capital is often characterized as a broad set of cognitive and noncognitive skills and knowledge that is necessary in modern economies. See Kirsch et al. (2016) for a discussion of the growing importance of human and social capital and their connections to opportunity.

3. Advanced TIMSS was intended to assess students close to the end of compulsory education (e.g. Grade 12 in the United States). It was administered in 1995, 2008, and 2015.
4. Each paying participant obtains a database with its results, and the results are published in the accompanying reports.
5. More exactly, PIRLS is administered to students in the upper of the two grades that enroll the most 9-year-olds. This is the fourth grade in most countries.
6. For PIRLS 2021, ePIRLS booklets were more fully integrated into the booklet structure of digital PIRLS.
7. These assessments were the Young Adult Literacy Survey 1984, the Department of Labor Study of JTPA Trainees and UI recipients 1992, and the National Adult Literacy Survey 1993.
8. The ICILS was conducted in 2013, 2018, and 2023.
9. ICCS assesses students enrolled in the eighth grade, provided that the average age of students at this year level is 13.5 years or above. In countries where the average age of students in Grade 8 is less than 13.5 years, Grade 9 is defined as the target population. The ICCS was conducted in 2009, 2016, and 2022.
10. Beginning in 2003, succeeding assessments retained the acronym TIMSS, with "Trends" replacing "Third" in the title.
11. BICSE was established in 1988 and disbanded in 2002. Throughout its existence, the National Center for Education Statistics provided substantial funding.
12. With a statutory mandate to collect data in the United States, the National Center for Education Statistics, a division within the U.S. Department of Education, has also provided crucial, sustained support for many of the international assessments beyond the contribution required for U.S. participation.
13. These dates mark the beginning of data collection for the surveys. Of course, preparatory work began years earlier.
14. For a contemporary treatment of constructs and test scores, see Haertel (2018).
15. This and other issues are reviewed as part of the adjudication process that is completed at the end of the project cycle.
16. The separate estimation of item parameters for a country (or set of countries) improves the estimated score distributions for those countries. If the number of such items is small, this policy should not materially impact the comparability of the described proficiency scale across all countries. However, if the number of items requiring separate estimation exceeds a certain threshold (yet to be determined), then it may be advisable to construct a unique scale for that country (based on its data only) and sacrifice comparability with the main scale.
17. For a more complete discussion of this activity, see Dept et al. (2025).
18. For example, in school-based surveys, the aim is to estimate score distributions for students, but the school is a relevant level of analysis. Consequently, the

sample design must allow for selecting multiple students within each school, as well as large numbers of schools. Schools constitute the first stage of the design, and students within schools constitute the second stage.

19. To prepare for school nonresponse, replacement schools are selected at the same time as the main sample.
20. For example, in PIRLS 2016 the target population was students in the fourth year of formal schooling. In most countries this was Grade 4. However, if the mean age in Grade 4 was less than 9.5 years, Grade 5 was chosen. For further details, see Martin et al. (2017, chap. 3).
21. If the school has fewer than 35 eligible students, all eligible students are selected for administration.
22. As noted earlier, countries vary in the proportions of the birth cohorts enrolled in school, with implications for the kinds of (comparative) inferences that can be made from the data collected.
23. The problem has become more salient with the greater participation of middle- and low-income countries.
24. Blocks are systematically varied by position within booklets to remove any order effects.
25. More exactly, the students responding to each item have known probabilities of selection so that the raw results can be modified by appropriate weighting. With adaptive testing, the students exposed to different item blocks are not randomly equivalent. Nonetheless, the designs do enable unbiased estimation of item parameters.
26. In many contexts, such item sets are called *testlets*.
27. Items requiring human scoring cannot contribute to the real-time decision-making process; however, they may be essential for construct representation.
28. When there is interest in reporting information at the subscale level, items are grouped by subscale and the item parameters are obtained from the unidimensional scaling employed in scale construction. This strategy possesses advantages when comparing subscale profiles across countries.
29. For more technical detail on addressing item–country interactions, see, for example, Martin et al. (2016, chap. 11) for TIMSS 2015, and Martin et al. (2017, chap. 10) for PIRLS 2016. On occasion, a single separate item calibration is carried out for a group of countries.
30. See also Mullis et al. (2021, chap. 4).
31. As a general rule, the number of principal components retained is limited to no more than 5% of a country's student sample size, thereby reducing the percentage of variance accounted for to avoid overspecification of the conditioning model. (See Martin et al., 2020, chap. 12.)
32. Various versions of the expectation–maximization algorithm are used to obtain parameter estimates for the latent regression model (von Davier & Sinharay, 2014).

33. This discussion assumes there are no missing data in the background questionnaire, an assumption that is rarely fulfilled. Recent research has investigated methods for carrying out the estimation procedures in the face of missing data (e.g., Grund et al., 2021).
34. See also OECD, 2017, chap. 9.
35. More exactly, the key requirement is that there is sufficient information (i.e., large enough sample size) on each trend item so that it is possible to judge whether it is working well in each population for which results are reported—typically a (large) language group within a country.
36. In this context, "freeing item parameters" entails estimating the item parameters separately for each country, rather than having one set of item parameters for all countries. For more technical detail on addressing item-by-country interactions, see Martin et al. (2020, chap. 10) for TIMSS 2019 and Martin et al. (2017, chap. 10) for PIRLS 2016.
37. See also OECD (2017, chap. 9).
38. NAEP no longer uses this approach, having shifted to employing achievement-level descriptors.
39. More generally, in PISA, students who are administered items in the major domain and only one of the minor domains receive PVs for the other minor domain.
40. Recall that sampling weights are used to control the proportional representation of the cases in the estimation of population parameters.
41. See International Association for the Evaluation of Educational Achievement (n.d.-b).
42. Both tools are considered proprietary software developed specifically for use with ILSA data. The current version of the microdata analyzer is called IDBA.
43. Attendees may be required to pay fees to attend, depending on the level of governmental support available.
44. See International Association for the Evaluation of Educational Achievement (n.d.-a).i
45. See, for example, Martin et al. (1999).
46. Canada, France, Norway, Switzerland, England, and Shanghai.
47. Educational attainment is typically quantified in terms of either years of schooling completed or reaching certain milestones, such as completion of secondary school.
48. Developing an appropriate cross-national measure of socioeconomic status is itself a challenging endeavor (Avvisati, 2020; Marks & O'Connell, 2021).
49. One concern with such studies is that the focal construct may be operationally defined differently in each assessment. Such a "construct shift" limits the kinds of conclusions that can be drawn.
50. Although the transition from paper-and-pencil administration to digitally based assessments occurs at a point in time, it may prove necessary to replenish the

paper-and-pencil administration item pools if there is an ongoing need to offer the paper-and-pencil administration option.

51. Maintaining item parameter invariance is one way to achieve such comparability.
52. The lack of familiarity is mitigated somewhat with tutorials and practice exercises that familiarize respondents with new item response formats.
53. The issue is complicated by the fact that the construct definitions and the corresponding assessment frameworks do not always make explicit which types of prior knowledge are part of the construct and which are not.
54. Denmark, Finland, Iceland, Norway, and Sweden.
55. In the ILSA context, an extreme case of the aptitude–achievement paradox occurs when the relationship between a cognitive outcome and a behavioral (or attitudinal) outcome is positive within countries but negative among countries. See Ainley and Ainley (2019) for further discussion.
56. Some countries employed incentives in Cycle 1 of PIAAC and others are planning to use them in Cycle 2. An incentive study conducted in conjunction with the U.S. National Adult Literacy Study found that monetary incentives increased participation rates but not performance (Mohadjer et al., 1997).
57. Cognitive laboratories were conducted for PIAAC, with particular attention on interface issues related to usability. However, they are not standard for school-based ILSAs.
58. We thank Xue Jiang and Shinji Katsumoto, Teachers College, Columbia University, for the translations.
59. To this point, M. Smith (cited in Singer et al., 2018, p. 28, footnote) asserted that in the past, legislators found state-level differences more compelling than country-level differences.
60. Quoted in Singer et al. (2018), pp. 28–29.
61. Consult Maehler et al. (2020) for a comprehensive bibliography.
62. For an extended example using U.S. PIAAC data, see the report by Sands et al. (2021).
63. For more on Shanghai's PISA results, see Loveless (2014).