Kirsch and Braun *Large-scale Assess Educ* (2020) 8:10
https://doi.org/10.1186/s40536-020-00088-9

Large-scale Assessments in Education

RESEARCH Open Access

Changing times, changing needs: enhancing the utility of international large-scale assessments

Irwin Kirsch[1*] and Henry Braun[2]

*Correspondence:
ikirsch@ets.org
[1] Educational Testing Service, Princeton, NJ, USA
Full list of author information is available at the end of the article

Abstract

Mounting concerns about the levels and distributions of human capital, as well as how they are associated with outcomes for individuals and societies, have contributed to an increase in the number of national and international surveys. These surveys not only examine skills among school-age and adult populations, they also facilitate evaluation of the relationships among these skills and various background factors. At this juncture, the main ILSAs are making the transition to becoming fully digitally based assessments (DBAs). With the transition rapidly progressing, this is a propitious moment to consider the history of large-scale national and international assessments and to reflect on both what has contributed to their increased salience and growth, and how best to enhance their constructive impact on both policy and policy research in the future. We argue this can be done by utilizing a comprehensive, multidimensional framework that establishes a set of design criteria against which these assessments, present and future can be evaluated. The original framework was proposed more than 30 years ago by Messick (European Journal of Psychology of Education II 2:157–165, 1987) when all large-scale assessments were developed and delivered with paper-based instruments. Messick's framework provided an organizational structure for how to think about and evaluate the potential utility of all large-scale assessments. After presenting a brief historical perspective on the development and growth of large-scale assessments, we review and extend Messick's original framework. We then describe how the transition to DBAs can impact each of the framework's three criteria. We also offer recommendations on how ILSAs' development and innovations can best be deployed so that they are of greater utility to policy makers and other stakeholders worldwide.

Introduction

Some say that technology changes everything. Both in the workplace and in day-to-day activities, individuals are increasingly required to navigate, critically analyze, and problem solve in data-intensive, complex digital environments. Beyond the loss of millions of jobs in the United States and elsewhere, many have warned about the larger impact artificial intelligence (AI) and robotics will likely have on both labor markets and everyday life (Brynjolfsson and McAfee 2014). And, according to a recent report from the Metropolitan Policy Program at the Brookings Institution (Muro et al. 2019), automation and

AI will affect virtually all occupations in the future, but their effects will vary considerably across occupations, places, and demographic groups in the United States. In general, however, those seeking jobs with better wages and benefits will need increasingly higher levels of education and skills.

These escalating demands can be conveniently organized into three related categories. First, valued skill profiles now comprise not only essential competencies such as literacy and numeracy, but also so-called 21st century skills that include critical thinking, problem solving, collaboration, and creativity (Autor et al. 2019; Bughin et al. 2018; World Economic Forum 2019). Second, as AI-powered software is implemented to carry out more complex tasks, individuals will need higher levels of skills, as well as the ability to apply those skills in new settings. Finally, employers seek workers who can keep pace with rapidly changing technologies. As a result, they are looking for individuals who have both the ability and initiative to learn on their own and continuously upgrade what they know and can do.

Mounting concerns about the levels and distributions of human capital[1] and how they are associated with outcomes for individuals and societies have contributed to an increase in the number of national and international surveys that examine skills for adults and school age populations, as well as the number of participating countries/jurisdictions that take part in these surveys (Heyneman and Lee 2014; Wagemaker 2014). Concomitantly, there has been an exponential increase in the quantity of policy-relevant research drawing on the data generated by such assessments. These secondary analyses are carried out by researchers around the world, representing a wide range of disciplines and organizations, and are accompanied by greater efforts by various entities to make these data more accessible.[2]

This growing interest is also evident in greater media attention that, in turn, raises the salience of international large-scale assessments (ILSAs) in policy circles and among the public at-large. Not surprisingly, greater salience has led to increased scrutiny and criticism of ILSAs with respect to both the methodologies employed and the uses of ILSA data for policy. Methodological critiques have focused on the assumed cross-cultural equivalence of both the cognitive instruments and the background questionnaires (van de Vijver 2018). Others have centered on the heterogeneity of sample quality among participating countries (Kirsch et al. 2018).

Policy critiques have focused on the overemphasis on, and misuse of, ILSAs' country rankings, especially the common tendency to interpret differences in ranks as being credible indicators of differences in the quality of the corresponding education systems. Some have stressed the overinterpretation of cross-sectional data and frequent use of language suggesting causal inferences (Rutkowski and Delandshere 2016; Singer and Braun 2018). Others doubt the relevance of one country's experiences and policies to another's (Carnoy et al. 2015). Yet others cite the tendency to adopt apparently successful

[1] Human capital is often characterized as a broad set of cognitive and noncognitive skills and knowledge that is necessary in modern economies. See the ETS report titled *Choosing Our Future* written by Kirsch et al. (2016) for a discussion around the growing importance of human and social capital and their connections to opportunity. https://www.ets.org/research/report/opportunity.

[2] Examples of the types of papers and reports that have been developed using the PIAAC data can be found in the following reference listed at the end of this paper (Maehler et al. 2018). PIAAC bibliography—2008–2017.

policies without due consideration of the appropriateness of those policies in different settings. One example is the recent decision by England to equip more than half of its elementary schools with mastery textbooks for mathematics based on the observation that high ranked East Asian countries use such books.[3]

On the positive side, continuing and growing participation in ILSAs indicates a general consensus that these surveys provide jurisdictions with generally comparable results and credible, policy-relevant information. ILSAs enable developed countries to assess their relative standing with respect to the distributions of human capital, the relationships between specific skills and background factors, as well as to the contexts for learning and for using these skills. This information enables policy makers to understand better the factors that (may) contribute to skill trends and patterns (e.g., strengths of relationships of skills to demographic characteristics). For many commentators, the relevance of the policy choices identified through secondary analyses is almost self-evident. This leads, in turn, to the strategies lower-ranked jurisdictions can consider adopting or adapting (Schleicher 2018). For developing countries there are similar motivations, as well as the prospect of support from donor organizations to develop, implement and maintain the infrastructure needed to obtain this kind of information on a systematic basis (Lockheed 2013).

At this juncture, the largest ILSAs are making the transition to becoming fully digitally based assessments (DBAs). With the transition rapidly progressing, this is a propitious moment to consider the history of large-scale national and international assessments and to reflect on both what has contributed to their increased salience and growth, and how best to enhance their constructive impact on both policy and policy research in the future. We argue that can be done by utilizing a comprehensive, multidimensional framework that establishes a set of design criteria against which these assessments, present and future, can be evaluated. The original framework was proposed more than 30 years ago by Messick (1987) when all large-scale assessments were developed and delivered with paper-based instruments. Messick's framework provided an organizational structure for how to think about and evaluate the potential utility of all large-scale assessments. After presenting a brief historical perspective on the development and growth of large-scale assessments, we review and extend Messick's original framework. We then describe how the transition to DBAs can impact each of the framework's three criteria, for good or for ill. We also offer recommendations on how ILSAs' development and innovation can best be deployed so that they are of greater utility to policy makers and other stakeholders worldwide.

Historical perspective

Before the late 1950s, little if any systematic data relating to the outcomes of education were collected at either the national or international level. Then, in 1958, a group of scholars met at the UNESCO Institute for Education in Hamburg, Germany, to discuss issues associated with trying to evaluate schools and student learning. Notwithstanding the many challenges, these individuals saw the rich possibilities of developing

[3] https://www.gov.uk/government/news/south-asian-method-of-teaching-maths-to-be-rolled-out-in-schools.

international surveys as a source of relevant evidence regarding important factors that influence educational outcomes among the participating countries. They hypothesized that the variation and patterns found in the data would provide important insights into how best to influence change (Foshay et al. 1962).

The first international study, conducted in 1960, involved 12 countries and assessed 13-year-old students in 5 domains including reading and mathematics. Among other things, the study, known as the Pilot Twelve-Country Study, demonstrated the feasibility of developing and conducting international large-scale assessments and led to the formal establishment of IEA (International Association for the Evaluation of Educational Achievement) in 1967. It should be noted that this and other early studies were motivated more by an interest in comparative education research and less by collecting information that could be used to inform education policy making. The shift away from more traditional academic research toward informing policy makers came later, with the recognition of ongoing societal changes and the need for different types of information.

In the United States, a parallel effort began around the same time. Francis Keppel, the then-U.S. Commissioner of Education, was responsible for reporting to Congress on the condition of education.[4] Among his concerns was the fact that national data focused primarily on the inputs to education, with no information on what students were learning. This led to the establishment of two planning committees charged with thinking about the development of a national assessment of students. The work of these two committees in the early–to mid-1960s led to the creation of what is now known as the National Assessment of Educational Progress (NAEP), or the Nation's Report Card. NAEP conducted its first assessment of in-school 17-year-olds in 1969, with assessments in the domains of science, writing, and citizenship. This was followed later that same year by an assessment of 9- and 13-year-olds in the same subjects (Jones and Olkin 2004).

NAEP assessments continued through the 1970s with both growing interest and concerns by policy makers and other stakeholders. The growing interest in the NAEP data came from the recognition of the importance of education and skills in a society that was undergoing significant changes. The concerns were twofold. First, many policy makers and other stakeholders felt the data showed that, across the country, schools were neither adequately serving the needs of particular subpopulations nor meeting the changing needs of the workplace or society. Second, there was a consensus that NAEP, given the changes that were taking place in the country, was no longer providing interpretable, policy-relevant information. These concerns led to the establishment of a national commission. Its report (Wirtz and Lapointe 1982) emphasized the limitations of the NAEP design and proposed directions for addressing a new set of questions.

Following the work of this commission, a team of scholars working at the Educational Testing Service (ETS) developed a proposal that featured an entirely new design for NAEP (Messick et al. 1983). The evident strengths of the proposed design resulted in the transfer of the program of work from the Education Commission of the States (where it had been since NAEP's inception) to ETS. The proposed design was motivated by a set

[4] It is worth noting that Keppel's reporting to Congress in the 1960s was soon after the Soviet Union had launched a satellite into space for the first time. This achievement by a Cold War adversary resulted in a growing desire in the U.S. to invest in education with a renewed emphasis on mathematics and science.

of policy questions that focused on the need to examine the extent to which all groups of students were being equally well prepared and whether what they were learning would enable them to contribute to the nation's economic and social development in the 1980s and beyond.

In order to address these questions, Messick et al. (1983) argued that NAEP had to abandon its original design and methodologies that yielded interpretations of the data that were tied to individual items and constrained by the limited student demographic information available. Central to the new design was the introduction of BIB-spiraling, which is a combination of a balanced incomplete block design (a powerful variant of matrix sampling) and random allocation of specially designed item blocks (constructed from a large pool of items) to students within a school (Beaton and Barone 2017). The principal goal was to broaden the item pool used to measure each target construct without increasing student response burden. They also advocated the use of item response theory (IRT) for the analysis of the response data (Lord 1980; Carlson and von Davier 2017). IRT has the advantage of enabling the creation of a common proficiency scale across the multiple forms of an instrument that result from a BIB design. In contrast to item level data, scales constructed from sets of items are more stable and provide a stronger basis for making comparisons across groups or individuals. In addition to IRT scaling, technical innovations included marginal estimation procedures that optimized the construction of the proficiency scales based on data collected through complex designs along with the introduction of plausible value methodology in order to estimate the measurement error associated with NAEP estimates (Mislevy 1987, 1991; Mazeo et al. 2006; Braun and von Davier 2018).

Large-scale assessments of student populations

Over the next 20 years, the methodologies and techniques developed for and implemented in NAEP in the 1980s, were adopted and adapted by organizations conducting national and international surveys. These innovations contributed to the growth in both the number of such surveys and the number of participating countries that, in turn, stimulated increased interest among policy makers, researchers, and other key stakeholders. Examples include studies conducted by IEA such as TIMSS (Trends in Mathematics and Science Studies) and PIRLS (Progress in Reading Literacy Study), as well as PISA (Program for International Student Assessment) conducted by the Organisation for Economic Co-operation and Development (OECD). These surveys not only obtained estimates of the distributions of key foundational skills among students at various grades and ages, but also enabled the examination of these skills in connection with student demographic characteristics, as well as teacher, school and home variables.

Assessments such as TIMSS, PIRLS, and PISA are arguably the largest and most widely discussed comparative international assessments. Although they were developed in response to the interests and participation of developed countries, they have continued to expand with the participation of low- and middle-income countries. The participation of these countries is primarily facilitated by the financial and technical support provided by international donor organizations such as the World Bank, the United

Nations Development Program (UNDP), and the United Nations Educational, Scientific and Cultural Organization (UNESCO).

In addition to the expansion of large international surveys, a number of smaller, regional assessments began in the 1990s (Lockheed 2013; Wagemaker 2014). For example, organizations affiliated with UNESCO conduct three additional regional comparative studies: One is the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ); a second is the Programme d'analyse des systèmes éducatifs de la CONFEMEN (PASEC), which is a program designed to assess student abilities in mathematics and reading French; and, the third is the Latin American Laboratory for Assessment of the Quality of Education (LLECE). SACMEQ, involving 15 countries, focuses on understanding and improving the educational outcomes in the Anglophone Africa region. It assesses grade 6 students in reading and mathematics. LLECE is similar in design but assesses students in grades 3 and 6 in Spanish- speaking countries in Central and South America and the Caribbean. PASEC is typically administered to students in second and fifth grades at the beginning and end of the same school year and is designed to assess student growth in 13 Francophone West African countries.

Large-scale assessments of adult populations

Beginning in the 1990s, policy makers expressed a growing appreciation for the critical role that skills acquired through education, workforce participation, and lifelong learning play in the outcomes for adults and the societies in which they live.[5] As a result, they generated a new set of questions that included: how do skills relate to participation in the labor force; how are educational attainment and skills related to one another; how are essential skills (e.g., numeracy and literacy) related to outcomes such as health and well-being; and, what is the relationship between essential skills and the ability to benefit from employer-supported training and lifelong learning? In an effort to answer such questions, a series of international adult assessments were conducted under the auspices of the OECD and implemented through household surveys. The first was the International Adult Literacy Survey (IALS), a survey of adults 16–65 years of age conducted in multiple rounds from 1994–1999. It was followed by the survey of Adult Literacy and Life Skills (ALL), conducted between 2003 and 2008 (Kirsch and Lennon 2017).

Both IALS and ALL were designed to estimate the distributions of literacy and numeracy skills among adult populations in participating countries. Unlike the school-based surveys that focused on particular age- or grade-based samples of in-school populations, IALS and ALL targeted both in- and out-of-school adults. These paper-based assessments included a background questionnaire administered through individual interviews that provided researchers with data to explore the relationships among skills and important educational, social and labor market outcomes.

In 2012, the Programme for the International Assessment of Adult Competencies (PIAAC) was launched and has been conducted in several rounds through 2018 (Kirsch and Lennon 2017; Kirsch et al. 2017; Kirsch et al. 2020). Like IALS and ALL, this is a household survey of adults ages 16–65 and includes a cognitive assessment and

[5] In 1990, Elizabeth Dole, then Secretary of Labor, established the Secretary's Commission for Achieving Necessary Skills (SCANS), and charged it with identifying the skills needed by young people for the modern workplace.

extensive background questionnaire. Unlike IALS and ALL, PIAAC is a computer-based assessment focused on assessing literacy, numeracy and problem solving and is designed as an ongoing program conducted every 10 years, with new countries being able to join in new rounds of data collection within each 10-year cycle. The first cycle of PIAAC began in 2008 and was followed by two additional rounds in 2014 and 2017. A second PIAAC cycle began in 2018.

The work done (and lessons learned) in developing and implementing IALS and ALL informed the development and implementation of PIAAC. Many of the processes and procedures used in assessment design and item development, along with procedures associated with translation and adaptation of both the background questionnaire and cognitive instruments, were refined based on the experience gained from these earlier surveys. In addition, analytical procedures were enhanced to allow for the treatment of differential item functioning by country. From IALS through PIAAC, each assessment expanded what was measured. For example, the development of IALS was informed by the National Adult Literacy Survey used in the United States (Kirsch et al. 1993) that measured prose, document and quantitative literacy, while ALL included the prose and document literacy scales, but replaced quantitative literacy with numeracy and added an optional measure of problem solving. In addition, the international surveys substantially extended the range of information collected through the background questionnaires.

Much like the 1983 design for NAEP, which put large-scale assessments on a new trajectory, PIAAC marked the beginning of a new and significant cycle of innovation. As the first computer-based international large-scale assessment, PIAAC expanded what could be measured in order to better reflect the changing ways in which individuals access, use, and communicate information. In addition, PIAAC introduced methodological innovations, such as multistage adaptive testing and more flexible routing for the background questionnaire, that improve data quality and assessment efficiency and laid the foundation for future computer-based assessments (Kirsch et al. 2017).

Messick's framework for evaluating the utility of large-scale assessments

In 1987, Messick published a short but illuminating paper in which he addressed the proper role of large-scale assessments. He asserted that large-scale assessment constitutes a form of policy research and its success should be judged on the basis of its policy utility. Building on the work of Lerner and Laswell (1951), he noted that a degree of uncertainty always exists when moving from research to action, due to context effects and the contingent relationships among diverse factors. This uncertainty creates a gap between policy research and policy formulation—a gap that must be bridged by expert judgment, which is the responsibility of policy makers. Messick argued, therefore, that large-scale assessment is a type of policy research with a primary role of informing this judgmental process through the evidence generated by careful design and development of instrumentation, followed by proper reporting of results. For large-scale assessments to fulfill this role, Messick proposed three design criteria. Together they constitute a valuable framework for thinking about critical aspects in the design, development, and implementation of both student and adult assessments. The design criteria are *comparability, interpretability,* and *relevance.*

According to Messick, *comparability* refers to the extent to which skill measures (represented by score distributions) can be meaningfully compared across participating groups both within and across assessment cycles. He argued that reporting scales derived from the application of IRT, as proposed in Messick et al. (1983), yielded psychometrically sound scales that substantially improved such comparability over the item-level reporting that heretofore had been the hallmark of NAEP.

Interpretability refers to the extent to which the reported scores can be meaningfully related to the target construct and to various background characteristics or contextual factors. Messick argued that the combination of BIB spiraling and IRT scaling, along with an expanded background questionnaire, permitted the calculation of intercorrelations among all exercises in the enlarged pool of exercises, as well as between overall performance in the cognitive domains and the full set of background characteristics, attitudinal variables, and contextual factors. Patterns in the intercorrelations provide the basis for meaningful interpretations of the data.

Finally, *relevance* reflects the extent to which the estimates of the relationships between the cognitive measure(s) and the various constructs targeted by the background questionnaire can inform policy judgments and decisions. Singly and jointly, comparability, interpretability, and relevance contribute to policy utility.

These design criteria are key to understanding the rationale for the methodological innovations introduced into NAEP to address the concerns regarding its apparent lack of utility in the context of the changing educational, social, and political landscape of the time (Messick et al. 1983). In relatively short order, these innovations sparked an increased interest in national assessments such as NAEP and, in due course, greater attention to, and increased participation in, the international assessments conducted by IEA and the OECD.

Some 30 years later, paper-and-pencil instruments have begun to give way to new digitally based assessments. These assessments build on the past, but also represent a response to increasing globalization and the rapid increase in the use of new technologies to access, use, and communicate information. Globalization refers to the increased interdependencies (and competition) among countries that drives the growing salience of education and skills to their economic and social development. Digital content and tools impact how we learn, work and interact with this new form of information and, therefore, is increasingly important to understand how students and adults use such materials. Technological innovations also present new opportunities in the assessment context, including 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, multilinguistic, multidimensional data.

Although Messick's original framework for understanding and evaluating the utility of large-scale assessments grew out of the NAEP design and thinking about the types of information it could provide for policy makers and key stakeholders, we believe that this time of transition is an opportune moment to discuss and explore how these three design criteria apply in the era of digitally based assessments. In that light, we acknowledge that Messick's three design criteria have not changed—they remain valuable but can be extended or expanded to better reflect today's context.

Design criteria of the framework extended

The principal rationale for the very substantial global investment in ILSAs is the need to generate credible, policy-relevant evidence that is useful to a broad and growing range of stakeholders. Following Messick, we believe that the design criteria of comparability, interpretability, and relevance can function as a powerful organizing structure to guide the design, development, and implementation of all large-scale assessments. An overall judgment of an assessment's policy utility is largely dependent upon evaluating its strength with respect to these design criteria, with due regard to purpose, context, and the population(s) of concern. As we shall see, the requirement to meet these design criteria is particularly germane to the criteria of relevance and interpretability. We now describe our extensions of these criteria and discuss the impact of technology on each of them—with the goal of indicating how policy utility can be enhanced.[6]

Comparability refers to the degree to which the results obtained in different jurisdictions have the same meaning in relation to the underlying constructs being assessed. Achieving comparability in an ILSA context 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.[7] In the first instance, every effort must be made to ensure that the samples of respondents in the different jurisdictions are approximately equivalent in their representativeness of the corresponding target populations. Success depends on a number of factors, including the quality of the auxiliary information available for the survey design, the degree of cooperation among sampled units, and the fidelity of implementation. Each ILSA defines the target population along with the standards for determining how well the sample represents the target population. For those instances where standards are not met, countries' data may undergo further analyses and, in some instances, the data are not included in the final report.

Further, the cognitive instruments and background questionnaires undergo an extensive process of development, translation/adaptation, verification, and review so that the data generated have equivalent meanings (measurement invariance) across cultures and languages (von Davier and Sinharay 2014; Dept, in press; Kuger et al. 2016; van de Vijver 2018). Following data collection, psychometric analyses reveal the extent to which the goal of measurement invariance has been achieved. If certain items appear to function substantially differently in a particular jurisdiction, appropriate adjustments are made to the psychometric models that are used to estimate item parameters and generate the reported scale scores for that jurisdiction. Additionally, recent surveys have developed IRT-based scales for the background questionnaires, where items within each scale are evaluated using similar standards to those applied to the cognitive scales. Comparability across countries depends on the invariance of item parameters for the overwhelming majority of items from each reported domain. (OECD, 2016, 2017).

Comparability also has a chronological dimension; namely, that cognitive scale scores, as well as composite scale scores derived from the background questionnaire, can be

[6] Policy utility associated with ILSAs is not guaranteed. It requires intentional, strategic efforts to think about and apply advances that are taking place in measurement science, psychometrics, and survey operations in connection with technology. Each of the three design criteria can be characterized by a set of features that define them operationally.

[7] In TIMSS 2015, for example, 50 countries and 7 benchmarking jurisdictions participated in the 4th grade mathematics assessment. PISA 2018 had 79 participating countries/economies.

meaningfully related to scores from earlier administrations. In the cognitive domain, this generally entails carrying out scale linkage procedures that place trend items and new items from the current administration on the established scale. On occasion, a new scale is defined and the items from previous administrations are placed on this new scale. In either case, the relevant procedures are well known, as are the checks on the validity of the linkage (Mazeo and von Davier 2014; von Davier et al. 2019). 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. If significant changes are made, it may result in the need to establish a new scale, rather than treating it as a continuation of an earlier one.

Interpretability, in our view, depends on the extent to which the instruments administered have been developed through a fully coherent process and, as a consequence, the reported scores can be given substantive or normative meanings that are credible, defensible, and accessible to a range of stakeholders. By the term full coherence, we mean that the key processes associated with design, development, delivery, scoring, scaling, and reporting are not only each appropriately linked to the intended measurement goals, but also functionally integrated through continuing collaboration among different specialists. Thus, the term involves much more than "face validity" and goes well beyond the kinds of correlational relationships cited by Messick (1987). Rather, it implies strong, evidence-based support for the desired interpretation(s) or, in other words, demonstrable construct validity (Messick 1989).

PIAAC provides an instructive example of a coherent process for assessment development. Although it is likely that other ILSAs, including TIMSS, PIRLS, and PISA, follow similar processes, we are most familiar with the one carried out in the first cycle of PIAAC (OECD, 2016).[8] With regard to the development of the cognitive instruments, separate subject matter expert groups were established for each of the cognitive domains (e.g., literacy, numeracy, and problem solving in technology rich environments) and maintained throughout the PIAAC assessment cycle. Among other things, these experts had responsibility for creating an assessment framework that reflected current thinking in the field, while keeping in mind the goal of developing an instrument that would generate evidence valid in a cross-cultural context.

Taking a construct-centered or evidence-centered approach to assessment development requires agreement on an operational definition of the target construct, along with guidelines as to the nature of the evidence needed to locate individuals along a continuum or scale that is linked to the construct. Then follows specifications for item development, including the identification of key task characteristics to be varied singly and jointly (Mislevy et al. 1999; Messick 1994). For example, in the literacy assessment, the process included efforts to specify both the different purposes for reading and the types of texts to be employed. A set of context and content areas also was identified with target distributions specified for each area so that the instrument would involve variation in language structures, vocabulary, and background knowledge. These characteristics

[8] Throughout the paper, the authors tend to rely on examples from PIAAC. Many of the issues raised in the paper are relevant to all large-scale assessments. However, since PIAAC is the first ILSA to rely on a computer platform to design, develop, and deliver the assessment and because it is a household survey, it requires some processes and procedures that differ from school-based surveys.

were manipulated in different combinations by test developers to construct item sets that together fully represented the assessment framework and would elicit test takers' responses providing desired evidence about their knowledge and skills in the domain being assessed.

Once the pool of assessment items for a domain was developed, it was submitted for review and comment from the participating countries and the expert group. The experts also assisted in selecting existing items to serve as links with earlier adult surveys. They then mapped this augmented item pool to the framework to ensure full construct representation. This process resulted in a preliminary pool of field test items that underwent expert evaluation for translatability. This quality assurance process led to the identification and correction of potential translation issues before the items were submitted to participating countries for translation and adaptation.

The final pool of items was then administered by countries based on the field trial design. All national data were then processed and analyzed. The field trial analyses were used to evaluate the quality of the instruments and to recommend a final set of instruments to be deployed in the main study. Subsequent to the main study and application of scaling procedures, the expert group collaborated with test developers to create a "described proficiency scale." This activity employed the task characteristics identified in the framework, the performances of participating adults in each country and an item map[9] to create descriptions of the tasks that fall along different points of the established difficulty scale. Such descriptions made it possible to go beyond simply identifying that one item was more difficult than another and, instead, define levels of performance by articulating how the skills and knowledge required to complete the items progressively increased along the scale.

The primary goal in following a coherent process for assessment development is to enhance interpretability of the findings. In this regard, the assessment framework plays a key role. First, by providing a common language and an organizational structure, it serves as a vehicle for building consensus around the definition and approaches to measurement of the construct. Linking the framework to the evidence that is to be collected through the assessment leads to a deeper understanding of what is actually being measured. Once the reporting scale has been established and validated, this understanding can be employed 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, including the public-at-large.

Relevance is the extent to which (i) the evidence elicited by the cognitive instruments and the background questionnaire is germane to current policy questions and decisions, and (ii) the assessment design yields results that can be analyzed in such a way as to address current priorities. Relevance is strongly dependent on both comparability and interpretability—that is, deficiencies in either can directly undermine the utility of the data in addressing the questions of interest. Judgments of relevance are made by the various stakeholders in each jurisdiction, as well as by secondary analysts, and

[9] An item map is created by associating individual items with points along a scale, based on their statistical properties. Additional information about item mapping can be found at: https://www.nationsreportcard.gov/itemmaps

are contingent on the particular purposes at hand. The voluminous technical reports accompanying an ILSA provide useful descriptions and data to inform those judgments (OECD 2016, 2017).

In addition to assuring comparability and interpretability, ILSA designers adopt different strategies to enhance relevance. One example is improving construct representation of "legacy" constructs by introducing new item types that target neglected facets of that construct. Reading literacy is a good example as the introduction of digitally based assessments facilitates the introduction of electronic texts.[10] In some cases, texts are adapted from paper-based assessments for use in electronic environments. New items developed to reflect updated literacy frameworks are often based on continuous and non-continuous texts associated with digital environments, including web pages with links, emails, and interactive spreadsheets. Other examples of efforts to enhance relevance are the development of assessments for new constructs, such as collaborative problem solving, or developing measures of certain aspects of learning contexts (or other background factors) that research indicates may be associated with cognitive skills.

Transitioning to digitally based assessments

Successful implementation of digitally based assessments (DBAs) involves the introduction and integration of new tools, workflows, and processes that require the ongoing collaboration of experts across the different phases of ILSA, including assessment design, instrument development, translation and adaptation, survey operations and management, data collection and processing, and scaling and analysis. The digital platform that is developed or used to incorporate and support these innovations serves to both centralize and standardize the many activities that were previously carried out independently by participating project teams having various degrees of background and experience. Innovations brought about through the introduction of a digitally based platform lead to overall improvements in data quality, efficiency, and the scope of what can be assessed, thereby enhancing policy utility.

Although discussing all the innovations that have been introduced by ILSA that have already transitioned to DBA is beyond the scope of this paper, several are identified and discussed here. They relate to the new tools, processes, and workflows that: (i) improve the overall quality of the data; (ii) extend what can be measured so that what is assessed better reflects how survey participants access, use, and communicate information; and (iii) increase the overall efficiency of the management, delivery, and processing of the data. Taken together, the innovations that accompany the transition to DBA will continue to improve ILSA with respect to the design criteria of comparability, interpretability, and relevance—thereby enhancing the overall utility of ILSA (Rutkowski et al. 2014; von Davier et al. in press). These innovations in the development and/or expansion of the digital platforms used to manage, design, develop, and deliver ILSA to both student and adult populations include:

[10] It should be noted that as the framework for a given construct evolves, analyses need to be undertaken to check for the dimensionality of the scale and the corresponding changes must be made to the interpretation of proficiency levels.

- Electronic tools that are used to manage key processes and workflows of the survey that can add to data quality and strengthen comparability;
- Processes and procedures to enhance translation and adaptation, designed to improve comparability of the assessment instruments across participating countries;
- New item types such as scenario-based tasks in reading and simulation tasks in science to enhance construct representation, which contributes to the enhancement of relevance and interpretability;
- Creation of more complex assessment designs yielding data to address key policy questions that are aspects of the ILSA goals and, thus, add to data quality and interpretability;
- Enhancements to data capture and data processing systems designed not only to improve efficiency, but also to enhance data quality and perhaps interpretability; and,
- Introduction of dashboards, which can display paradata to monitor and manage administration of the survey, improving the identification of potential survey administration or response problems in real time.

Managing ILSA

With all ILSAs, there is a need to balance overall quality with project constraints, including costs and timelines; that is, the challenge is to choose a set of features and parameters to minimize total survey error[11] while meeting the overall goals of the project and respecting the externally imposed constraints. A digital platform can and should support the management of tasks needed to communicate with participating countries and contractors about specific activities. These include three basic functions: 1. centralized monitoring of tasks by which both country progress can be tracked and any potential problems with respect to timelines or standards can be identified; 2. content management so that item development files can be shared; and, 3. item previewing in which authored items can be examined in both source and target language(s) along with item layout. The platform also needs to support the development of the instruments (i.e., the cognitive items and the background questionnaires) that are included in the survey. This requires building functionality to support the design and delivery of the various item types, as well as the management of the various workflow processes associated with the instruments.

Translation and adaptation of the instruments

A key aspect of instrument development is the workflow associated with translation/adaptation and verification of the cognitive items and the context questionnaire(s). Although the instruments are typically translated by national teams, each language version undergoes verification by an independent expert in that language to assure the comparability of items across the many language versions. The platform thus needs to be able to accommodate the full range of languages used by participating jurisdictions

[11] Total survey error provides a framework that covers all types of errors that may arise in survey design, sample selection, data collection and processing, scaling and analysis, and the creation of data products. In putting this framework into practice the goal is to enhance data quality by reducing the various sources of sampling and nonsampling errors while operating within stated constraints, such as time and budget.

(including right-to-left and ideographic languages), along with the workflow needed to support all the specific features of the assessment. If robust linguistic quality assurance and quality control processes are implemented in an ILSA, they must rely on the coordinated work of test developers, linguists, and cross-cultural survey methodologists who produce informative translation and adaptation notes that both explain the underlying constructs and offer additional guidelines for use during the test translation and adaptation process (Dept, in press).

In particular, the final version of the item-by-item translation and adaptation notes (i) explains what the item is intended to measure; (ii) specifies which adaptations are mandatory, desirable, acceptable, or ruled out; (iii) draws the translators' attention to terminology problems, translation traps, patterns in response options and, for the cognitive items, provides 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; and (iv) in the case of recurring elements or elements already present in trend materials, indicates how to access previous translations of these segments.

The DBA platform makes it possible to import the translation and adaptation notes into the translation documents so that the notes appear in the translation tool when a translator (or reconciler, or verifier) processes a text segment. This is a technical innovation offering significant added value to the national translation teams—by streamlining processes, reducing the number of documents and tools required, and providing a translation environment that unites all relevant information and enables translators to better focus on key elements of the translation task—at no additional cost for countries.

DBA affords other technological innovations such as bilingual glossaries, which are useful when there are recurring terms and expressions that should be translated consistently, including standardized user instructions or prompts. Some organizations have developed searchable translation glossaries for various target languages that help to ensure consistency among participating ILSA countries. There are even web-based applications that can be used to verify consistent adherence to these glossaries. These are further examples of how technology can be used to improve the overall quality and efficiency of the translations/adaptations that are so vital to establishing comparability.

Introducing more complex assessment designs

Another function of the platform is to support the delivery of novel assessment designs. The advent of DBA has enabled the introduction of more complex designs that include more sophisticated routing in the questionnaires and the introduction of multistage adaptive designs for the cognitive instruments (Yamamoto, Shin et al. 2018). Measuring trends remains a key ILSA goal and, in order to meet this goal, the platform has to be able to handle existing items used in earlier paper-based cycles This presents challenges including replicating or adapting paper-and-pencil response modes, maintaining similar formatting and display for stimuli and, in some cases, supporting a move from human-scoring to computer-scoring. Such adaptations require investigation into potential mode effects and their impact on trend measurement. The platform must also support the implementation of new item types that are responsive to changes in the conceptualization of legacy constructs such as literacy and numeracy in PIAAC or reading,

mathematics, and science in PISA, PIRLS and TIMSS. Moreover, the platform must support the introduction of new constructs such as problem solving in technology rich environments in PIAAC.

With multistage adaptive testing, not only is the precision of estimates enhanced, but also it is possible to increase both the number of domains assessed overall and the number of domains assessed for each respondent. In this setting, however, the total number of domains will sometimes be somewhat greater than the number assessed per respondent (e.g., sometimes five domains overall in PISA but each student may respond to only two or three of these domains). This is due to constraints on sample size and/or respondent assessment time.[12] In any case, these more complex designs require more sophisticated statistical/psychometric models to carry out the calculations yielding the plausible values that are the basis for reporting and analysis. Because the technical challenges are considerable, measurement specialists are devising new strategies for carrying out differential item functioning analyses, as well as better measures of general model fit with accompanying diagnostics (Yamamoto, Shin et al. 2018; Yamamoto, Khorramdel et al. 2018; Chen et al. 2014; von Davier et al. 2019).

Inter-disciplinary teams are also working on modifying the structure of the background questionnaire so that it more resembles that of the cognitive instruments; namely, implementing a version of the BIB design so that no respondent takes the entire background questionnaire, although there would be a core set of questions administered to all. The implementation of this design facilitates the introduction of additional constructs with benefits for secondary analyses, without increasing the overall assessment time. On the other hand, such changes will have implications for the complex analysis methods (i.e. population modeling for generating plausible values) that are used in ILSAs. These implications must be understood, and appropriate modifications made to the models in order not to introduce unwanted bias into the results (von Davier 2014). It is worth noting that PISA is planning to introduce and evaluate the use of within construct rotation in the student background questionnaire for the 2021 cycle.

Enhancements to data capture and processing systems

Transitioning to DBA provides an opportunity to capture and employ important information associated with the assessment. The platform's role in recording and processing assessment data includes aspects or features related to the evaluation of item responses that are to be computer scored. The platform also needs to be able to provide support for the human scoring of open-ended items.

As item types with constructed response formats become more prevalent, automated scoring of such responses will be necessary to contain costs and maintain tight time schedules. To accomplish this, and to achieve greater efficiency, expert systems are being developed to carry out scoring for classes of item types (Lubaway et al. 2019). This requires close collaboration among test developers, psychometricians, content experts, computational linguists, and IT professionals. In fact, the demands on item development due to the requirements of the scoring engines can lead to improvements in item quality

[12] Having a goal to increase both precision and complexity of an assessment in terms of the number of domains being assessed likely requires an increase in sample size.

through greater attention to item models and their implementation (Mislevy et al. 1999; Bejar and Braun 1994).

The platform also needs to keep track of log-file data. This includes detailed timing information, which is easily captured in a DBA and can be used during data analysis to identify omitted items and items where responses appear to be guesses because they occur unusually quickly. Additional log-file data includes respondent behaviours and actions. Relevant events include any action or milestone during the course of a respondent's performance that the test developers or psychometricians believe is important for further review and analysis. Some events may be at a low level, such as recording the x–y coordinates of each mouse click or tapping location, while some may be domain-specific, higher-level actions such as those that reflect strategies or behaviors employed by students or adults responding to individual items or tasks (Kane and Mislevy 2017).

Beyond data capture, the platform must also be capable of exporting information into various software packages that accommodate country-specific data files—necessary because in some instances countries are allowed to adapt or add unique national items to the international background questionnaire. The software that has been developed to carry out this important function contributes to improved efficiency with respect to both time and cost, as well as to overall data quality.

Using paradata and dashboards to monitor and manage survey quality

A critical aspect of improving the quality of data from large-scale household assessments such as PIAAC is to establish processes that detect various sources of non-sampling error during data collection and to remedy them when possible. Such processes are usually developed by exploiting the information contained in the paradata (Mohadjer and Edwards 2018). Paradata are the survey-related data that are produced in the course of data collection. In the case of a household survey such as PIAAC, examples include the record of contacts information, instrument timings, voice recording of interviews, geo-location, and interviewer work activities (hours and travel routes). It provides indicators of data quality, costs, and interviewer effectiveness that can help survey managers react to operational challenges in a timely and well-informed manner. The examination of paradata in combination with survey data facilitates real-time error detection, thus increasing both accuracy and efficiency.

Paradata files may contain large amount of data, some of which may be unstructured. Thus, paradata need to be "mined" and presented in formats that makes it accessible to users. For example, performance dashboards for survey operations are a collection of control charts and statistical graphs arranged to present the data collection status in real-time, monitor sample yield and response rates, and highlight unusual outcomes. Such dashboards highlight the key process indicators and provide, at a glance, summary graphics of the indicators of interest, suggesting where there may be a need for further investigation or intervention.

Policy utility

In his discussion of policy research, Messick (1987) noted "…that in many instances its concrete forecasts are contingent upon variable and uncontrolled conditions." (p. 157). Consequently, "… if implications for action are to be drawn from the findings, one must

appraise the likelihood that relevant other factors will indeed remain constant and then empirically assess any changes in these factors and their likely impact on action alternatives" (pp. 157–158). He went on to argue that if large-scale educational assessments were to function effectively as policy research, they would have to possess certain characteristics exhibited in concert; namely, comparability, interpretability, and relevance. Implicit here is the contention that enhancement to one or more of these facets results in greater policy utility and, conversely, that degradation of any of the facets lessens policy utility. In the previous section, we have argued that the transition to DBA has generally led to such desirable enhancements, resulting in greater policy utility.

For example, improvements in both construct representation[13] and basic data quality, as well as reductions in measurement error through adaptive testing, yield better estimates of the distributions of cognitive skills, along with the relationships of those skills to various background characteristics and social/educational factors. This is particularly so for countries scoring at the lower ends of the proficiency scales and, consequently, are not well served by the ILSA equivalent of fixed form tests. Most germane for those countries, adaptive testing enables them to make meaningful distinctions among subpopulations of interest that can inform such policy decisions as resource allocation.

For adult populations, ILSA score distributions disaggregated by age and gender offer a more refined picture of a country's human capital landscape. Using demographic projections, it is possible to project score distributions into the future and to evaluate the discrepancy between what will be needed and what is likely to be available (Kirsch, Braun et al. 2007). Methodological advances facilitate making such projections at more granular levels.

Further, comparing score distributions at different levels of educational attainment (again disaggregated by age and gender) can provide policy makers with useful information regarding the contributions to skills provided by additional years of education. Indeed, for almost all countries, adult surveys reveal substantial overlap in the distributions of skill proficiencies at adjacent levels of educational attainment (e.g., high school diploma and bachelor's degree). This is not only rather surprising, but also suggests the need to examine more closely graduation standards at the secondary and tertiary levels. Moreover, the variability in skill proficiencies at each attainment level facilitates study of the returns to skills in a more fine-grained and, presumably, more policy-relevant manner (Kirsch et al. 2007; Fogg et al. 2019).

At the same time, we note that the implications for action of ILSA findings will depend on factors that vary across countries (or jurisdictions). Consequently, cross–country patterns of relationships, no matter how striking, should be supplemented by careful study of country-specific conditions, cultures, and policy history in order to appraise the implications for action (Meyer and Schiller 2013). Further, as Ritzen (2013) points out, realization of policy utility depends on the readiness of the relevant authorities to effect changes and to appropriately employ ILSA findings to inform the decision-making process. He offers numerous instances drawn from OECD countries, as does Lockheed

[13] It is important to note that improvements to construct representation includes both the background/context questionnaires as well as the cognitive domains and refers to enhancements to existing constructs as well as the introductions of new constructs.

(2013) with examples drawn from low- and middle-income countries. Thus, in point of fact, policy utility represents the *potential* for constructive use of the basic data generated by the ILSA, as well as of the patterns of relationships derived from that data.

Of course, there are numerous examples of the impact of ILSA findings on national education policies (Braun and Singer 2019, Heyneman and Lee 2014; Wagemaker 2014). For example, Wagemaker (2014) cites countries as disparate as Singapore and Qatar as having used their results from PIRLS and TIMSS to enact policy changes in relation to curriculum, pedagogy, and assessment. In a similar vein, Heyneman and Lee (2014) documents the impact of PISA on education policy in a number of countries. Perhaps the best-known example is Germany, a country that suffered a "TIMSS shock" followed by a "PISA shock." The relatively poor performance of German students prompted high-level changes in policy that resulted in significant reforms in the educations system of all German states (Ertl 2006). Other countries that also undertook major reforms in response to ILSA findings include Canada (Francophone), Iceland, Ireland, and Mexico.

ILSA results are also used for secondary analyses that can uncover patterns of relationships that either directly inform policy or suggest promising directions for further study. Using PIAAC data for OECD countries, Braun (2018) showed that women working full-time were much less likely than men working full-time to have incomes in the top quartile of the national income distribution, even after controlling for family background, measured cognitive skills, educational attainment, and occupational sector. The degree of disadvantage ranged from modest (United States) to very substantial (Japan, Netherlands). However, as noted above, implications for action depend on deeper, country-specific analyses (Schleicher 2018; Maehler et al. 2018).

Policy utility is manifested in other ways. For one, international funding agencies, such as the World Bank and the Inter-American Development Bank, often support the participation of lower income countries in which they have made education-related investments (Lockheed 2013). Donors rely heavily on ILSA results to monitor the success of their investments, particularly since locally produced data are sometimes problematic or even nonexistent. Many of these countries also take advantage of participation in ILSA (including regional large-scale assessments) by using it as an opportunity to build infrastructure capacity and expertise among education ministry staff and others (Wagemaker 2014). Indeed, with regard to a particular construct, its assessment framework, the associated documentation, as well as the released items, are valuable resources for countries that want to develop instructional and assessment capacity in relation to the construct domain.

Policy utility is enhanced (through greater relevance) when ILSAs are responsive to the emerging interests of stakeholders. One such interest is how proficiency in technology-enabled modes of information transmission differs from proficiency in more traditional modes (e.g., electronic reading vs. paper-based reading). Note that electronic reading comprises not only traditional text formats, but also hypertexts and information search in various web environments. As the conception of a legacy construct evolves, new items and item types are needed to adequately represent the evolving nature of the construct.

Sometimes achieving greater relevance requires a new assessment design and novel item types that can accommodate the measurement of an additional construct such as collaborative problem solving in PISA or problem solving in technology rich

environments in PIAAC. The administration protocols then must be modified so that there are sufficient sample sizes to estimate the relationships between all pairs of constructs, as well as their relationships to individual characteristics and contextual factors. Furthermore, in order to more fully realize the value of assessing the new construct, it is necessary to expand the background questionnaire to include new items related to a student's or adult's attitude toward, or use of, the construct at school, at home or at work.

Moreover, there is also the desire to include new constructs in the background questionnaires. A recent example is the plan for the new cycle of PIAAC to include a construct related to social-emotional learning. Incorporating a new item set to adequately assess this construct within the fixed time allocated to the overall assessment can impact the overall design and flow of the background questionnaire.

Longer term, ILSA policy utility also can be enhanced by augmenting the data collected through linkages with other databases. In the United States, the American Community Survey is one such repository. In Scandinavian countries, national registers contain very detailed information on individuals. With appropriate confidentiality safeguards, secondary analyses could be substantially enriched through expanding the data available at the individual and contextual levels, including school districts and larger administrative areas such as counties and states (Krenzke et al. 2020).

These examples nicely illustrate a general pattern: As the transition to DBA evolves, ILSA are likely to generate more information with greater relevance to stakeholder queries. This, in turn, leads to novel questions, placing new demands on ILSAs that require further innovations in tools and procedures. Such ongoing interactions between ILSA users and ILSA sponsors/contractors should lead to a long-term increase in ILSA policy utility.

Challenges

We argued earlier that the transition to DBA, incorporating technology-based tools and processes, has enabled contractors to increase efficiency in order to accommodate the substantial growth in the number of participating jurisdictions with concomitant increases in heterogeneity in tested languages (in the case of international assessments) and in distributions of proficiency. This has been accomplished often under the constraints of tight budgets and fixed reporting schedules, and all without compromising data quality. However, satisfying at least some of the future demands on the ILSA system may have to be delayed until constraints are relaxed or new tools and procedures are developed.

One example is the desire to deliver the assessment on multiple devices in order to be responsive to local capacities and resources. However, each device's technical characteristics determine its unique affordances and limitations for assessment delivery, data capture, and storage. Differences among devices can affect the nature of the assessment experience, introducing unwanted construct-irrelevant variance and, hence, reducing comparability. Although there may be some clever technical "fixes" that can mitigate the impact of some of these differences, eventually sponsors and participating jurisdictions will have to consider tradeoffs between flexibility and convenience on the one hand, and comparability on the other.

In the case of low- and middle-income countries, limited access to technology, to the internet and electricity, a lack of familiarity with computers or tablets, can all pose challenges to the implementation of DBA. As a result, countries are often given the option to participate in a large-scale assessment using a paper-and-pencil version of an instrument. When this occurs, it raises the issue of comparability over time as new development work typically focuses on new constructs (or new aspects of existing constructs) and expanded capabilities that can be supported in DBA. This means that new texts and items cannot be replicated in paper-based assessments. Therefore, over time, the link between the paper-based and digitally-based versions of an assessment grows weaker from both a statistical perspective, as well as from a construct point of view.

Another challenge arises with increased concerns with data privacy. Although the public's attention has focused primarily on medical records and social media, storage of, and access to, education data is also of concern. While some states (e.g., California) have passed relevant legislation, federal laws and regulations have yet to be promulgated. In the meantime, the European Parliament has passed the General Data Protection Regulation that is now in effect. It places substantial burdens on entities generating and storing data with respect to maintaining privacy and confidentiality of data. It remains to be seen how this and other regulations will affect access of secondary analysts to individual-level data collected by ILSA and, hence, the policy utility of the data.

Arguably, the full potential of ILSAs' data has not been realized. In particular, more could be done to communicate to educators at all levels about the nature of the constructs targeted by an ILSA and its implications for various stakeholders, including teachers. Typically, this type of information is available either through technical reports or some other publication. For example, ILSAs often make available to the public the full framework that was used to develop the assessment for each cognitive domain or an abridged version containing all the frameworks used in an assessment cycle (OECD, 2012). As described above, the framework can contain important information related to how the domain has been defined and operationalized. This is particularly true if the developers followed a construct-based or evidence centered design (Mislevy et al. 1999; Messick 1994). With this approach, the developers will have identified those key characteristics relating to the constructs that are used to develop item models. From the resulting pool of items, a subset of items is then assembled based on the relative weighting or importance of the features that have been identified to represent each of the task characteristics. This information provides the stakeholders with a deeper understanding of what is being measured and, hence, contributes to interpretability (Kirsch 2001).

Many argue that we are at the early stages of making good use of the log-file data that digitally based assessments provide. Many of the uses have been limited to obtaining timing information that has thus far been employed for information related to data quality (Yamamoto and Lennon 2018). This information is also being used in limited ways to improve item quality and to examine respondent motivation (Goldhammer et al. 2016). Some investigators are beginning to use these data for understanding response strategies and patterns as a way of adding meaning to what these assessments are measuring (Ercikan and Pellegrino 2017; Goldhammer et al. 2017; Greiff et al. 2015; Qiwei and Dandan 2019). At present, this work is in its early stages and, consequently, there is no clear agreement as to what data should be tracked and how it should be used. Carrying out the next

stage of analysis will require funding to initiate and sustain joint efforts by the sponsors, contractors and jurisdictional leaders to develop papers or workshops that can be held either online or face-to-face. Naturally, the interest and the capacity for such work will vary from country to country and over time. However, there are ample opportunities, at both the national and international levels, to work with various educational groups in disseminating a deeper understanding of how each domain has been operationalized—and the possible implications for changes in policies and practices to improve performance.

As the salience of ILSAs have grown, so have concerns in some quarters that, in many countries, they have become too influential in policy discussions and that they are a force for "homogenization" at the expense of national differences that are worth preserving (Carnoy 2015; Meyer and Benavot 2013). In part, the influence is due to the reputation of ILSA for generating data that are reliable and valid and, in part, to the arguments advanced by some that the best strategy is to emulate the policies of "high flyers." Indeed, in some instances, national education goals have been framed in terms of specific improvements in country rankings. This is both unfortunate and inappropriate, as there are many factors that determine a country's rank in a particular cycle, having little or nothing to do with the efficacy of its education system. In general, successful policy transfer across national boundaries can be difficult (Atkin and Black 1997; Braun 2008).

As Messick often quipped, "There is no such thing as a single-edged sword." The evident policy utility of ILSA should be complemented by explicit efforts to counter the overuse (or misinterpretations) of ILSA data. This requires sustained outreach at all levels but should take into account the judgments by the relevant authorities that adopting and adapting ILSA assessment frameworks and instruments to local needs is a positive step toward improving student learning. As Ritzen (2013) notes, utilization of ILSA results depends on a country's readiness (politically and otherwise) to institute changes, as well as having the capacity and commitment to conduct the change process. This may be particularly the case for many low- and middle-income countries whose participation has been mandated (and supported by) international donor organizations.

Conclusion

The transition to digitally based assessments marks an important inflection point in the evolution of ILSA to address policy makers' needs for actionable information in response to changes in the workplace and in society at large. We have argued that this transition has resulted in improvements in ILSAs' scope, efficiency, and data quality. These improvements, in turn, have enhanced the comparability, interpretability, and relevance of ILSA findings, resulting in greater policy utility. Although many challenges remain, we see a growing interest in these large-scale comparative assessments by policy makers and other key stakeholders, especially in response to the United Nations agenda for Sustainable Development Goals by 2030. We expect that the framework originally proposed by Messick and expanded here can be used both to guide future ILSA development and as a basis for making judgments about future utility. In that light, we expect this positive dynamic of increasing utility to continue, while acknowledging that both ILSA sponsors and national actors must remain vigilant in mitigating unintended negative consequences.

Acknowledgements
The authors would like to thank Larry Hanover of ETS for his editorial assistance in the final review of this paper. We would also like to thank the three internal ETS reviewers for their review and comments on the final draft of this paper.

Authors' contributions
This paper represents joint authorship between the two contributors. Each author provided important background information and made intellectual contributions to the development of the ideas and concepts presented in this paper. Each author also contributed to the writing and editing of this paper. Both authors read and approved the final manuscript.

Funding
No outside grant or contract monies from a third party were used in the development of this paper.

Availability of data and materials
Not applicable.

Competing interests
Not applicable.

Author details
[1] Educational Testing Service, Princeton, NJ, USA. [2] Boston College, Chestnut Hill, MA, USA.

Received: 8 January 2020 Accepted: 29 July 2020
Published online: 08 August 2020

References

Atkin, J. M., & Black, P. (1997). Policy perils of international assessments. *Phi Delta Kappan, 79,* 22–28.
Autor, D., Mindell, D. A., & Reynolds, E. B. (2019). *The work of the future: Shaping technology and institutions*. Cambridge: MIT Work of the Future.
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: Cham, Switzerland.
Bejar, I. I., & Braun, H. I. (1994). On the synergy between assessment and instruction: early lessons from computer-based simulations. *Machine Mediated Learning: An International Journal, 4*(1), 5–25.
Braun, H. I. (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. I. (2018). The relationships of family background to selected adult outcomes. *Large-scale Assessments in Education*. https://doi.org/10.1186/s40536-018-0058-x.
Braun, H. I., & Singer, J. D. (2019). Assessment for monitoring educational systems: International comparisons. *Annals of the Academy of Political and Social Sciences, 683,* 75–92. https://doi.org/10.1177/0002716219843804.
Braun, H. I., & von Davier, M. (2018). The use of test scores from large–scale assessment surveys: psychometric and statistical considerations. *Large-scale Assessments in Education*. https://doi.org/10.1186/s40536-017-0050-x.
Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress and prosperity in a time of brilliant technologies*. New York: W.W. Norton.
Bughin, J., Hazan, E., Lund, S., Dahlstrom, P., Wiesinger, A., & Subramaniam, A. (2018). Skill shift: Automation and the future of the workforce. New York: McKinsey and Company. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/future%20of%20organizations/skill%20shift%20automation%20and%20the%20future%20of%20the%20workforce/mgi-skill-shift-automation-and-future-of-the-workforce-may-2018.ashx
Carlson, J. E., & von Davier, M. (2017). Item response theory. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 133–178). Springer Open: Cham.
Carnoy, M. (2015). International test score comparisons and educational policy: A review of the critiques. https://nepc.colorado.edu/publication/international-test-scores (Accessed 3 December 2018).
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. https://www.epi.org/publication/bringing-it-back-home-why-state-comparisons-are-more-useful-than-international-comparisons-for-improving-u-s-education-policy/ (Accessed 3 December 2018).
Chen, H., Yamamoto, K., & von Davier, M. (2014). Controlling MST exposure rates in international large-scale assessments. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 391–409). Boca Raton: Chapman and Hall/CRC.
Dept, S. (in press). Translation in CB-ILSA: Not a stand-alone component. In L. Khorramdel, M. von Davier, & K. Yamamoto (Eds.), Computer-based international large-scale assessments: Concepts, methodologies and quality. Cham, Switzerland: Springer Open.
Ercikan, K., & Pellegrino, J. (Eds.). (2017). *Validation of score meaning for the next generation of assessments: The use of response processes. The NCME Applications of Educational Measurement and Assessment Book Series*. New York: Rutledge.
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
Fogg, N., Harrington, P., Khatiwada, I., Kirsch, I., Sands, A., & Hanover, L. (2019). If you can't be with the data you love: And the risks of loving the data you're with. Center for Research on Human Capital and Education. Princeton, NJ: Educational Testing Service. https://www.ets.org/research/report/love-the-data
Foshay, A. W., Thorndike, R. L., Hotyat, F., Pidgeon, D. A., & Walker, D. A. (1962). *Educational achievements of thirteen-year olds in twelve countries: Results of an international research project, 1959–1961*. Hamburg: UNESCO Institute for Education.

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

Goldhammer, F., Naumann, J., Rölke, H., Stelter, A., & Tóth, K. (2017). Relating product data to process data from computer-based competency assessment. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), Competence assessment in education (pp. 407–425). Cham: Springer. https://doi.org/10.1007/978-3319-50030-0_24

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. Computers & Education, 91, 92–105. https://www.journals.elsevier.com/computers-and-education

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). Boca Raton: CRC Press.

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

Kane, M. & Mislevy, R. (2017). Validating score interpretations based on response processes, 11–34. In K. Ercikan & J. Pellegrino, (Eds.) Validation of score meaning for the next generation of assessments: The use of response processes. The NCME Applications of Educational Measurement and Assessment Book Series. New York, NY: Rutledge.

Kirsch, I. S. (2001). The International Adult Literacy Survey (IALS): Understanding what was measured (Research Report No. RR-01–25). Princeton, NJ: Educational Testing Service.

Kirsch I., Braun, H. I., Yamamoto, K., & Sum, A. (2007). America's perfect storm: Three forces changing our nation's future. Princeton, NJ: Educational Testing Service. (2007). https://www.ets.org/perfect_storm

Kirsch, I., Braun, H., Lennon, M. L., & Sands, A. (2016). Choosing our future: A story of opportunity in America, ETS Center for Research in Human Capital and Education. Princeton, NJ: Educational Testing Service. https://www.ets.org/research/report/opportunity

Kirsch, I., & Lennon, M. L. (2017). PIAAC: A new design for a new era. Large-scale Assessments in Education, 5(11). Cham, Switzerland: Springer Open. https://doi.org/10.1186/s40536-017-0046-6

Kirsch, I., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). Adult literacy in America: A first look at the results from the National Adult Literacy Survey. Princeton, NJ: Educational Testing Service. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=93275

Kirsch, I., 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: Cham, Switzerland.

Kirsch, I., Thorn, W., & von Davier, M. (2018) (Eds). Special issue: Managing the quality of data collection in large-scale assessments. Quality Assurance in Education, 26(2). https://www.emerald.com/insight/publication/issn/0968-4883/vol/26/iss/2

Kirsch, I., Yamamoto, K., & Khorramdel, L. (2020). Design and key features of the PIAAC Survey of Adults. In D. B. Maehler & B. Rammstedt (Eds.), Large-scale cognitive assessment: Analysing PIAAC data. Cham: Switzerland: Springer International Publishing.

Krenzke, T., Mohadjer, L., Li, J., Erciulescu, A., Fay, R., Ren, W., Van de Kerckhove, W., Li, L. and Rao, J.N.K. (2020). PIAAC state and county indirect estimation methodology (NCES 2019-012). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Kuger, S., Kleime, E., Jude, N., & Kaplan, D. (Eds.). (2016). *Assessing contexts of learning: An International Perspective*. Cham: Springer International Publishing.

Lerner, D., & Laswell, H. D. (1951). *The policy sciences: Recent developments in scope and method*. Stanford: Stanford University Press.

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). Oxford: Symposium Books.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah: Lawrence Erlbaum Associates Inc.

Lubaway, E., Yamamoto, K., & Chen, C. (2019). Technical platforms for the improvement of coder reliability – OCES and OERS. Manuscript in preparation.

Maehler, D. B., Bibow, S., & Konradt, I. (2018). PIAAC bibliography—2008–2017. (GESIS Papers, 2018/03). Cologne, Germany (GESIS—Leibniz-Institut für Sozialwissenschaften. https://doi.org/10.21241/ssoar.56014

Mazeo, 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). Boca Raton: CRC Press.

Mazeo, J., Laser, S., & Zeiky, M. J. (2006). Monitoring educational progress with group-score assessments. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 681–699). Westport CT: Praeger Publishers.

Messick, S. J. (1987). Large-scale educational assessment as policy research: Aspirations and limitations. European Journal of Psychology of Education II, 2, 157–65. https://www.jstor.org/stable/23423437

Messick, S. J. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. J. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 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. (NEAP Report No. 83-1). Princeton, NJ: Educational Testing Service.

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). Oxford: Symposium Books.

Meyer, H.-D., & Schiller, K. (2013). Gauging the role of non-educational effects in large-scale assessments: Socio-economics, culture and PISA outcomes. In H.-D. Meyer & A. Benavot (Eds.), *PISA, Power and policy: The emergence of global educational governance* (pp. 207–224). Oxford: Symposium Books.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11,* 81–91. https://doi.org/10.1002/j.2330-8516.1986.tb00173.x.

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., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-centered assessment design*. Princeton: Educational Testing Service.
Mohadjer, L. and Edwards, B. (2018), Paradata and dashboards in PIAAC. In I. Kirsch, W. Thorn, & M. von Davier. (2018) (Eds). Special issue: Managing the quality of data collection in large-scale assessments. Quality Assurance in Education, 26(2). https://www.emerald.com/insight/publication/issn/0968-4883/vol/26/iss/2
Muro, M., Maxim, R., & Whiton, J. (2019). Automation and artificial intelligence: How machines are affecting people and places. Washington, DC: Brookings Institution. https://www.brookings.edu/research/automation-and-artificial-intelligence-how-machines-affect-people-and-places/
Organisation for Economic Co-operation and Development. (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD Survey of Adult Skills*. Paris: OECD Publishing. https://doi.org/10.1787/9789264128859-en.
Organisation for Economic Co-operation and Development (2016). Technical report of the Survey of Adult Skills (PIAAC), 2nd ed. Paris, France: OECD Publishing. 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. Paris, France: OECD Publishing. https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf
Qiwei, H., Dandan, L., & Hong J. (2019). Clustering behavioral patterns using process data in PIAAC problem-solving items. In B. P. Veldkamp & C. Sluijter (eds.), Theoretical and practical advances in computer-based educational measurement, methodology of educational measurement and assessment, https://doi.org/10.1007/978-3-030-18480-3_10
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). New York: Springer.
Rutkowski, D., & Delandshere, G. (2016). Causal inferences with large scale assessment data: using a validity framework. *Large-scale Assess Educ, 4,* 6. https://doi.org/10.1186/s40536-016-0019-1.
Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). (2014). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton: CRC Press.
Schleicher, A. (2018). *World class: How to build a 21st-century school system, strong performers and successful reformers in education*. Paris: OECD Publishing. https://doi.org/10.1787/9789264300002-en.
Secretary's Commission on Achieving Necessary Skills. (1991). *What work requires of schools: A SCANS report for America 2000*. Washington: U.S. Department of Labor.
Singer, J. D., & Braun, H. I. (2018). Testing international education assessments: rankings get headlines, but often mislead. *Science, 360*(6384), 38–40. https://doi.org/10.1126/science.aar4952.
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.
von Davier, M. (2014). Imputing Proficiency Data under Planned Missingness in Population Models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 175–201). Boca Raton: CRC Press.
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*. Boca Raton: CRC Press.
von Davier, M., Yamamoto, K., Shin, H., Chen, H., Khorramdel, L., Weeks, J., … & Kandathil, M. (2019). Evaluating item response theory linking and model fit for data from PISA 2000-2012. Assessment in Education: Principles, Policy & Practice. https://doi.org/10.1080/0969594X.2019.1586642
von Davier, M., Khorramdel, L., He, Q., Shin, H., & Chen, H. (in press). Developments in psychometric population models for technology based-large scale assessments—An overview of challenges and opportunities. Journal of Educational and Behavioral Statistics.
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). Boca Raton: CRC Press.
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.
World Economic Forum (2019). Strategies for the new economy: Skills as the currency of the labour market. Prepared by the Centre for the New Economy and Society within the World Economic Forum. Switzerland: http://www3.weforum.org/docs/WEF_2019_Strategies_for_the_New_Economy_Skills.pdf
Yamamoto, K., & Lennon, M.L. (2018), Understanding and detecting data fabrication in large-scale assessments. In I. Kirsch, W. Thorn, & M. von Davier. (2018) (Eds), Special issue: Managing the quality of data collection in large-scale assessments. Quality Assurance in Education, 26(2). https://www.emerald.com/insight/publication/issn/0968-4883/vol/26/iss/2
Yamamoto, K., Khorramdel, L., & Shin, H. J. (2018). Introducing multistage adaptive testing into international large-scale assessments designs using the example of PIAAC. Psychological Test and Assessment Modeling, 60, 347–368. https://www.researchgate.net/publication/328175530_Introducing_multistage_adaptive_testing_into_international_large-scale_assessments_designs_using_the_example_of_PIAAC
Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018b). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice, 37,* 16–27. https://doi.org/10.1111/emip.12226.

Publisher's Note