Chapter 1
On the Growing Importance of International Large-Scale Assessments

Irwin Kirsch, Marylou Lennon, Matthias von Davier, Eugenio Gonzalez and Kentaro Yamamoto

Large-scale assessments that compare the skills and knowledge demonstrated by populations across countries are relatively recent endeavors. These assessments have expanded in scope over time in response to increasing concern about the distribution of human capital and the growing recognition that skills contribute to the prosperity of nations and to better lives for individuals in those nations. Broadly defined, large-scale assessments are surveys of knowledge, skills, or behaviors in a given domain. The goal of large-scale assessments is to describe a population, or populations, of interest. As such, these assessments focus on group scores and can be distinguished from large-scale testing programs that focus on assessing individuals. The major themes laid out here—that these large-scale assessments have expanded over the past 50 years to include a greater number of surveys focusing on a broader range of populations and skill domains, that this work has led to new methodologies and modes of assessment, and that these assessments have grown to address the increasingly challenging questions posed by researchers and policymakers around the world—will be addressed in greater detail in each of the remaining chapters. We begin here by providing a general overview of the history of international large-scale assessments and the broadening role that these surveys have played in influencing policymakers around the world.

I. Kirsch (✉) · M. Lennon · E. Gonzalez · K. Yamamoto · M. von Davier
Educational Testing Service, Rosedale Road, MS 13-E, 08541 Princeton, NJ, USA
e-mail: ikirsch@ets.org

M. Lennon
e-mail: mlennon@ets.org

M. von Davier
e-mail: mvondavier@ets.org

E. Gonzalez
e-mail: egonzalez@ets.org

K. Yamamoto
e-mail: kyamamoto@ets.org

M. von Davier et al. (eds.), *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research,*
DOI 10.1007/978-94-007-4629-9_1,

Large-Scale Assessments of Student Populations

Prior to the late 1950s, no systematic or standardized comparative data focusing on skills and knowledge had been collected at national or international levels. The foundational work in this area began with a focus on student skills. In 1958, a group of scholars met at the UNESCO Institute for Education in Hamburg to discuss issues associated with collecting systematic data about schools and education systems in a cross-country context. That meeting led to a study designed to investigate the feasibility of developing and conducting an assessment of 13-year-olds in 12 countries. The pilot 12-country study focused on five domains including mathematics, reading comprehension, geography, science, and non-verbal ability and was conducted between 1959 and 1962. The results of this pioneering study demonstrated the feasibility of conducting a large-scale international survey in which common cognitive instruments worked in a comparable manner across different cultures and languages (Naemi et al. in press).

A parallel effort in the United States began around this same time under the leadership of several prominent American scholars and policymakers. Francis Keppel, the US Commissioner of Education in the mid-1960s, was responsible for reporting to Congress about the condition of education in America. Keppel was concerned about the lack of systematic data on the educational attainment of students in the country. As he pointed out, most of the information that had been collected to date focused on the inputs of education—such as the number of classrooms, dollars spent, and school enrollment figures—rather than on the output of education in terms of skills and knowledge. This concern led Keppel to invite Ralph Tyler, Director of the Center for Advanced Study in the Behavioral Sciences at Stanford University, to develop a plan for the periodic national assessment of student learning. With Tyler as chair, the Carnegie Foundation funded two planning meetings for national student assessments in 1963 and 1964. A technical advisory group was formed in 1965 and chaired by John Tukey, head of the Department of Statistics at Princeton University and Associate Executive Director of Research Information Systems at AT&T Bell Laboratories. This work led to the National Assessment of Educational Progress (NAEP), which conducted its first assessment of in-school 17-year-olds in citizenship, science, and writing in 1969.

Rather than build an assessment around classical test theory models that focused primarily on measuring individual differences, Tyler's vision for NAEP was to focus on what groups of students knew and could do. In this scheme, groups were defined by educationally relevant variables such as gender, immigrant status and ethnic background. Tyler's idea was to convene panels of subject-matter experts, to have them identify key educational objectives within the domains to be assessed, and then to develop test items based on those objectives. Reports from these assessments would then focus on the performance of national populations or subgroups rather than individual students. Additionally, Tyler was adamant that assessment results not be based on any type of norm-referenced perspective such as grade-level norms.

As surveys such as NAEP progressed, one of the criticisms that arose was that interpretations were quite limited because they were fixed to the individual items used in the assessments. In the 1980s, Educational Testing Service (ETS) bid on and won the contract to conduct NAEP based on a monograph written by Samuel Messick, Albert Beaton, and Frederic Lord. In "National Assessment of Educational Progress reconsidered: a new design for a new era," they introduced the idea of using Item Response Theory (IRT), an analytic approach with important advantages compared to the classical methods used previously in that it directly supports the creation of comparable scales across multiple forms of a test. In addition to incorporating IRT-based methodology, the work on NAEP led to developments of new methodologies including marginal estimation procedures that could optimize the reporting of proficiency scales based on very complex designs (von Davier et al. 2006).

NAEP and other surveys began by using a version of matrix sampling, an approach that is based on utilizing multiple, partially overlapping test forms. The introduction of balanced incomplete block (BIB) spiraling to large-scale assessment was another important innovation introduced in the 1980s. The goal of these developments was to broaden the item pool represented in the BIB-spiraled test forms in order to maximize the coverage of the constructs of interest. As an example, NAEP 8th grade mathematics assessments include a large number of test items across five subdomains of mathematics: number properties and operations; measurement; geometry; data analysis, statistics and probability; and algebra. Using BIB spiraling, each student is asked to respond to only a small subset of these items, reducing the burden on the test taker. Striking this balance of construct coverage and the reduction of test taker burden requires utilizing covariance information to create proficiency scales and the ability to generalize to populations of interest.

The use of IRT in combination with BIB-spiraling and covariance information among domains has made it possible to both broaden content coverage to include relevant facets of the cognitive constructs of interest and to extend inferences beyond individual items to the underlying construct. Just as we sample individuals and then make generalizations to populations, these scales, constructed with the help of IRT, represent a construct broadly and therefore make it possible to generalize beyond the specific items in the assessment to the construct domain that those items represent. These methodologies originally developed for NAEP are utilized in all the large-scale assessments covered in this volume, including the studies currently conducted by the International Association for the Evaluation of Educational Achievement (IEA) and the Organisation for Economic Co-operation and Development (OECD) that will be described next. Methodological innovations such as these have contributed to the growth and expansion of international large-scale assessments and allowed us to move beyond the questions raised by Tyler and others in the 1960s and 1970s and focus on increasingly complex questions raised by policymakers today.

Following the initial work that occurred from the 1960s through the 1980s, international large-scale assessments of student skills have expanded tremendously in terms of the number of assessments and participating countries. IEA continued to conduct important periodic large-scale international studies and, starting in 1995,

began to conduct continuous assessment cycles for the Trends in Mathematics and Science Study (TIMSS) followed by the Progress in Reading Literacy Study (PIRLS) in 2001. TIMSS is conducted every 4 years and focuses on achievement in mathematics and science at the fourth and eighth grades. PIRLS runs on a 5-year cycle and assesses how well children read after 4 years of primary school. By 2007, some 60 countries participated in TIMSS and over 40 countries participated in PIRLS. At the end of the 1990s, the OECD began the Programme for International Student Assessment (PISA) cycle of studies. PISA assesses the skills of 15-year-olds with the goal of gathering information about how well students have acquired the knowledge and skills essential for full participation in society. The first assessment was conducted in 2000 in over 30 countries and focused on the domains of reading, mathematics, and science. Since then, PISA has expanded in terms of the number of participating countries, with over 65 in the 2009 cycle, as well as the range of domains assessed, with cross-curricular areas such as problem solving and financial literacy being added to the assessment.

Large-Scale Assessments of Adults

In the 1990s, policy interest in the skills of adult workers and citizens led to the first international large-scale assessment focusing on adults ages 16–65. Working with Statistics Canada, ETS conducted the International Adult Literacy Survey (IALS) between 1994 and 1999, with 22 countries participating over three cycles. This assessment focused on prose, document, and quantitative literacy skills[1] and demonstrated the feasibility of conducting a household survey of adult literacy skills in an international context, maintaining comparability across countries and cultures. As such, IALS laid the foundation for subsequent surveys of adult skills and knowledge. The Adult Literacy and Life Skills Survey (ALL), which focused on a somewhat expanded set of adult skills including literacy, numeracy, and analytical problem solving, was conducted between 2003 and 2008 with some 11 countries participating.[2] The most recent adult survey, the OECD's Programme for the International Assessment of Adult Competencies (PIAAC), was conducting its first cycle in 2012 with 25 countries participating in 33 languages. PIAAC is a significant step forward in that it is the first computer-based household survey of adults, with interviewers taking laptops into people's homes and asking respondents to complete a background questionnaire and cognitive items on the computer. A parallel paper instrument is utilized for adults who are unable or unwilling to use the laptop equipment. For those adults taking the assessment on the computer, electronic reading tasks as well as scenario-based tasks assessing problem solving in technology envi-

[1] For definitions of these three literacy domains see Organisation for Economic Co-operation and Development and Statistics Canada (2000).

[2] For definitions of the ALL domains and more information about the survey see Statistics Canada and OECD (2005).

ronments complement the more traditional literacy and numeracy tasks that utilize texts, tables and static print-based stimulus material. PIAAC expands large-scale assessments by utilizing technology to administer the survey and, at the same time, embracing the fact that today's literacy-related tasks often take place in technology-based contexts such as web-based environments, spreadsheets and databases, or electronic mail.

The countries participating in today's student and adult large-scale surveys represent the overwhelming majority of GDP in the world and interest in the data these surveys yield continues to grow. For example, within the context of large-scale assessments, many countries now include special studies focusing on populations of particular interest such as the elderly, immigrants, and incarcerated adults. There is also interest in longitudinal studies as is the case in Canada, which is planning to use PIAAC to measure skills over time. Given that countries are looking more and more toward these assessments for data to drive and inform policy, it is likely that we will see international large-scale surveys continue to expand over time.

The Expanded Range of Large-Scale Assessments

As the aforementioned studies demonstrate, not only have we seen an expansion of who is assessed in terms of the range of participating countries and populations within those countries, but international large-scale assessments are also broadening the horizons in terms of what is being assessed. Earlier studies focused on in-school populations and measured typical academic domains such as mathematics, reading, and science. While these continue to be areas of interest, student assessments have expanded to measure a wider range of competencies and interests, reflecting a growing recognition of the need for lifelong learning as a tool to succeed in rapidly changing economies. Large-scale comparative surveys of adult populations began with a focus on literacy and quantitative skills and have expanded to include numeracy and problem solving in everyday adult contexts. With the growing importance of information technologies, measures of Information and Communication Technology (ICT) literacy skills, digital reading, and problem solving in technology environments have also been included in a number of studies.

The growing interest in assessing technology-related skills and knowledge has led to a growing interest in delivering assessments via computer. As has been mentioned, PIAAC is a household survey delivered on laptops. The call for tenders for PISA 2015 also focuses on moving that assessment more fully towards a computer-based platform. Computer-based assessments are making it possible to include new and innovative item types such as interactive scenario-based items and to collect a broader range of information including timing data and information about the processes test takers engage in when completing assessment tasks. This capability is, in turn, leading to a broadening of the cognitive constructs being measured. Additionally, computer-based assessments make it possible to take advantage of

psychometric advances such as the use of adaptive testing, which allows for more targeted and time efficient measures.

Another significant development in the history of international large-scale assessments has been the growing interest in broadening the information gained from cognitive measures through the use of extensive background questionnaires. Recent student and adult surveys typically include quite extensive background questionnaires. Student questionnaires address a range of topics including general attitudes and interests, day-to-day learning and leisure activities, and educational resources at home.

For adult assessments, questions about job requirements, literacy related activities at home and at work, and social outcomes such as engagement in civic activities have been included. Applying IRT scaling methodologies to these questionnaires has made it possible to create derived scales based on attitude and interest questions as well as on self-reported literacy-related activities and uses of technology. The use of IRT allows us to study differences across participating countries in terms of background characteristics along the same scales and in the same detail as is possible for the cognitive scales. Data from these questionnaires, in conjunction with the cognitive measures, are being used to inform increasingly complex policy questions about the relationships among learning, skills and outcomes.

Both the broadening of the cognitive constructs being addressed in large-scale comparative surveys and the interest in expanded coverage of policy relevant information collected in background questionnaires have driven the need to develop new methodologies for survey design and data analysis. What began as a basic desire to collect descriptive data in the 1960s and 1970s has now expanded to a much broader range of questions of policy interest. There is clearly growing interest on the part of stakeholders from different disciplines to address policy and research questions that are of interest both at the national and the international level.

Evidence-Based Policy Information

It is important to remember that the foundation of international large-scale assessments has always been some call for comparable information about the skills possessed by populations of interest and an understanding of how those skills are related to educational, economic and social outcomes. As such, the development of international large-scale assessments represents a cycle, as shown in Fig. 1.1. The initial work is motivated by policy questions which then drive the development of assessment frameworks and the design of instruments to address those questions. The desire to assess new aspects of existing constructs as well as to include new domains leads to advances in design and methodology that, finally, facilitate the analysis and interpretation of the survey data. This assessment data and the possibility of assessing new constructs as an outcome of more advanced methodologies leads, in turn, to new questions that then form the basis of the next cycle of assessment.

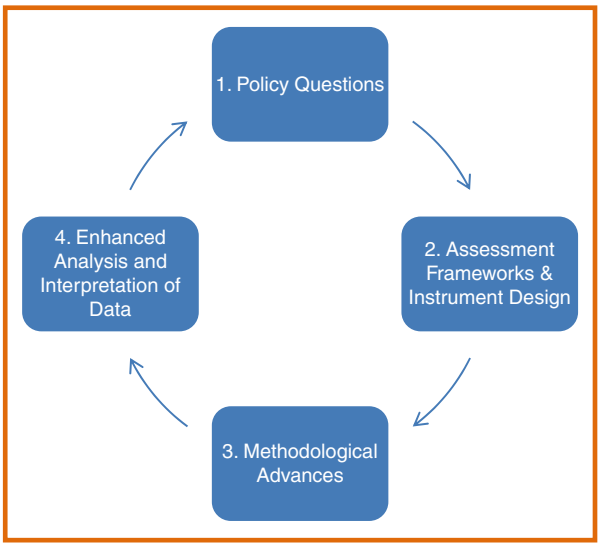

Fig. 1.1 The development cycle of evidence-based educational policy utilizing data from large-scale international assessments of skills

Current and future assessments will continue this cycle, enriching the databases available to researchers around the world to address questions arising in educational research and policy. And the most recent step towards computer-based assessments leads us to the next era of international large-scale assessments. While opening up new sources of information about how students and adults solve technology-based tasks, this move towards computer-based assessments also poses challenges. First and foremost, the ongoing development of new cycles of assessments and the move to new modes of technology-based delivery have to be reconciled with the desire of policymakers to measure trends—particularly for student populations—by comparing skills over cycles of these assessments. In addition, as technology develops, the platforms used for delivery of large-scale assessments will surely evolve and change more quickly than their paper-based predecessors. Finally, we can expect that new domains and new constructs may be added to the ones currently assessed, and additional information will be collected in the background instruments. These innovations will continue the development cycle, likely leading to new assessment methodologies, new interpretation models, and new areas of focus for large-scale assessments.

Perspectives on International Large-Scale Assessments

The contributions in the remaining chapters of this volume are based on invited presentations given during the International Large-Scale Assessment (ILSA) conference, held at Educational Testing Service in March of 2011. The range of perspectives reflected by the presenters and participants reflects the ever broadening

range of stakeholder groups with an interest in international large-scale assessments and the role that these assessments play in educational policy. The present volume includes the views of thought leaders from a variety of disciplines, all of whom have profound interest in international large-scale assessments.

In Chap. 2 "International Large-Scale Assessments as Change Agents," Jo Ritzen, Chair of the NGO Empower European Universities, focuses on ILSAs from the policy perspective. He discusses the impact that international large-scale assessments have had on educational policies, using PISA as a focus. He presents a series of national examples, illustrating how data from PISA 2006 resulted in a range of policy outcomes across participating countries. In addition, the role of transmission mechanisms, with a focus on the media in particular, in translating the results of assessments into institutional change is discussed. Ritzen concludes that international large-scale assessments provide transparency by allowing countries and institutions to evaluate the quality of their educational programs and compare where they stand compared to others. It is this transparency, he argues, that spurs change in policy and practice.

Michal Beller, Director-General of the Israeli National Authority for Educational Measurement and Evaluation (RAMA), focuses on the connection between technology-based assessments and learning. In Chap. 3, "Technologies in Large-Scale Assessments: New Directions, Challenges, and Opportunities," Beller contends that the goals for schools today are much broader than helping students master a set of core subjects. Educational systems in the twenty-first century must assist students in becoming critical thinkers, problem solvers, good communicators and good collaborators. In addition, schools strive to help students become information- and technology-literate, innovative and creative and globally competent. Beller argues ICTs have the potential to enhance the assessment of this broad range of skills. Her chapter addresses the computerized revolution of large-scale assessment and considers whether these new developments will be merely a technological leap forward, or serve as a catalyst for a more profound pedagogical change, influencing the way instruction and assessment will be conducted in the next era. Beller also discusses the role of computer-based large-scale assessments in the integration of twenty-first century skills into all content areas, as well as in creating new methodologies for a better use of technology in the service of learning.

Eric Hanushek of Stanford University brings his focus on economic analyses of educational issues, including efficiency, resource usage, and economic outcomes of schools, to the discussion of international large-scale assessments in his chapter. In Chap. 4, "The Role of Assessing Cognitive Skills in International Growth and Development," Hanushek notes that while most analyses of growth and development emphasize the central role of human capital, measurement issues have plagued both research and policy development. Specifically, attention to school attainment and enrollment rates appears misdirected. In contrast, Hanushek explains, recent work has shown that measures of cognitive skills derived from international assessments greatly improve the ability to explain differences in economic growth rates across countries. Moreover, higher levels of cognitive skills appear to have dramatic impacts on the future economic well-being of a country, suggesting that policy actions

should focus directly on school quality and other means of improving cognitive skills.

Henry Levin of Teachers College, Columbia University also looks at large-scale assessments in the context of the economics of education. However, Levin extends that focus beyond measures of traditional cognitive skills in Chap. 5, "The Utility and Need for Assessing Noncognitive Skills in Large-Scale Assessments." Levin notes that most attention in large-scale assessments of educational progress and outcomes is focused on cognitive measures of student proficiency. In part, this focus is due to the assumption that "skills" are cognitive in nature and have a high predictive value in terms of productivity. However, the predictive value of cognitive scores on worker productivity and earnings is more modest than commonly assumed. In fact, attempts to relate cognitive test scores from surveys to economic output, although meritorious, require substantial liberties in the interpretation of data. At the same time, there is considerable evidence that noncognitive skills are as important as—or even more important than—cognitive attributes in predicting both school outcomes and economic productivity. Noncognitive outcome measurement is more challenging to assess than cognitive because of its highly diverse dimensions and difficulties in sampling performance on these dimensions. This chapter addresses the developing knowledge base on the potential importance of noncognitive aspects of students and schools, issues of measurement and assessment, and their predictive value on adult outcomes.

In Chap. 6, "The Contributions of International Large-Scale Assessments in Civic Engagement and Citizenship," Judith Torney-Purta of the University of Maryland shares the perspective of researchers and policymakers interested in social policy. She focuses on three studies conducted since the 1970s that have addressed the topic of measuring civic engagement and citizenship and examined patterns of student achievement in attitudes and skills as well as knowledge. Because of the complexity of preparing for citizenship and workplace readiness in different democratic systems, these civic education projects have had an innovative edge in both assessment development and the analysis undertaken. Results from these studies have led to insights into political events, such as the difficulty of establishing civic education after a dictatorship, the rise of anti-immigrant parties, and changes in the political participation of young adults in Europe and the United States. These studies provide information about how students are able to get along with others in society, acquire norms, and participate via democratic means to implement change. In addition to considering civic studies in an international perspective, this chapter presents results of secondary analyses of CIVED data to illustrate the utility of these studies, and discusses analyses relevant for policy and for researchers in political science and psychology.

Eckhard Klieme, Director of the Center for Research on Educational Quality and Evaluation at the German Institute for International Educational Research, brings the perspective of researchers interested in school quality and school development to the discussion of international large-scale assessments. In Chap. 7, "The Role of Large-Scale Assessments in Research on Educational Effectiveness and School Development," Klieme contends that policymakers are primarily interested

in large-scale educational assessments as indicators that monitor the functioning, productivity and equity of educational systems, while researchers tend to perceive large-scale assessments as a kind of multi-group (i.e., multi-country) educational effectiveness study. Aside from describing strengths and challenges with regard to student performance and the conditions of teaching and schooling in participating countries, researchers also want to understand why students achieve certain levels of performance. Klieme argues that, because large-scale assessments provide only observational data, it is exceedingly difficult to draw causal inferences, such as concluding that a particular educational policy or practice has a direct or indirect impact on student performance. He proposes that a productive interplay between large-scale assessments and effectiveness research may be established in several ways by implementing enhancements to the assessment design. Two examples of such enhancements are presented and discussed: (1) a national large-scale assessment on language competencies in Germany reassessed students 1 year after the first large-scale assessment, allowing researchers to study the impact of school-level factors on classroom instruction and student growth and (2) a reassessment of Germany's schools performed 9 years after initial participation in PISA.

In Chap. 8 of this volume, "Prospects for the Future: A Framework and Discussion of Directions for the Next Generation of International Large-Scale Assessments," Henry Braun of Boston College describes a set of essential conditions that international large-scale assessments must satisfy in order to contribute to constructive change. These include that reported outcomes of these assessments must be credible, relevant, and correspond to national goals, that stakeholders must be spurred by the results to propose new policies and allocate (or reallocate) resources, and that policymakers must maintain a sustained but flexible focus on these policies. Braun situates the topics covered in this volume within this framework and concludes by examining how the positive impact of international large-scale assessments can be increased. In his view, forward-looking strategies require that these assessments provide more useful information, enhance the value of that information, and extend their reach through approaches such as developing strategies to allow a wider range of jurisdictions to participate and to share resources and expertise to build capacity.

As these brief summaries show, the authors in this volume cover a range of topics and perspectives related to the role and impact of international large-scale assessments. As such, they very much reflect the range of questions that international large-scale assessments will be asked to address as we strive to better understand where we are in terms of educational effectiveness and human capital development and how we might best move into an increasingly interconnected and challenging future.

References

Messick, S., A. Beaton, and F. Lord. 1983. *National assessment of educational progress reconsidered: A new design for a new era* (Report 83–10). Princeton: Educational Testing Service.

Naemi, B., E. Gonzalez, J. Bertling, A. Betancourt, J. Burrus, P. Kyllonen, J. Minsky, P. Lietz, E. Klieme, S. Vieluf, J. Lee, and R.D. Roberts, (in press). Large-scale group score assessments: Past, present, and future. In *Oxford handbook of psychological assessment of children and adolescents,* eds. Saklofske D., and Schwean V. (Cambridge, MA: Oxford University Press).

Organisation for Economic Co-operation and Development and Statistics Canada. 2000. *Literacy in the information age: Final report of the international adult literacy survey.* Paris: OECD Publishing.

Statistics Canada and Organisation for Economic Co-operation and Development (OECD) 2005. *Learning a living: First results of the adult literacy and life skills survey.* Paris: OECD Publishing.

von Davier, M. Sinharay, S., Oranje, A. and Beaton, A. 2006. Statistical Procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In *Handbook of statistics,* eds. Rao C. R., and Sinharay S., (Vol. 26): Psychometrics. Amsterdam: Elsevier.