

MEANINGFUL MEASUREMENT

The Role of Assessments in Improving High School Education in the Twenty-First Century

June 2009



© 2009 Alliance for Excellent Education. All rights reserved.

Suggested citation:

L. M. Pinkus, ed., *Meaningful Measurement: The Role of Assessments in Improving High School Education in the Twenty-First Century* (Washington, DC: Alliance for Excellent Education, 2009).

Ordering information:

Copies of *Meaningful Measurement: The Role of Assessments in Improving High School Education in the Twenty-First Century* can be downloaded from the Alliance's website at www.all4ed.org. To request print copies of the report, please visit http://www.all4ed.org/publication_material/order_form. The first copy of the report is complimentary. Additional copies are available at a charge of \$1 per copy to cover shipping and handling costs.

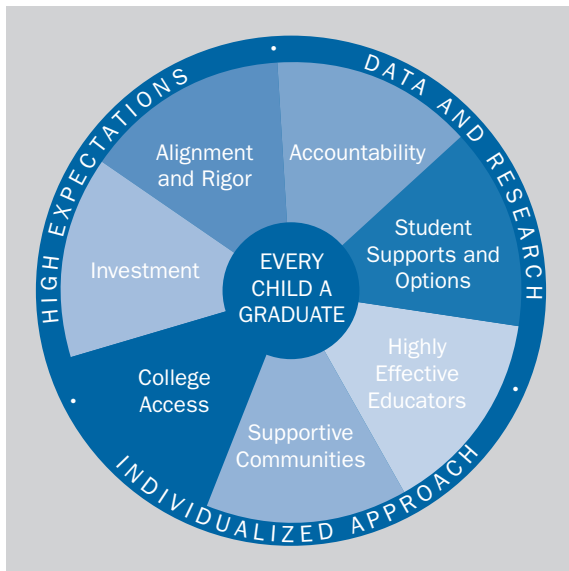
About the Alliance for Excellent Education

The mission of the Alliance for Excellent Education is to promote high school transformation to make it possible for every child to graduate prepared for postsecondary learning and success in life.

The Alliance for Excellent Education is a national policy and advocacy organization, based in Washington, DC, working to improve national and federal policy so that all students can achieve at high academic levels and graduate high school ready for college, careers, and citizenship in the twenty-first century.

The Alliance has developed a “Framework for Action to Improve Secondary Schools” that informs a set of federal policy recommendations based on the growing consensus of researchers, practitioners, and advocates about the challenges and solutions for improving secondary student learning.

The framework, shown graphically here, encompasses seven policy areas that represent key leverage points in ensuring a comprehensive, systematic approach to improving secondary education. The framework also captures



three guiding principles that apply to all of the policy areas. Although the appropriate federal role varies from one issue area to another, they are all critically important to reducing dropouts and increasing college and career readiness.

About the Editor

Lyndsay M. Pinkus is director of strategic initiatives at the Alliance for Excellent Education. Since joining the Alliance in January 2002, she has served in a variety of research, coordination, and advocacy roles, where her work has included managing policy and grant work on a range of issues including graduation rates, data, secondary school accountability, and secondary school improvement, and authoring a number of publications for the Alliance. Prior to rejoining the staff in January 2006, Ms. Pinkus served as a legislative associate at Washington Partners, LLC, providing government relations and policy research and analysis for a variety of clients, including the Alliance. She is a graduate of the School of Public Affairs at American University as a presidential scholar; the Public Affairs and Advocacy Institute at the Center for Congressional and Presidential Studies; and the Institute for Educational Leadership's Education Policy Fellowship program.

Acknowledgments

The Alliance for Excellent Education is greatly appreciative of the authors for sharing their time and expertise in writing the following chapters, as well as of the multiple Alliance staff members and advisors whose dedication contributed significantly to this volume.

The Alliance for Excellent Education is also grateful to Carnegie Corporation of New York for the financial support that made this publication possible.

The views expressed in this volume are those of the authors and do not necessarily represent those of the Alliance for Excellent Education or the funder.

CHAPTER

International Assessments of Student Learning Outcomes

Andreas Schleicher

Organisation for Economic Co-operation and Development

Introduction

Parents, students, and educators who teach and run education systems seek good information on how well their education systems prepare students for life. Most countries now monitor students' learning and the functioning of schools in order to provide answers to this question: among the thirty Organisation for Economic Co-operation and Development (OECD) countries and six other countries with comparable data, twenty-two countries undertake student examinations and/or assessments and seventeen require schools to be evaluated (either self-evaluations and/or inspections by an external body) at regular intervals. For student performance measures, student assessments (evaluations without direct consequences for the individual student) are used in seventeen countries, whereas national examinations (with direct consequences for the individual student) are used in ten OECD countries.

Comparative international assessments can extend and enrich the national picture by providing a larger context within which to interpret national

performance. They have gained prominence over recent years because the benchmarks for public policy in education are no longer solely national goals or standards, but increasingly the performance of the most successful education systems internationally.¹ International assessments can provide countries with information that allows them to identify areas of relative strengths and weaknesses and monitor the pace of progress of their education system. They can also stimulate countries to raise aspirations by showing what is possible in education in terms of the quality, equity, and efficiency of educational services provided elsewhere, and they can foster better understanding of how different education systems address similar problems.

Following a brief introduction to the history of international assessments, this chapter sets out the potential that international assessments offer for educational policy and practice as well as some of the methodological challenges they face in providing valid, comparable, and reliable evidence.

History of International Assessments

While efforts to compare education systems internationally can be traced back to the early nineteenth century,² the discourse on international comparisons of learning outcomes started to emerge during the 1950s and '60s. In 1958, an expert group led by William Douglas Wall and including prominent researchers such as Benjamin Bloom, Robert Thorndike, Arthur Wellesley Forshay, Arnold Anderson, Gaston Mialaret, and Torsten Husen met under the auspices of UNESCO's International Institute of Education in Hamburg, Germany, to launch a feasibility study to compare student performance internationally. The feasibility study involved twelve thousand thirteen-year-olds in twelve countries, and its results were published in 1962.³ The International Association for the Evaluation of Educational Achievement (IEA) emerged out of this collaboration, and later conducted a series of international assessments. The most prominent regular survey carried out by the IEA is the quarterly Trends in Mathematics and Science Study (TIMSS), which assesses fourth- and eighth-grade students' acquisition of math and science skills, and the Progress in Reading Literacy Study (PIRLS), which is given five times a year and measures reading literacy achievement of fourth-grade students.

The U.S. Education Testing Service conducted the International Assessment of Educational Progress (IAEP)⁴ in 1998 and a follow-up study in 1991.⁵ The latest generation of international assessments has been developed by the OECD as part of the Program for International Student Assessment (PISA). PISA surveys have been given every three years since 2000 in key content areas such as reading, mathematics, and science, but they also cover cross-curricular domains such as problem solving and a range of noncognitive outcomes. PISA is currently the most rigorous and also the most comprehensive international assessment, not least in terms of its coverage of subject areas and its geographic coverage, with the latest survey in 2009 testing more than 400,000 students in over seventy countries that together comprise close to 90 percent of the world economy. To implement the assessment, each country draws a random sample of between 3,500 and 50,000 fifteen-year-olds enrolled in school. Each participating student spends two hours carrying out pencil-and-paper tasks, solving electronically delivered problems, and answering multiple-choice questions. Students also answer a questionnaire focused on their personal background, their learning habits, and their engagement with and motivation at school. Principals complete a questionnaire about their school that includes demographic characteristics and an assessment of the quality of the school's learning environment.

Research Frameworks of International Assessments

The international assessments of the OECD and IEA seek to contextualize measures of student learning outcomes with background information collected from students, principals, and sometimes teachers and parents in order to interpret the observed variation in learning outcomes between students, classrooms, schools, and education systems.

To facilitate this, the tests operate with research frameworks that typically address **three research areas** (learning outcomes, policies shaping education outcomes, and factors that constrain policies and outcomes) with data at up to **four levels of the education system** (individual learners, classrooms or instructional settings, educational institutions and providers of educational services, and the education system as a whole). (See Table 1.)

These international assessments can then be used to address a variety of research issues from different perspectives relating, for example, to the quality of educational outcomes, to issues of equality of educational outcomes and equity in educational opportunities, or to the adequacy, effectiveness, and efficiency of resource management.

Table 1: Research Frameworks for International Assessments

		Research areas		
		Education and learning outputs and outcomes	Policy levers and contexts shaping educational outcomes	Constraints that contextualize policies and outcomes
Levels of education system	Individual participants in education and learning	The quality and distribution of individual educational outcomes	Individual attitudes, engagement, and behavior	Background characteristics of the individual learners
	Instructional settings	The quality of instructional delivery	Curriculum, pedagogy and learning practices, and classroom climate	Student learning conditions and teacher working conditions
	Providers of educational services	The output of educational institutions and institutional performance	School environment and organization	Characteristics of the service providers and their communities
	The education system as a whole	The overall performance of the education system	System-wide institutional settings, resource allocations, and policies	The national educational, social, economic, and demographic contexts

The Potential of International Assessments for Policy and Practice

The design and conduct of international assessments was originally motivated by research objectives. More recently, governments have begun to attribute growing importance to international assessments and have invested considerable resources into their development and implementation. This interest derives from several considerations:

- There is increasing recognition in many countries that the yardstick for educational success is no longer improvement by national standards but the performance of the best-performing education systems internationally. By **revealing what is possible in education** in terms of the performance levels demonstrated in the countries that perform strongest in international comparisons, international assessments can enhance the quality of existing policies but also create a debate about the paradigms and beliefs underlying policies. While international assessments alone cannot identify cause-and-effect relationships between inputs, processes, and educational outcomes, they can shed light on key features in which education systems show similarities and differences, and make those key features visible for educators, policymakers, and the general public. This, in turn, can generate powerful hypotheses for further analysis and research.
- In some countries, international assessments are also used to **set policy targets** in terms of measurable goals achieved by other systems, and seek to identify policy levers and establish trajectories as well as delivery chains for reform. In a number of countries, international assessments are also used to contextualize national standards and assessments.
- International assessments can assist with gauging the pace of educational progress, through assessing to what extent **achievement gains** observed nationally are in line with achievement gains observed elsewhere.
- Finally, international assessments can **support the politics** of educational reform, which is a major issue in education, where

any payoff to reform almost inevitably accrues to successive governments, if not generations.

These issues are examined more closely in the remainder of this section.

Revealing what is possible in education and identifying factors that contribute to educational success

International assessments seem to impact more on countries whose performance is comparatively low. Although it is sometimes argued that weighing the pig does not make it fatter, diagnosing underweight can be an important first step toward treatment. Also, as the level of public awareness was raised by international comparisons, it has in some countries created an important political momentum and engaged educational stakeholders, including teacher and/or employer organizations, in support of policy reform.⁶

Equally important, international assessments have had a significant impact in some countries that did not do poorly in absolute terms but found themselves confronted with results that differed from how educational performance was generally perceived in that country. (See, for example, the profile on Germany's experiences in the box on the opposite page.)

Showing that strong educational performance and improvement are possible seems to be one of the most important aspects of international assessments. Whether in Asia (like in Japan, Korea, or Singapore), in Europe (like in Finland or the Netherlands), or in North America (like in Canada), many countries display strong overall performance in PISA, and, equally important, some of these countries also show that poor performance in school does not automatically follow from a disadvantaged socioeconomic background. In addition, some countries show that success can become a consistent and predictable educational outcome. In Finland, for example, the country with the strongest overall results in PISA, the performance variation between schools amounted in 2006 to only 5 percent of students' overall performance variation on PISA. So parents can rely on high and consistent performance standards in whatever school they choose to enroll their children. Considerable research has been invested in the features of these education systems. In some countries, governments have used

knowledge provided by PISA as a starting point for a peer review to study policies and practices in countries operating under similar context that achieve better results.⁷ Such peer reviews, each resulting in a set of specific policy recommendations for educational improvement, are also being carried out by the OECD, the results of which have been published so far for Denmark and Scotland.⁸

Profile: Germany

In Germany, equity in learning opportunities across schools was historically often taken for granted, as significant efforts were devoted to ensuring that schools were adequately and equitably resourced. The results from the PISA 2000 assessment, however, revealed large socioeconomic disparities in educational outcomes between schools. Further analyses separated equity-related issues between those that relate to the socioeconomic heterogeneity within schools and those that relate to socioeconomic segregation through the school system. These results taken together suggested that German students from more privileged social backgrounds were being directed into the more prestigious academic schools, which yielded superior educational outcomes, while students from less privileged social backgrounds were being directed into less prestigious vocational schools, which yielded poorer educational outcomes, even where their performance on the PISA assessment was similar.

This raised the specter that the German education system was reinforcing rather than moderating socioeconomic background factors. Such conclusions, and the ensuing vivid public debate, inspired a wide range of equity-related reform efforts in Germany, some of which have been transformational in nature. These include

- giving early childhood education, which had hitherto been considered largely an aspect of social welfare, an educational orientation;
- establishing national educational standards for schools in a country in which regional and local autonomy had long been the overriding paradigm;
- introducing full-day schooling in a system where half-day schooling had been the norm for centuries; and
- enhancing the support for disadvantaged students, such as students with a migration background.

For many educators and experts in Germany, the socioeconomic disparities that PISA revealed were unsurprising. However, it had often been taken for granted and outside the scope of public policy that disadvantaged children would fare less well in school. The fact that PISA revealed that the impact that socioeconomic background has on students and school performance varied considerably across countries, and that other countries appeared to moderate socioeconomic disparities much more effectively, showed that improvement was possible—and provided the momentum for policy change.

As a result, the benchmarks for public policy in education are no longer national goals or standards alone, but increasingly the performance and achievement gains of the most successful education systems measured internationally. International assessments have at times raised awareness, leading to a public debate about education in which citizens have recognized that their country's educational performance will not just need to match average performance, but will have to do better if their children want to justify above-average wages.

Putting national targets into a broader perspective

International assessments can also play an important role in putting national performance targets into perspective. Educators are often faced with a dilemma: if, at the national level, the percentage of students achieving good exam scores in school increases, some will claim that the school system has improved. Others will claim that standards must have been lowered, and behind the suspicion that better results reflect lowered standards is often a belief that overall performance in education cannot be raised. International assessments allow those perceptions to be related to a wider reference framework by allowing schools and education systems to compare themselves with schools and education systems in other countries. Some countries have actively embraced this perspective and systematically related national performance to international assessments. Australia and Germany, for example, have embedded national items into the PISA assessments in order to relate what is considered important nationally to what is valued in other countries. Conversely, Japan has embedded PISA-type questions into its national assessment. By their very nature, international assessments assess aspects of students' skills and knowledge that are not *completely* covered by *all* national curricula, simply because curricula vary across countries. So they require national experts and authorities to examine what are the dimensions covered and uncovered in their schools, then to decide whether the uncovered ones should or should not be taught. When a country discovers that its students are unable to do things that students in other countries can do, the crucial question is, "Do *our* students need these skills too, to be able to survive in our modern society?" If the answer is yes, there is an opportunity to review and improve the standards, assessments, and curriculum.

Assessing the pace of change in educational improvement

A third important aspect is that international comparisons provide a frame of reference to assess the pace of change in educational development. While a national framework allows progress to be assessed in absolute terms, an internationally comparative perspective allows an assessment of whether that progress matches the pace of change observed elsewhere. Indeed, while all education systems in the OECD area have seen quantitative growth over past decades, international comparisons reveal that the pace of change in educational output has varied markedly.

For example, among fifty-five- to sixty-four-year-olds, the United States is well ahead of all other OECD countries in terms of the proportion of individuals with both school and university qualifications. However, international comparisons show that this lead is largely a result of the “first-mover advantage” that the United States gained after World War II by massively increasing school enrollments. This gain has eroded over the last few decades as more and more countries have reached and surpassed qualification levels in the United States in younger cohorts. While many countries are now close to ensuring that virtually all young adults leave schools with at least a high school qualification—which the OECD benchmarks highlight as the baseline qualification for reasonable earnings and employment prospects—the United States has stood still on this measure, and among OECD countries only New Zealand, Spain, Turkey, and Mexico now have lower secondary school completion rates than the United States.⁹

In contrast, two generations ago, South Korea had the economic output of Afghanistan today and was ranked twenty-fourth in terms of schooling performance among today’s OECD countries. Today it is the top performer in the proportion of successful school leavers, with 96 percent of an age cohort obtaining a high school degree. While progress from a national perspective matters, in this global framework the internationally comparative perspective is having a growing impact not just on public policy, but on institutional behavior as well. The results of international assessments of student performance are beginning to demonstrate similar influence.

A tool for changing the politics of education reform

International assessments can also affect the politics of education reform. For example, in the 2007 Mexican national survey of parents, 77 percent of parents interviewed reported that the quality of educational services provided by their children's school was good or very good. But in OECD's PISA 2006 assessment, roughly half of the Mexican fifteen-year-olds who were enrolled in school performed at or below the lowest level of proficiency established by PISA.¹⁰ There may be many reasons for this kind of discrepancy between perceived educational quality and performance on international assessments—it may be due in part, for instance, to the fact that the educational services that Mexican children receive are significantly better than what their parents experienced. However, the point here is that justifying the investment of public resources in areas for which there seems no public demand poses difficult challenges. One recent response by the Mexican presidential office was to include a “PISA performance target” in the new Mexican education reform plan. This performance target—based on the outcome of international assessments, and set to be achieved by 2012—will serve to highlight the gap between national performance and international standards, and monitor how educational improvement feeds into closing this gap. It is associated with a reform trajectory and delivery chain of support systems, incentive structures, and improved access to professional development to assist school leaders and teachers in meeting the target. Such reforms draw on the experience of other countries. Brazil has taken a similar route, providing each secondary school with information on the level of progress that is needed to perform at the OECD average performance level on PISA in 2021.

Japan is one of the best-performing education systems on the various international assessments. However, PISA results revealed that while Japanese students tended to do very well on tasks that require reproducing subject matter content, they did much less well on open-ended constructed tasks requiring them to demonstrate their capacity to extrapolate from what they know and apply their knowledge in novel settings. Conveying that situation to parents and a general public used to certain types of tests providing the gateway to further education poses a challenge for reform too. The policy response in Japan has been to incorporate “PISA-type” open-constructed tasks into the national assessment, with the aim that

skills that are considered important internationally will become valued in the national education system. Similarly, Korea has recently incorporated advanced PISA-type literacy tasks in its university entrance examinations, in order to enhance excellence in the capacity of its students to access, manage, integrate, and evaluate written material. In both countries, these changes represent transformational change that would have been much harder to imagine without the challenges revealed by PISA.

Design Issues and Challenges for International Assessments

The design of international assessments of learning outcomes needs to fulfill different and sometimes competing demands.

- International assessments need to ensure that their **outcomes are valid** across cultural, national, and linguistic boundaries, and that the target populations from which the samples in the participating countries are drawn are comparable.
- International assessments need to **offer added value** to what can be accomplished through national assessment and analysis.
- While international assessments need to be as comparable as possible, they also need to **be country specific** so they can adequately capture historical, systemic, and cultural variation among countries.
- The measures need to be as simple as possible to be widely understood, while remaining as complex as necessary to **reflect multifaceted educational realities**.
- While there is a general desire to keep any set of performance measures as small as possible, the picture should not be reduced to a small common denominator that no longer represents the variability of approaches and policy issues across countries, since this variability provides the foundation for countries to learn from each other's experiences.

Important issues that arise in meeting these demands are examined in the remainder of this section in more detail.

Cross-country validity and comparability in the assessment instruments

International assessments necessarily are limited in their scope for several reasons. First, there is no overarching agreement, across countries, on what students in a particular grade or at a particular age should know and be able to do—often referred to as “competencies.” Second, any single assessment can only measure a selection of such competencies. Lastly, there are various methodological constraints that limit the kinds of competencies that currently can be measured through large-scale assessment.

International assessments have made considerable progress toward assessing knowledge and skills in content areas such as mathematics, reading, science, and problem solving. However, they are still limited in the coverage of important cognitive outcomes, in particular the assessment of creative competencies. Similarly, achieving high degrees of objectivity in the assessments, which favor multiple-choice tasks that can be scored without human judgment, tends to detract from the assessment of the higher-order competencies and the production of knowledge, which require open-ended assessment tasks. At times, in order to make the assessments affordable to lower-income countries, international assessments have also sacrificed validity gains over efficiency gains, by giving undue weight to assessment tasks that can be easily administered and scored. Even less progress has been made to assess interpersonal dimensions of competencies that are often recognized as of increasing importance, such as the capacity of students to relate well to others or to manage and resolve conflicts. Last but not least, international assessments provide only very crude self-reported measures of intrapersonal dimensions of competencies.

Establishing the assessment domains

Even in established content areas, internationally comparative measurement poses major challenges. Countries vary widely in their intended, implemented, and achieved curricula. Inevitably, international assessments need to strike a balance between narrowing the focus to what is common across the different curricula of school systems, on the one hand, and capturing a wide enough range of competencies to reflect the content domains to be assessed adequately, on the other. Leaning toward the former—as has been the tendency for the assessments of the IEA—ensures

that what is being tested internationally reflects what is being taught in all countries. This is an important aspect of fairness, but there is a risk that the assessment reflects just the lowest common denominator of national curricula. It also lacks important aspects of curricula that are not taught in all of the countries, as well as the content validity that is required to faithfully represent the relevant subject area. Leaning toward the latter—as is the case for the assessments of the OECD, with their focus on the capacity of students not merely to reproduce what they have learned but to extrapolate from what they have learned and apply their knowledge and skills in novel settings—enhances content validity but risks that students are being confronted with assessment material they may not have been taught in their national context.

In whatever way the various international assessments have struck these balances, they have tried to build them through a carefully designed interactive process between the agencies developing the assessment instruments, various international expert groups working under the auspices of the respective organizations, and national experts charged with the development and implementation of the surveys in their countries. Often, a panel of international experts, in close consultation with participating countries, has led the identification of the range of knowledge and skills in the respective assessment domains that have been considered to be crucial for students' capacity to fully participate in and contribute to a successful modern society. A description of the assessment domains—the assessment framework—was then used by participating countries and other test development professionals as they contributed assessment materials.

For example, in the development of PISA, this involved

- the development of a working definition for the assessment area and a description of the assumptions that underlay that definition;
- an evaluation of how to organize the set of tasks constructed in order to report to policymakers and researchers on performance in each assessment area among fifteen-year-old students in participating countries;
- the identification of a set of key characteristics to be taken into account when assessment tasks were constructed for international use;

- the operationalization of the set of key characteristics to be used in test construction, with definitions based on existing literature and the experience of other large-scale assessments;
- the validation of the variables, and assessment of the contribution that each made to the understanding of task difficulty in participating countries; and
- the preparation of an interpretative scheme for the results.

The PISA assessment is defined through three interrelated dimensions: the knowledge or structure of knowledge that students need to acquire (e.g., familiarity with scientific concepts); competencies that students need to apply (e.g., carrying out a particular scientific process); and the contexts in which students encounter scientific problems and relevant knowledge and skills are applied (e.g., making decisions in relation to personal life, understanding world affairs). (See Table 2.)

Once the assessment framework is established and agreed upon (which tends to be the most challenging aspect of an international assessment), assessment items are developed to reflect the intentions of the frameworks, and they need to be carefully piloted before final assessment instruments can be established. To some extent, the question of to what extent the tasks in international assessments are comparable across countries can be answered empirically. Analyses to this end were first undertaken for the IEA Trends in Mathematics and Science Study.¹¹ The authors compared the percentage of correct answers in each country according to the international assessment as a whole with the percentage correct in each country on the items said by the country to address its curriculum in mathematics. Singapore, for example, had 144 out of 162 items that were said to be covered by the Singaporean curriculum. The percentage of items correct on the whole test and on the items covered in the curriculum was seventy-nine in both cases.

Singapore also scored between 79 and 81 percent correct on the items that other countries considered covered in their own curricula. These ranged from seventy-six items in Greece to 162 items in the United States. For most countries, the results were similarly consistent, suggesting that the composition of the tests had no major impact on the relative standing of countries in the international comparisons. Such analyses have also been conducted for PISA, and have yielded similar results.

Table 2: Defining an Assessment Domain—An Example from PISA

	Science
<p>Definition and its distinctive features</p>	<p>The extent to which an individual</p> <ul style="list-style-type: none"> • possesses scientific knowledge and uses that knowledge to identify questions, acquire new knowledge, explain scientific phenomena, and draw evidence-based conclusions about science-related issues; • understands the characteristic features of science as a form of human knowledge and inquiry; • shows awareness of how science and technology shape our material, intellectual, and cultural environments; and • engages in science-related issues and with the ideas of science, as a reflective citizen. <p>Scientific literacy requires an understanding of scientific concepts, as well as the ability to apply a scientific perspective and to think scientifically about evidence.</p>
<p>Knowledge domain</p>	<p>Knowledge <i>of</i> science, such as:</p> <ul style="list-style-type: none"> • “Physical systems” • “Living systems” • “Earth and space systems” • “Technology systems” <p>Knowledge <i>about</i> science, such as:</p> <ul style="list-style-type: none"> • “Scientific inquiry” • “Scientific explanations”
<p>Competencies involved</p>	<p>Type of scientific task or process:</p> <ul style="list-style-type: none"> • Identifying scientific issues • Explaining scientific phenomena • Using scientific evidence
<p>Context and situation</p>	<p>The area of application of science, focusing on uses in relation to personal, social, and global settings such as</p> <ul style="list-style-type: none"> • “Health” • “Natural resources” • “Environment” • “Hazard” • “Frontiers of science and technology”

Reflecting national, cultural, and linguistic variety

International assessments pay close attention to reflecting the national, cultural, and linguistic variety among participating countries. OECD's PISA assessments employ the most sophisticated and rigorous process to this end. The agency charged with the development of the instruments uses professional test item development teams in several different countries. In addition to the items developed by these teams, assessment material is contributed by participating countries and is carefully evaluated and matched against the framework. Furthermore, each item included in the assessment pool is rated by each country: (1) for potential cultural, gender, or other bias; (2) for relevance to the students to be assessed in school and nonschool contexts; and (3) for familiarity and level of interest.

Selecting assessment nature and form

Also important is the nature and form of the assessment, as reflected in the task and item types. While, as noted before, multiple-choice tasks are the most cost-effective way to assess knowledge and skills, and have therefore dominated earlier international assessments, they have important limitations in assessing more complex skills, particularly ones that require students not just to recall but to produce knowledge. Moreover, since the nature of assessment tasks, and in particular student familiarity with multiple-choice tasks, varies considerably across countries, heavy reliance on any single item type such as multiple-choice tasks can be an important source of response bias. The PISA assessments have tried to address this through employing a broad range of assessment tasks, with about 40 percent of the questions requiring students to construct their own responses. Other tasks require students to either provide a brief answer (short-response questions) or construct a longer response (open-constructed-response questions), allowing for the possibility of divergent individual responses and an assessment of students' justification of their viewpoints. Partial credit can be given for partly correct or less complex answers, with answers judged by trained specialists (or "coders") using detailed scoring guides. Open-ended assessment tasks, however, raise other challenges, in particular the need to ensure inter-rater reliability in the results. For PISA, there are a number of checks in place to ensure reliability. First, samples of the assessment booklets are coded independently by four coders and examined

by the international contractor. Second, an inter-coder reliability study and a homogeneity analysis are currently being implemented to examine the consistency of this coding process in more detail within each country, and to estimate the magnitude of variance associated with the use of coders. Third, an international coding review is now examining how consistently the response-coding standards are being applied across all participating countries, with the goal of estimating potential bias (either leniency or harshness) in the coding standards applied in participating countries. Lastly, in order to measure the intended broad range of content while meeting the limits of individual assessment time, PISA, like most modern international assessments, is now using multiple test forms within a country's test population.

Ensuring external validity

Ensuring that international assessments are comparable across countries is one thing, but the more important challenges relate to their external validity, which involves verifying that the assessments measure what they set out to measure. An important question is whether the knowledge and skills that are being assessed are predictive for the future success of students. In the case of PISA, the Canadian Youth in Transition Survey (YITS), a longitudinal survey that investigates patterns of and influences on major educational, training, and work transitions in young people's lives, provided a way to examine this empirically. In 2000, 29,330 fifteen-year-old students in Canada participated in both YITS and PISA. Four years later, the educational outcomes of the same students, then aged nineteen, were assessed, and the association of these outcomes with PISA reading performance at age fifteen was investigated.¹² The results showed that students who had mastered PISA performance Level 2 on the PISA reading test at age fifteen were twice as likely to participate in postsecondary education at age nineteen as those who performed at Level 1 or below, even after accounting for school engagement, gender, mother tongue, place of residence, parental education, and family income. The odds increased to eightfold for those students who had mastered PISA Level 4 and to sixteenfold for those who had mastered PISA Level 5. A similar study undertaken in Denmark led to similar results, in that the percentage of youth who had completed post-compulsory, general, or vocational upper-secondary education by the age of nineteen increased significantly with

their reading ability as assessed by PISA at age fifteen (see <http://www.sfi.dk/sw19649.asp>). Last but not least, the International Adult Literacy Study allowed reading and numeracy skills (defined in similar ways to those measured by PISA) to be related to earnings and employment outcomes in the adult population, and the analyses showed that such indicators were generally a better predictor for individual earnings and employment status than the level of formal qualification individuals had attained.¹³

Comparability of the target populations

Even if the assessment instruments are valid and reliable, meaningful comparisons can only be made if the target populations being assessed are also comparable. International assessments therefore need to use great care when defining comparable target populations, ensuring that they are exhaustively covered with minimal and well-defined population exclusions, and ensuring that the sampled students do participate in the assessment.

As regards defining target populations, important trade-offs need to be made between international comparability and relating the target populations to national institutional structures. Differences between countries in the nature and extent of pre-primary education and care, the age of entry to formal schooling, and the institutional structure of educational systems do not allow the establishment of internationally comparable grade levels. Consequently, international comparisons of educational performance typically define populations with reference to a target age group. International assessments of the IEA have defined these target groups on the basis of the grade level that provides maximum coverage of a particular age cohort (such as the grade in which most thirteen-year-olds are enrolled). The advantage of this is that a grade level can be easily interpreted within the national institutional structure and provides a cost-effective way toward assessment, with minimal disruption of the school day. However, a disadvantage is that slight variations in the age distribution of students across grade levels often lead to the selection of different target grades in different countries, or between education systems within countries, raising serious questions about the comparability of results across, and at times within, countries. In addition, because not all students of the desired age are usually represented in grade-based samples, there may be a more serious potential bias in the results if the unrepresented students

are typically enrolled in the next higher grade in some countries and the next lower grade in others. This excludes students with potentially higher levels of performance in the former countries and students with potentially lower levels of performance in the latter. To address these problems, the assessments of the OECD use an age-based definition for their target populations. For example, PISA assesses students who were between fifteen years and three (complete) months and sixteen years and two (complete) months at the beginning of the assessment period and who were enrolled in an educational institution, regardless of the grade level or type of institution in which they were enrolled and whether they were in full-time or part-time education. The disadvantages of this age-based approach is that it is costly, that the assessment process becomes more disruptive, and that it is more difficult to relate the results of individual students to teachers and classrooms.

The accuracy of any survey results also depends on the quality of the information on which national samples are based as well as on the sampling procedures. For the latest international assessments, advanced quality standards, procedures, instruments, and verification mechanisms have been developed that ensure that national samples yielded comparable data and that the results could be compared with confidence.

Even the best international samples will only translate into comparable results if the sampled schools are willing to take part in the assessment. While most countries participating in PISA now achieve high response rates, some countries, most notably the United States, have faced major challenges in securing school participation. At times, schools do not perceive sufficient benefit from an assessment that only yields national outcomes. Some countries have started to link PISA more closely to participating schools, either through providing them with school-level outcomes from the assessment or the related questionnaires, or through the provision of better information on the objectives and nature of these assessments. Incentives or feedback that have been deployed or are being considered by countries include

- better explanation of the context and usefulness of PISA at the start of the process to help engage teachers and schools;

- preparing a briefing pack to prepare teachers, schools, pupils, and parents to overcome pupils' initial anxieties and stimulate better communication within schools;
- setting up an international buddies scheme with schools doing PISA in other countries, in particular for sharing ideas for using the results to improve education;
- giving out student certificates on the day, perhaps as part of a small awards ceremony;
- encouraging PISA to be seen as a whole-school issue and to ensure corresponding dissemination;
- preparing electronic versions of feedback—perhaps in PowerPoint format to allow easier dissemination among staff;
- sharing good practice on what schools did with the feedback on the PISA website; and
- making the student questionnaire accessible so that schools can use it for benchmarking whenever they want and with a wider range of students.

Comparability in survey implementation

Well-designed international assessment needs to be well implemented to yield reliable results. The process begins with ensuring consistent quality and linguistic equivalence of the assessment instruments across countries. PISA, which provides the most advanced available procedures to this end, seeks to achieve this through providing countries with equivalent source versions of the assessment instruments in English and French and requiring countries (other than those assessing students in English and French) to prepare and consolidate two independent translations using both source versions. Precise translation and adaptation guidelines are supplied, including instructions for the selection and training of the translators. For each country, the translation and format of the assessment instruments (including test materials, marking guides, questionnaires, and manuals) is verified by expert translators appointed by the agency charged with the development of the assessment instruments (whose mother tongue was the language of instruction in the country concerned and who were knowledgeable about education systems) before they are used.

The assessments then need to be implemented through standardized procedures. Comprehensive manuals typically explain the implementation of the survey, including precise instructions for the work of school coordinators and scripts for test administrators for use during the assessment sessions. Proposed adaptations to survey procedures, or proposed modifications to the assessment session script, are reviewed internationally before they are employed at a national level. In the case of PISA, specially designated quality monitors visited all national centers to review data-collection procedures and school quality. Monitors from the international agency visited a sample of fifteen schools during the assessment. Marking procedures are designed to ensure consistent and accurate application of the internationally agreed-upon marking guides.

Recommendations and Conclusion

In a globalized world, the benchmarks for public policy in education are no longer national goals or standards alone, but increasingly the performance of the most successful education systems internationally. International assessments can be powerful instruments for educational research, policy, and practice by allowing education systems to look at themselves in the light of intended, implemented, and achieved policies elsewhere. They can show what is possible in education in terms of quality, equity, and efficiency in educational services, and they can foster better understanding of how different education systems address similar problems. Most importantly, by providing an opportunity for policymakers and practitioners to look beyond the experiences evident in their own systems and thus to reflect on some of the paradigms and beliefs underlying these, they hold out the promise to facilitate educational improvement. As this chapter has shown, designing and implementing valid and reliable international assessments poses major challenges, including defining the criteria for success in ways that are comparable across countries while remaining meaningful at national levels, establishing comparable target populations, and carrying the surveys out under strictly standardized conditions. However, more recently, international assessments such as PISA have made significant strides toward this end.

Some contend that international benchmarking encourages an undesirable process of degrading cultural and educational diversity among institutions and education systems, but the opposite can be argued as well: in the dark,

all institutions and education systems look the same, and it is comparative benchmarking that can shed light on differences on which reform efforts can capitalize. Who took notice of how Finland, Canada, or Japan ran their education systems before PISA revealed their success in terms of the quality, equity, and coherence of learning outcomes?

Of course, international assessments have their pitfalls, too. Policymakers tend to use them selectively, often in support of existing policies rather than as instruments to challenge them and to explore alternatives. Moreover, highlighting specific features of educational performance may detract attention from other features that are equally important, thus potentially influencing individual, institutional, or systemic behavior in ineffective or even undesirable ways. This can be like the drunk driver who looks for his car key under a street lantern and, when questioned whether he lost it there, responds that he didn't—but that it was the only place where he could see. This risk of undesirable consequences of inadequately defined performance benchmarks is very real, as teachers and policymakers are led to focus their work on the issues that performance benchmarks value and put into the spotlight of the public debate.

While the development of international assessments is fraught with difficulties and their comparability remains open to challenges, cultural differences among individuals, institutions, and systems should not suffice as a justification to reject their use, given that the success of individuals and nations increasingly depends on their global competitiveness. The world is indifferent to tradition and past reputations, unforgiving of frailty, and ignorant of custom or practice. Success will go to those individuals, institutions, and countries that are swift to adapt, slow to complain, and open to change. The task for governments will be to ensure that their citizens, institutions, and education systems rise to this challenge, and international benchmarks can provide useful instruments to this end.

There are specific actions U.S. leaders can take to use international assessments as a tool in evaluating progress and establishing effective policies toward the goal of graduating every student from high school prepared for college and the demands of the twenty-first-century global economy. These include

- measuring state-level education performance globally by examining student achievement and attainment in an international context to ensure that, over time, students are receiving the education they need to compete in the twenty-first-century economy (this can be achieved through participating at both the national and state levels in international studies like PISA that serve to collect data about student performance as well as related policies and practices);
- creating an ongoing public awareness and interest in the importance of international education comparisons by communicating results widely, encouraging discussion about findings, and partnering with key stakeholders;
- embedding international indicators into policy goals and decisionmaking processes; and
- employing the PISA framework as a tool in evaluating and improving U.S. standards and assessments.

The views expressed in this chapter are those of the author and do not necessarily represent those of the Alliance for Excellent Education.

About the Author

Andreas Schleicher is head of the Indicators and Analysis Division (Directorate for Education) at the Organisation for Economic Co-operation and Development (OECD). He also holds an honorary professorship at the University of Heidelberg in Germany. As division head at the OECD, his responsibilities include directing the Program for International Student Assessment (PISA) and the Indicators of Education Systems program (INES) and steering the development of new projects such as the OECD Teaching and Learning International Survey (TALIS) and the OECD Programme for the International Assessment of Adult Competencies (PIAAC). At the OECD, Mr. Schleicher has also held the posts of deputy head of the Statistics and Indicators Division in the former Directorate for Education, Employment, Labour and Social Affairs (1997–2002) and project manager in the OECD Centre for Educational Research and Innovation (CERI) (1994–1996).

Before joining the OECD, he served as director for analysis at the International Association for Educational Achievement (IEA) within the Institute for Educational Research in the Netherlands (1993–1994) and international coordinator for the IEA Reading Literacy Study, at the University of Hamburg, Germany (1989–1992).

In 2003, Mr. Schleicher was awarded the Theodor Heuss Prize, named after the first president of the Federal Republic of Germany, for “exemplary democratic engagement” in association with the public debate on PISA. In 2002, he was awarded the *educación y libertad en el ámbito educativo* prize by the Spanish national association of private schools. Mr. Schleicher earned a bachelor’s degree in physics and a master’s degree in mathematics from Deakin University in Australia, where his master’s thesis received the Bruce Choppin Award.

¹ D. Hopkins, D. Pennock, and J. Ritzen, *Evaluation of the Policy Impact of PISA* (Paris: OECD, 2008).

² M. A. Jullien, *Esquisse et vues préliminaires d'un ouvrage sur l'éducation compare* (Paris: L. Colas, 1817).

³ A. W. Foshay, R. L. Thorndike, F. Hotyat, D. A. Pidgeon, and D. A. Walker, *Educational Achievement of Thirteen-Year-Olds in Twelve Countries* (Hamburg: UNESCO Institute for Education, 1962).

⁴ A. E. LaPointe, N. A. Mead, and G. W. Phillips, *A World of Differences: An International Assessment of Mathematics and Science* (Princeton, NJ: Educational Testing Service, 1989).

⁵ A. E. LaPointe, N. A. Mead, and J. M. Askew, *The International Assessment of Educational Progress Report* (Princeton, NJ: Educational Testing Service, 1992).

⁶ Hopkins, Pennock, and Ritzen, *Evaluation of the Policy Impact of PISA*.

⁷ H. Döbert, E. Klieme, and W. Sroka, *Vertiefender Vergleich der Schulsysteme ausgewählter PISA-Teilnehmerstaaten* (Frankfurt: Deutsches Institut für pädagogische Forschung, 2004).

⁸ OECD, *Reviews of National Policies for Education—Denmark: Lessons from PISA 2000* (Paris: OECD, 2004); OECD, *Reviews of National Policies for Education: Quality and Equity of Schooling in Scotland* (Paris: OECD, 2007).

⁹ OECD, *Education at a Glance—OECD Indicators 2007* (Paris: OECD, 2008).

¹⁰ IFIE-ALDUCIN, *Mexican National Survey to Parents Regarding the Quality of Basic Education* (Mexico City: IFIE-ALDUCIN, 2007); OECD, *Reviews of National Policies for Education: Quality and Equity of Schooling in Scotland*.

¹¹ A. E. Beaton, I. V. S. Mullis, M. O. Martin, E. J. Gonzales, D. L. Kelly, and T. A. Smith, *Mathematics Achievement in the Middle School Years* (Chestnut Hill, MA: Boston College Center for the Study of Testing, Evaluation, and Educational Policy, 1996).

¹² T. Knighton and P. Bussiere, *Educational Outcomes at Age 19 Associated with Reading Ability at Age 15* (Ottawa: Statistics Canada, 2006).

¹³ OECD and Statistics Canada, *Literacy Skills for the Information Age* (Ottawa and Paris: Authors, 2000).