17

International Assessments

Sauli Takala
University of Jyväskylä, Finland

Gudrun Erickson
University of Gothenburg, Sweden

Neus Figueras
University of Barcelona, Spain

Introduction

Assessment and evaluation are pervasive features of human activity: We evaluate everything and are being evaluated all the time. Education is no exception. While education generally aspires to goals of individual growth and development, it is also expected to serve social, cultural, and economic policies. One of the present top policy priorities is to enable the nations and their citizens to take full advantage of an increasingly globalized economy. This requires provision of high quality and sustainable education, with an acceptable degree of equity in the distribution of opportunities to learn (OTL) and with clear incentives for achieving greater efficiency in schooling.

Successful educational policy and well-informed planning and implementation depend on indicators showing how well the educational systems are functioning. During recent decades, many countries have set up monitoring systems of various kinds: revised national examinations or sample-based national assessment to monitor students' learning and the performance of schools (e.g., National Assessment of Educational Progress [NAEP], designed in the late 1960s). In addition to national assessments, international yardsticks were called for. Systematic international assessments emerged in 1958 when the International Association for the Evaluation of Educational Achievement (IEA) was set up, and expanded when the Organization for Economic Cooperation and Development (OECD) launched the intergovernmental Programme for International Student Assessment (PISA) project. International assessments have since proliferated. As indicated above, international assessment is understood here to refer to assessments undertaken by an international team or organization to obtain comparative

The Companion to Language Assessment, First Edition. Edited by Antony John Kunnan.

DOI: 10.1002/9781118411360.wbcla052

information on educational performance through a jointly planned approach and methodology. This means that, for example, widely used international tests are not covered in this chapter.

Previous Views or Conceptualization

Descriptive Phase in International Comparisons

Throughout the long history of formal education and long before the emergence of the IEA and PISA international assessments, the quality of education had been of interest and an object of comparison to students, parents, and scholars. As a consequence, many students chose to study abroad in well-reputed international educational institutions. When national educational systems were being developed, it was common for educationalists to visit other countries to observe how education was conducted elsewhere and what appeared to be the outcomes. Such visits to "educational laboratories" provided useful stimuli, although data were not gathered in a consistent and standardized fashion.

This comparative approach was often ethnographic (in a broad sense), setting the descriptive national case studies in a cultural context, paying particular attention to the curricular arrangements (what was being taught), the organization of the educational system, teacher education, and teaching methods. Successful pedagogic approaches were copied and adapted (Pestalozzi, Herbart, Montessori, Waldorf, and so forth). Occasionally a more explicit exploration followed, when it was perceived that some particular country was doing particularly well in a subject. For instance, Brown (in 1915) reported to his interested American readers "how the French boy learns to write."

Comparative education developed also as a discipline (e.g., Noah, 1973) and acquired special journals, the flagship of which, *Comparative Education Review*, started in 1957.

From early on, examinations had been a burning pedagogical problem in many countries. At a world congress in 1927, a committee was set up to study the question. This committee met five times from 1931 to 1938. At the final conference in 1938, members from the participating countries, namely England, Finland, France, Germany Norway, Scotland, Sweden, Switzerland, and the United States, presented reports confirming problems concerning the marking of essays, highlighting the common inadequacies of the prevailing examinations in all countries, and stressing the need for intensive research to improve such measures (see Spolsky, 1995, pp. 66–73 for a succinct review). In spite of such activity, empirical comparative education was in short supply.

Emergence of a Systematic Approach: IEA

In the late 1950s, a group of internationally minded scholars initiated discussions within the IEA on the idea that doing systematic empirical research on educational achievement in a comparative perspective and using the same data collection methods and instruments might provide useful theoretical and practical

information on patterns of variables related to the levels of achievement across countries. The variation in educational systems was seen to provide a "natural laboratory," a natural "experimental setting."

The IEA studies, the main focus of this section, measure performance among students of different countries and thereby indirectly highlight the question of whether certain policies in a particular educational system have a positive or negative impact on learning.

> Through its comparative research and assessment projects, IEA aims to:
>
> 1. provide international benchmarks to assist policy-makers in identifying the relative strength and weaknesses of their education systems
> 2. provide high-quality data to increase policy-makers' understanding of key school- and non-school-based factors that influence teaching and learning
> 3. provide high-quality data that will serve as a resource for identifying areas of concern and action, and for preparing and evaluating educational reforms
> 4. develop and improve the capacity of education systems to engage in national strategies for educational monitoring and improvement
> 5. contribute to the development of a worldwide community of researchers in educational evaluation. (IEA, *n.d.*)

The early IEA international assessments reflected the influential views of Ralph W. Tyler, and the Chicago measurement school more generally, on the triangular relationship between goals of education (curriculum), modes of instruction, and the assessment of outcomes. In the assessments conducted in the 1980s, a distinction between the intended curriculum, the implemented curriculum, and the realized curriculum (systemic, instructional, and student levels, respectively) became an important design feature.

Since 1958, IEA has conducted more than twenty comparative surveys focusing on student performance (see Papanastasiou, Plomp, & Papanastasiou, 2011). The main purpose of the massive Six Subject Survey (Walker, 1976), including a quarter of a million students in about 10,000 schools and stretching from the late 1960s to the mid-1970s, was to study the relationship between *input* factors in the social, economic, and instructional domains and *output* as measured by international tests covering both cognitive (student performance) and affective behavior (questionnaires on student attitudes and motivation). These relationships were studied in some twenty national systems of education and, as a rule, at three different levels (populations) within each educational system, aiming at generalizable findings.

The IEA studies used a common design (see Table 17.1) where achievement (dependent variable) was predicted by a variety of societal, institutional, instructional, and personal characteristics, using multivariate methods such as regression analysis and path analysis. The independent variables were arranged in "blocks" with the home background entered as the first block in analyses, followed by type of school or program (degree of selectivity) and school instruction variables. This order was considered to reflect the causal sequence in influencing school achievement (see also Figure 17.1). Walker (1976) provides an informative summary of the six studies.

Table 17.1 A summary of early IEA language-related surveys, late 1960s to early 1990s

Study	*Target populations*	*Participants*	*Tests*	*Questionnaires*	*Reports*
Reading Comprehension (1968–72)	* 14-year-olds * Final grade of secondary school	Belgium, Chile, England, Finland, Hungary, India, Italy, Netherlands, New Zealand, Scotland, Sweden, United States	* Verbal ability * Reading comprehension * Speed of reading * Word knowledge	* Out-of-school literacy environment * Educational practices * Size and type of school * Interests, attitudes * Study and reading habits	Thorndike (1973)
Literature Education (1968–73)	* 14-year-olds * Final grade of secondary school	Belgium, Chile, England, Finland, Hungary, Italy, New Zealand, Sweden, United States	* Measures of literary response * Literary comprehension	* Attitudes to literature * Interest in literature	Purves (1973)
French as a Foreign Language (1968–73)	* 14-year-olds * Final grade of secondary school	Chile, England, Netherlands, New Zealand, Romania, Scotland, Sweden, United States	* Reading * Listening * Speaking * Two writing tests ("objective" and directed composition)	* Student * Teacher * School	Carroll (1975)
English as a Foreign Language (1968–73)	* 14-year-olds * Final grade of secondary school	Belgium, Chile, Finland, Germany (FRG), Hungary, Israel, Netherlands, Sweden, Thailand	* Reading (six subtests, including vocabulary and grammar) * Listening (sound discrimination, sentence comprehension, dictation)	* Place of English in the educational system * Student * Teacher * School * Context	Lewis and Massad (1975)
Written Composition (1983–9)	Students: * near the end of primary schooling (A) * near the end of compulsory schooling (B) * near the end of academic secondary school (C)	Chile, England, Finland, Germany (FRG), Hungary, Indonesia, Italy, Netherlands, New Zealand, Nigeria, Sweden, Thailand, United States, Wales	See Figures 17.1 and 17.2	See Figures 17.1 and 17.2	Gorman, Purves, and Degenhart (1988); Purves (1992)

Language Education (1993–6)	* End of compulsory schooling (ages 15/16) * End of upper secondary schooling (ages 17/ 18) Planned as a three-phase project ending up with testing of performance, but lack of funding limited information gathering to phase 1	Austria, Cyprus, Czech Republic, Denmark, England, Finland, France, Hong Kong, Hungary, Iran, Israel, Italy, Latvia, Netherlands, Norway, Philippines, Portugal, Russian Federation, Slovenia, South Africa, Spain, Sweden, Switzerland, Thailand, United States	* Data for four languages commonly taught as a school subject (English, French, German, and Spanish) collected in 1995	* Language education (sociolinguistic context, language policy, curriculum and assessment) * Language teaching and professional support) * School level (characteristics of schools and teachers) * Student level (proficiency, attitudes, and aspirations)	Dickson and Cumming (1996)
Reading Literacy (1985–94)	* 9-year-olds * 14-year-olds	Belgium (French), Botswana, Canada (British Columbia), Cyprus, Denmark, Finland, France, Germany (FRG), Germany (GDR), Greece, Hong Kong, Hungary, Iceland, Indonesia, Ireland, Italy, Netherlands, New Zealand, Nigeria, Norway, Philippines, Portugal, Singapore, Slovenia, Spain, Sweden, Switzerland, Thailand, Trinidad & Tobago, United States, Venezuela, Zimbabwe	* Word recognition (only 9-year-olds) * Narrative * Expository * Documents	* Student * Teacher * School * National case study	Elley (1994)

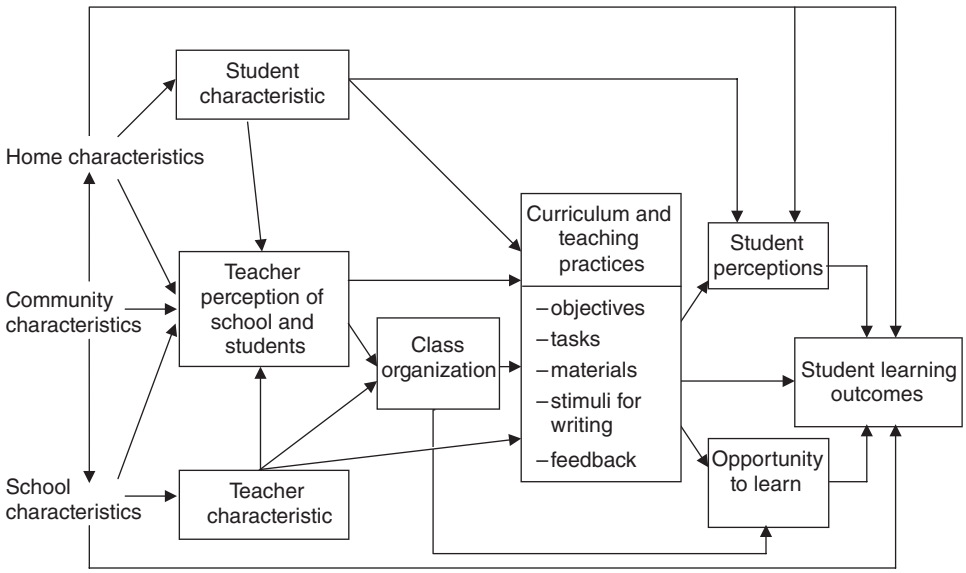

Figure 17.1 **The design of the IEA Study of Written Composition (adapted from Gorman, Purves, & Degenhart, 1988, p. 10) © Elsevier**

Figure 17.1, based on a design used in the Study of Written Composition, illustrates the approach to the IEA study designs. This kind of model is still basically applied in broad outline. For an up-to-date conceptualization in the Progress in International Literacy Study (PIRLS), consult http://timss.bc.edu/pirls2011/downloads/PIRLS2011_Framework.pdf

In addition to the prioritized international studies of mathematics and sciences, the IEA carried out studies of English and French as a foreign language and of reading and literature, published in the early 1970s, and of writing in the late 1980s. Studies of reading have continued, focusing on 10–11-year-olds (PIRLS) with a cycle of five years (2001, 2006, and 2011).

The language-related IEA studies are presented in Table 17.1.

The wealth of results cannot be reported in any detail (see Walker, 1976). Therefore, only two studies are discussed briefly below: the study of French (second language [L2]) and the study of written composition (first language [L1]) as summarized on the IEA website (http://www.iea.nl/completed_studies.html). As a prominent psychometric expert, Carroll (1975) was able to apply state-of-the-art methodology and, incidentally, also explore the validity of his 1973 model of school learning. The main findings were these:

- General proficiency in learning French was strongly related to performance on a word knowledge test in the student's mother tongue, which was used as a measure of verbal ability.
- The student's aspiration to understand spoken French contributed more to listening achievement than to reading achievement. Aspiration to learn to read French contributed more to reading scores than to listening scores.

- In all four fields of performance (reading, listening, speaking, and writing) there was a strong linear relationship between country mean score and the average number of years the students had studied French.
- Time spent on homework had an influence on reading scores, but much less effect on listening scores, which were only indirectly influenced by amount of homework. Classroom activities were much more important for listening. Students achieved higher scores when French was used for a substantial part of the time in the classroom, and when the use of the mother tongue was reduced but not eliminated.
- Neither the amount of university training nor the amount of travel or residence in a French-speaking country by the teacher led to any differences in students' French achievement.

Carroll found that the French study was very successful in identifying predictors of achievement in French. As an innovation in methodology, he pooled the data across countries and used canonical regression analyses to explore the "international French classroom." He estimated, among other things, that 5–6 years with three or four weekly lessons were required to achieve a satisfactory level of reading comprehension (Carroll, 1975, pp. 227–64).

The domain specification and the sampling of tasks for the three populations (A, B, and C) of the Study of Written Composition are presented in Table 17.2.

The key findings of the study of written composition, again as summarized on the IEA website, were as follows:

- The construct "written composition" was found to be sited in a cultural context and so cannot be considered a general cognitive capacity or activity. Marked variation across the countries existed both in the ideology of the teachers and in instructional practices. Written performance was also found to be task dependent.
- Good compositions from different countries shared common qualities of handling of content and appropriateness of style, but these qualities had their national or local characteristics in organization, use of detail, and other aspects of rhetoric.
- Students across educational systems had in common a sense of the importance of the written product and its surface features. Beneath that commonality, however, there was national variation in the perception of what is valued.
- In most countries, girls were treated differently than boys in the provision of writing instruction and in the rating of writing performance, particularly at the primary and lower secondary school levels, where women largely provided instruction. In such a milieu, the most successful students were girls, and gender itself, or gender in combination with certain home variables, was the most powerful predictor of successful performance, particularly on the more "academic" tasks.
- Differences between the ratings of student writing were not explained by differences in instruction. They were, however, accounted for by factors involving the characteristics of the home, the reinforcement provided by parents, and the cultural values of the community.

Table 17.2 Domain specification and distribution of tasks among the three populations in the IEA Study of Written Composition

Dominant intention/ Purpose	*Primary cognitive demand*		
	Reproduce	*Organize/Reorganize*	*Invent/Generate*
1. To learn (metalingual/ mathetic)		* Summary (B, C) * Paraphrasing (A)	
2. To convey emotions (emotive)		* Narrative/personal story (A, B)	* Open essay (B, C)
3. To inform (referential)		* Letter to uncle describing a bike (A, B) * Self-description in a letter to pen-pal (A, B) * Formal note to head of school (A, B) * Message to family (A) * Application letter (B, C) * Letter of advice to a younger student (B, C) * Describing an object (B, C) * Describing a process (B, C)	* Reflective essay (B, C)
4. To convince/persuade (conative)		* Application letter (B, C) * Letter of advice to a younger student (B, C)	* Persuasive/ argumentative essay (A, B, C)
5. To entertain (poetic)			* Open essay (B, C)

Note. Several tasks were common for two populations and one task for all three populations.

The IEA studies have been, and continue to be, an important source for considering how to enhance students' learning at the international, national, and local levels. By reporting on a wide range of topics and subject matters, the studies contribute to a deeper understanding of educational processes within individual countries, and across a broad international context.

Current Views or Conceptualization

When the IEA Six Subject Survey was conducted, several participating countries had no prior experience in large-scale assessment. For this reason, national centers were provided with very detailed instructions on sampling and test administration. The order for the actions by test administration instructions were spelled out in minute detail. In fact, the survey served as an effective hands-on training in large-scale assessment methodology.

Since then, there has been considerable methodological progress in international assessments ranging across the whole process: conceptualization (assess-

ment frameworks), domain specification, sampling and design of task rotation, scoring guides, scorer training, data analysis methods, and presentation of results.

By administering different subsets of items to different subsamples of students, broad coverage can be achieved with a reasonable amount of testing time for each student. Such matrix sampling designs have been used in most of the international studies, and they have been implemented in several different ways, such as administration of different forms to different subsamples, and administration of a common core of items to all students along with different forms to different subsamples (Linn, 2002). Current studies, such as the Trends in International Mathematics and Science Study (TIMSS) and PISA, use different versions of balanced incomplete block designs, in which blocks of items are combined into booklets to obtain a balanced order of presentation and to obtain links among the different blocks.

Results of early international assessments were reported in terms of total number of correct scores or average percentage of correct scores until the late 1980s. However, when matrix sampling designs are used such reporting tends to be complicated and inefficient. Starting with the TIMSS 1995 study, the international studies have relied on item response theory (IRT) techniques to put results obtained by students taking different combinations of items onto a common scale. These techniques model the probability of a correct answer in terms of invariant item characteristics such as difficulty and discrimination, along with student ability, and they provide a basis for estimating performance on a common scale even when students have been given different subsets of items. Given that there is an overlap of items in successive assessments, IRT can also be used to put these onto the same scale, thereby allowing investigations of trends in performance.

Starting with the IEA Reading Literacy Study (Elley, 1994) the international studies reported their results on a scale with a mean of 500 and a standard deviation of 100. This study did not use a matrix sampling design, but it was the first international study that relied on IRT techniques (the Rasch model) to scale the data. Such scaling results in both positive and negative scores, and before publication these results needed to be transformed into more meaningful numbers. While the choice of the mean of 500 and standard deviation of 100 was arbitrary, it carries the advantage that results can be reported in terms of integer values without any decimals, and it has been adopted as a standard scale for reporting results from international studies.

Much of the reporting of international studies focuses on means, but there is also great interest in measures of variability, and in levels of performance at different percentiles. All this information can be obtained with the IRT-based scales, and it is regularly provided in the international reports. However, the simplicity and accessibility of the reporting are somewhat deceptive, because it is based on complex techniques that are not easy to apply in secondary analyses. Thus, the estimation of different statistics computed from matrix sampling designs requires the use of several so called "plausible values" computed for each student, and user-friendly software to support such analyses has only recently become available.

While the main emphasis in reporting is typically put on a single score representing the general level of performance in the domain under investigation, the international studies generally also report separate scores for different subdomains.

This information can, for example, be used to describe achievement profiles within countries in relation to different curricular emphases.

Both PISA and PIRLS have devoted a lot of attention to the scoring of constructed response answers. For instance, PIRLS provides, for each constructed response item, an analysis of what aspect of the construct it measures and what characterizes an acceptable, unacceptable, partial, or complete answer. In addition, authentic examples are provided to further clarify the qualitative differentiation between different responses. Such procedures have improved the reliability of scoring in international assessments.

Translation has also become a topic of growing priority. This will be discussed in more detail below.

Current Research

Summary of Current International Assessments

This section presents the main features of the current PISA, PIRLS, and the European Survey on Language Competences (ESLC), mandated by the European Council of the EU. For economy and comparability, these most recent large-scale international assessments in the domain of languages are presented in Table 17.3. Several new aspects will be discussed below.

Recent European Studies of Foreign Language Proficiency

Over the years, compared to other subjects, international surveys of foreign language proficiency have been sparse. Among them, a few should be mentioned.

The Assessment of Pupils' Skills in English in Eight European Countries In 2002, a European survey of English proficiency at the end of compulsory education was performed in eight countries: Denmark, Finland, France, Germany (partly), the Netherlands, Norway, Spain, and Sweden. The survey was initiated by the European Network of Policymakers for the Evaluation of Education Systems and was an expanded repeat of a 1996 study. All in all, around 12,000 students took part in the 2002 study, which comprised tests, a set of self-assessment questions, an extensive student questionnaire, and a questionnaire for teachers (Bonnet, 2004). In spite of certain problems with construct coverage and student representativeness, the study generated data of considerable interest, most of all for national analyses. As for international comparisons, the report emphasizes that the approach taken was to provide broad indications about pupils' performance, and it was not attempted to benchmark countries. Consequently, the comparative perspective was toned down (see http://www.reva-education.eu/spip.php?page=article&id rubrique=213&id article=203&lang=en).

The EBAFLS Project In 2002, a decision was taken by the European Council to develop a linguistic competence indicator for foreign language learning. This decision brought about an initiative by institutions in eight EU countries (France,

Table 17.3 Current major international assessments: PISA, PIRLS, and ESLC

Feature	*PISA: Reading literacy (OECD)*	*PIRLS (IEA)*	*ESLC (EU)*
Construct definition	"An individual's capacity to: understand, use, reflect on and engage with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society" (OECD, 2009a).	Reading literacy is defined as the ability to understand and use those written language forms required by society, valued by the individual, or both. Young readers can construct meaning from a variety of texts. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment.	The Common European Framework of Reference for Languages (CEFR) serves as the framework for the Survey. A sociocognitive model based on the CEFR's model of language use and learning has been adopted, identifying two dimensions: *the social* (functional language use in real life) and *the cognitive* (language as a developing set of competences, skills and knowledge).
Target groups	* 15-year-old students * OECD members and several non-members participate	* Students in their fourth year in school, at least 9.5 years old * Approximately 50 countries participate	* Final year of lower secondary education (15–16-year-olds) * Two most popular foreign languages studied (English, French, German, Italian, Spanish) * Approximately 1,500 per language or country; 14 EU member states
Skills and domains tested	* Continuous * Noncontinuous * Also: mixed texts and multiple texts	Two overarching purposes for reading: * reading for literary experience * reading to acquire and use information Four types of comprehension processes: * focus on and retrieve explicitly stated information * make straightforward inferences * interpret and integrate ideas and information * examine and evaluate content, language, and textual elements	* Listening, reading, writing * Short routing test to select the appropriate test booklet level.
Background	Questionnaires	Questionnaires	Questionnaires
Result reporting	Five processes or aspects of comprehension or reading literacy, condensed in 2009 into three broad categories: * access and retrieve * integrate and interpret * reflect and evaluate	* Continuous texts * Noncontinuous texts	Related to the CEFR levels, using standard-setting procedures.

Germany, Hungary, Luxembourg, the Netherlands, Scotland, Spain, and Sweden) to seek funding for a project aimed to investigate the possibility of producing banks of calibrated anchor items. The project, referred to as Building a European Bank of Anchor Items for Foreign Language Skills (EBAFLS), was granted financial support by the EU for three years (2004–7) and was organized on a cooperative basis, coordinated by Cito in the Netherlands. The project undertook to provide items focusing on reading and listening comprehension in English, French, and German. A large number of items from existing tests in the participating countries were collected, scrutinized, pretested, standard-set, and analyzed. Considerable differential item functioning (DIF) was found, meaning that item difficulties tended to vary considerably among the participating countries. Thus, one of the conclusions of the project was that identical test items could not automatically be used across countries and contexts (www.cito.com/research_and_development/participation_international_research/ebafls.aspx).

Challenges

International assessments have faced and are facing many challenges requiring critical analyses, solid research, and continuous development work.

Translation

Translation guidelines have been an essential part of international assessments. In the late 1960s, the IEA Six Subject survey established a methodology that has been followed and adapted in subsequent assessments. It recommended that two translators be employed who were to be specialists in the subject matter and experienced in item writing. In case of disagreement, a third opinion was to be heard. If possible, back translation was recommended. Literature survey texts were to be translated by a literary translator.

In PISA, high requirements are set for the translators. They are to be professional translators with a good command of the two source languages and cultures (English or French), and to be familiar with the educational systems and cultures of the countries involved and with the topics covered in the assessment.

The translation process recommended by PISA is double (forward) translation but from two parallel source texts, followed by national and international verification (Grisay, 2003; see Figure 17.2). Two calibrated source versions (source texts, STs), English and French, are used. Two translators produce two independent versions (TT_1 and TT_2) in the target language. These are reconciled by a third translator into one national version, verified by still a fourth, independent translator from the International Project Centre. Test booklets are sent to the International Project Centre for a final optical check of the layout of the texts.

Specific instructions are given concerning layout, choice of vocabulary and syntax, and avoidance of irrelevant clues. The translators are reminded that the guidelines provide advice and that cumbersome translations are avoided. The translators are also provided with specific translation notes attached to the texts. For every question item, it is explained whether answering the item requires

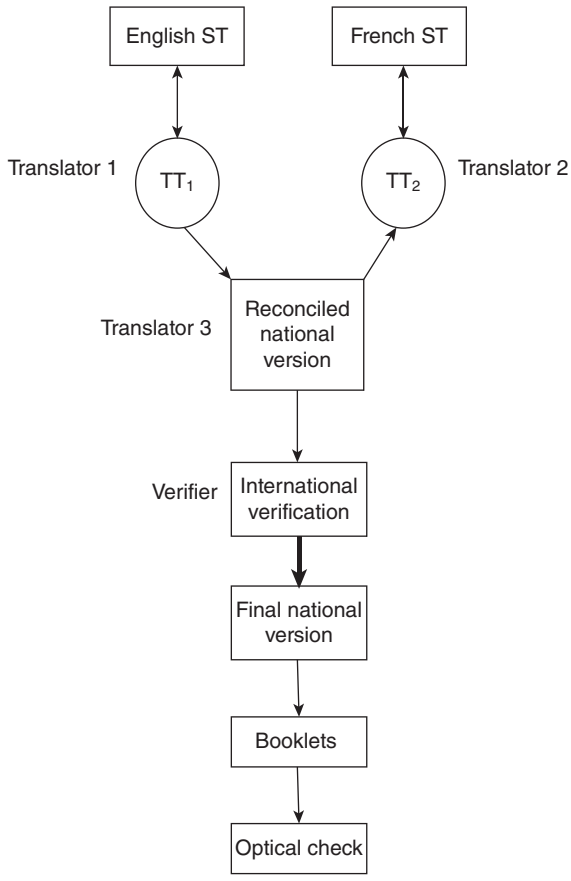

Figure 17.2 PISA translation and verification process (Arffman, 2007, p. 107) © University of Jyväskylä, Institute for Educational Research. Reprinted with permission

general understanding, retrieving information, developing an interpretation, reflecting on the content of the text, or reflecting on the form of the text. This is to avoid changing the nature of the questions and the strategies required to answer them correctly, because such modifications have been found to be one of the most typical reasons leading to shifts in difficulty (see Bechger, van Schooten, de Glopper, & Hox, 1998).

Valid results presuppose that all the different-language texts and translations are equivalent with each other, and hence equally easy or difficult to understand. Given this, it is unexpected that Arffman's (2007) linguistic analysis appears to be the first to explore in depth the equivalence of translations (PISA 2000 reading texts in Finnish). Statistical analyses of item "behavior" across countries have usually been considered sufficient. Another technique used extensively up to the early 1990s is back translation. If the original and the back-translated versions are similar, the target text is deemed to be of high quality and equivalent with the source text. This technique is relatively effective in detecting, for example, miscomprehensions and mistranslations. However, it may put too much weight on

the source text, surface structure phenomena, and literal translation (Grisay, 2003, pp. 227–8), as a back-translated text that is formally equivalent may sound strange and awkward and be difficult to understand. Thus back translation alone cannot guarantee high quality and equivalence with the source text (Brislin, 1986), and more recent reading literacy studies have not utilized it.

Some critical studies have been reported on recent international assessments (e.g., Bechger et al., 1998; Bonnet, 2002). They have pointed out significant shortcomings in the implementation of the studies and cited translations as one potential source of error, bias, and invalidity. This criticism has mainly concerned differences between languages and cultures, and claims that, due to these differences, translations will never be able to ensure full linguistic and cultural comparability. While the critics acknowledge that international reading literacy studies have improved during the last few years, they maintain that the distortions, including defects in the translations, still jeopardize the validity of the assessments.

Scaling Models and DIF

One problem is the effect of the aforementioned DIF on the interpretation of results. Kreiner (2011) claims that the fit of item responses to PISA's scaling model is often inadequate and that the ranking of countries is confounded by this. He offers two ways of dealing with the problem: (1) modeling departures from the scaling model so that measurement can be adjusted for DIF and other problems before countries are compared, and (2) purification by elimination of items that do not agree with the scaling model. Kreiner's criticism was promptly countered by the OECD (Adams, 2011), claiming that the fundamental flaw in Kreiner's argumentation is that he confounds two primary issues: (1) Do the outcomes of PISA depend upon the set of items that are developed and chosen, and (2) does the use of the Rasch model provide misleading results because the data do not *fit* the Rasch model? The conclusion drawn by the OECD is that Kreiner's analyses do not offer a better and more viable alternative than the one used in the regular PISA analyses.

Use of Computer Technology

The use of computer technology at the national and international levels offers great potentials for using a greater variety of more real-life tasks and achieving better cost-effectiveness. However, a certain cautious reflectiveness is called for, concerning theoretical as well as practical implications. Examples of matters to be considered thus range from construct definition to format effects and student computer literacy. Moreover, conducting technology-based assessments internationally poses formidable challenges due to variations in the level of infrastructure and the technological competence of the school staff. All these aspects are related to validity in an expanded sense and need to be discussed and analyzed as such (e.g., Björnsson, 2008).

Several computer-based studies have been conducted as part of large international surveys, e.g., within PISA (the Computer-Based Assessment of Science

[CBAS] in 2006 and the digital reading study in 2009), with full-scale studies being planned for the near future. Thus, it will be of considerable interest to see what the experiences of the IEA 2013 International Computer and Information Literacy Study (ICILS) project and PISA's plan to extend the use of computer-based assessment dramatically in all aspects of the 2015 survey will yield. Furthermore, the ESLC, conducted in 2011 and with a final report delivered in 2012, was offered in both print and digital versions, thereby generating data for interesting analyses.

Volume VI of *PISA 2009 Results* (OECD, 2009b) reports the experiences and results of the digital reading component of the reading literacy study.

Future Directions

As in all types of assessment, at least five fundamental questions need to be continuously addressed, namely *Why*, *What*, *How*, *Who*, and *And . . .?* This means that the different aims of international studies must be clarified and modified, constructs analyzed and problematized, and rubrics scrutinized and elaborated on; the same obviously goes for methodology at all stages of the process, for example test development, translation, and analyses of results. The role of different stakeholders is another crucial aspect of the assessment process. However, what may need the most intense attention is the interpretation and use of the results, and, in a wide sense, the various consequences—the impact—that they may have at different educational and societal levels, and perhaps even for individual students and teachers (e.g., Simola, 2005; Novóa & Yariv-Mashal, 2003; Hopmann, Brinek, & Retzl, 2007).

The alignment of content with assessment is likely to be one of the strongest priorities in both national and international assessments. Porter, McMalen, Hwang, and Yang (2011) is a good example of this trend, as it discusses the US core curriculum in mathematics and language arts and compares the results with three "international benchmark countries" with high student achievement: Finland, New Zealand, and Sweden.

As in all assessment, the definition of the constructs and their credible representation is a perennial challenge in international assessments. The breadth and depth of construct coverage are an obvious challenge, but the increased use of computer technology may ameliorate the situation in the future. Noncognitive factors may be expected to receive considerably more attention in national and international assessments. Motivation, liking of school, attitudes, interests, and so forth have been part of many designs in the past, but it is likely that there will be clear progress in doing a better job in future assessments.

Another probable trend is an increase in elaborative studies using the national and international assessment databases. Verhelst (2012) can be cited as an illustrative example. Using a newly developed method of profile analysis, he takes a closer look at the PISA 2000 Reading Data and reports interesting new findings. Sophisticated analyses such as structural equation modeling (SEM) are being used, but it is probable that new approaches will be further elaborated. Increasing attention will most probably be paid to the description and analyses of trends over

time in individual countries, thereby perhaps, to some extent, decreasing the interest shown in international comparisons that, so far, has often been the focal point of many comments and analyses. It can also be expected that there will be closer links to the educational effectiveness research (EER), which can be expected to have a positive impact on international assessments.

Large-scale assessments, both national and international, are here to stay. If the past fifty-odd years are anything to go by, the number of both assessments and participants will increase. International assessment is a "growth industry" (see ETS, 2011).

In spite of the growth of the international assessments and the increasing interest in the outcomes at many levels of stakeholders, there has been an undercurrent of critical response. As expected, the research community has found several grounds for critical views, especially concerning the methodology used and the validity of the findings. There has been hand-wringing and occasionally some drastic policy measures in countries that have done less well than expected, and admiration and envy of the high achieving countries, but it would appear that there has been little complacency in the latter. For instance in Finland, which has done well in PISA, the good results have caused a pleasant surprise but the dangers of complacency have often been voiced. It has been pointed out that the educational system has a number of problems to cope with, requiring continuous and consistent development work. Indeed, it would be useful to conduct systematic analyses of what discussions have emerged and what actions have been taken in well-performing and especially in less well-performing countries. Are there any signs of adapting teaching, testing, and examinations, and even national curricula, to be aligned with the PISA approach—"teaching to the test" in order to obtain a higher ranking? In other words, what is the inevitable impact of large-scale, comparative studies, whether perceived as positive or as negative?

There is widespread agreement that international assessments are extremely challenging and complex, posing questions about validity ranging from construct definition and coverage to interpretation, use, and consequences. Since large-scale assessments of the kind dealt with in this chapter have considerable influence at pedagogical, political, and personal levels, issues of impact must be given continuous attention. Equally important, however, is the fact that viewing the world as an "educational laboratory" or "educational experiment" holds promise for exploring and generating hypotheses, testing them, and gaining a better understanding of systemic and cultural effects. This means that, at best, international assessments can inform policy in positive directions concerning the learning of students as well as teachers, decision makers, and politicians.

The authors wish to acknowledge the valuable comments and suggestions by Professor Jan-Eric Gustafsson, University of Gothenburg, on the methodological discussion.

SEE ALSO: Chapter 4, Assessing Literacy; Chapter 32, Large-Scale Assessment; Chapter 66, Fairness and Justice in Language Assessment; Chapter 76, Differential Item and Testlet Functioning Analysis

References

Adams, R. (2011). *Comments on Kreiner 2011: Is the foundation under PISA solid? A critical look at the scaling model underlying international comparisons of student attainment*. Retrieved January 25, 2013 from http://www.oecd.org/dataoecd/21/58/47681954.pdf

Arffman, I. (2007). *The problem of equivalence in translating texts in international reading literacy studies: A text analytic study of three English and Finnish texts used in the PISA 2000 reading test*. University of Jyväskylä, Finland: Institute for Educational Research.

Bechger, T., van Schooten, E., de Glopper, C., & Hox, J. (1998). The validity of international surveys of reading literacy: The case of the Reading Literacy Study. *Studies in Educational Evaluation*, *24*(2), 99–125.

Björnsson, J. (2008). *The PISA computer based assessment of science: What did we learn?* Reykjavik, Iceland: Educational Testing Institute.

Bonnet, G. (2002). Reflections in a critical eye: On the pitfalls of international assessment. *Assessment in Education*, *9*(3), 387–99.

Bonnet, G. (Ed.). (2004). *The assessment of pupils' skills in English in eight European countries*. Retrieved January 30, 2013 from http://www.eva.dk/projekter/2002/evaluering-af-faget-engelsk-i-grundskolen/projektprodukter/assessmentofenglish.pdf

Brislin, R. (1986). The wording and translation of research instruments. In W. Lonner & J. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137–64). Beverly Hills, CA: Sage.

Brown, R. W. (1915). *How the French boy learns to write*. Cambridge, MA: Harvard University Press.

Carroll, J. B. (1975). *The teaching of French as a foreign language in eight countries*. Stockholm, Sweden: Almqvist & Wiksell.

Dickson, P., & Cumming, A. (Eds.). (1996). *Profiles of language education in 25 countries*. Slough, England: National Foundation for Educational Research.

Elley, W. B. (Ed.). (1994). *The IEA Study of Reading Literacy: Achievement and instruction in thirty-two school systems*. Oxford, England: Pergamon Press.

ETS. (2011). *International Large-Scale Assessment Conference, March 16–18, 2011*. Retrieved July 13, 2011 from http://www.ets.org/sponsored_events/ilsa_conference/agenda

Gorman, T. P., Purves, A., & Degenhart, R. E. (Eds.). (1988). *The IEA Study of Written Composition I: The international writing tasks and scoring scales*. Oxford, England: Pergamon Press.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing*, *20*(2), 225–40.

Hopmann, S. T., Brinek, G., & Retzl, M. (Eds.). (2007). *PISA zufolge PISA: Hält PISA, was es verspricht?/PISA according to PISA: Does PISA keep what it promises?* Vienna: LIT Verlag.

IEA. (*n.d.*). *Mission statement*. Retrieved January 25, 2013 from http://www.iea.nl/?id=72

Kreiner, S. (2011). *Is the foundation under PISA solid? A critical look at the scaling model underlying international comparisons of student attainment*. Retrieved January 25, 2013 from https://ifsv.sund.ku.dk/biostat/biostat_annualreport/images/c/ca/ResearchReport-2011-1.pdf

Lewis, E. G., & Massad, C. E. (1975). *The teaching of English as a foreign language in ten countries*. Stockholm, Sweden: Almqvist & Wiksell.

Linn, R. L. (2002). The measurement of student achievement in international studies. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement*, (pp. 27–57). Washington, DC: National Academy Press.

Noah, H. J. (1973). Defining comparative education: Conceptions. In R. Edwards, B. Holmes, & J. van der Graf (Eds.), *Relevant methods in comparative education. International Studies in Education, 33* (pp. 109–17). Hamburg, Germany: UNESCO Institute for Education.

Novóa, A., & Yariv-Mashal, T. (2003). Comparative research in education: A mode of governance or a historical journey? *Comparative Education*, *39*(4), 423–38.

OECD. (2009a). *PISA 2009 results: Learning trends. Changes in student performance since 2000. Vol. V*. Retrieved July 13, 2011 from http://browse.oecdbookshop.org/oecd/pdfs/free/9810111e.pdf

OECD. (2009b). *PISA 2009 results: Students on line. Digital technologies and performance. Vol. VI*. Retrieved July 13, 2011 from http://www.oecd.org/dataoecd/46/55/48270093.pdf

Papanastasiou, C., Plomp, T., & Papanastasiou, E. C. (Eds.). (2011). *IEA 1958–2008: 50 years of experience and memories*. Nicosia, Cyprus: Research Centre of the Kykkos Monastery.

Porter, A., McMalen, J., Hwang, J., & Yaong, R. (2011). Curriculum core standards: The new U.S. intended curriculum. *Educational Researcher*, *40*, 103–16.

Purves, A. C. (1973). *Literature education in ten countries*. Stockholm, Sweden: Almqvist & Wiksell.

Purves, A. C. (Ed.). (1992). *The IEA Study of Written Composition II: Education and performance in fourteen countries*. Oxford, England: Pergamon Press.

Simola, H. (2005). The Finnish miracle of PISA: Historical and sociological remarks on teaching and teacher education. *Comparative Education*, *41*(4), 455–70.

Spolsky, B. (1995). *Measured words*. Oxford, England: Oxford University Press.

Thorndike, R. L. (1973). *Reading comprehension education in fifteen countries: An empirical study*. Stockholm, Sweden: Almqvist & Wiksell.

Verhelst, N. (2012). Profile analysis: A closer look at the PISA 2000 reading data. *Scandinavian Journal of Educational Research*, *56*, 315–32.

Walker, D. A. (1976). *The IEA Six-Subject Survey: An empirical study of education in twenty-one countries*. Stockholm, Sweden: Almqvist & Wiksell.

Suggested Readings

Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education*, *11*, 319–30.

Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioural Statistics*, *32*, 252–86.

Grisay, A., de Jong, J. H. A. L., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, *8*(3), 249–66.

Hambleton, R. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Mahwah, NJ: Erlbaum.

van de Vijver, F. (2003). Bias and equivalence: Cross-cultural perspectives. In J. Harkness, F. van de Vijver, & P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 143–55). Hoboken, NJ: John Wiley & Sons.

Online Resources

IEA. (*n.d.*). *PIRLS 2006 Encyclopedia*. Retrieved July 13, 2011 from http://tims.bec.edu.pirls2006/encyclopedia.html

OECD. (*n.d.*). *OECD Programme for International Student Assessment (PISA)*. Retrieved July 13, 2011 from http://www.oecd.org/document/61/0,3746,en_32252351_32235731_46567613_1_1_1_1,00. html