International Assessments: Merits and Pitfalls

Anat Ben-Simon & Yoav Cohen

National Institute for Testing & Evaluation, Israel

Paper presented at the 30th Annual Conference of the International Association for Educational Assessment (IAEA), Philadelphia, June 2004

Abstract

Forty years have passed since educational achievements were first compared on an international scale. What began as a hesitant and sporadic attempt to compare scholastic achievements in various countries has grown into a well-established enterprise encompassing close to 50 countries worldwide. Perhaps as a function of globalization and increasing awareness of the role human capital plays in furthering economic development, policy makers around the world are expressing growing interest in the results of such surveys, realizing their importance for precipitating educational reform.

The quality of international comparisons of educational achievements has improved consistently as experience in the field has accumulated. Nevertheless, policy makers in many countries still fail to interpret the results of cross-national surveys in an accurate and useful manner, partly because they are unaware of the potential influence that diverse methodological factors have on the results of the tests.

The present paper discusses the impact of various factors, other than achievement, on test performance. These factors include sampling, administration, translation and adaptation, factors associated with the structure of the test (e.g., item format and test specifications) and lastly, response evaluation and data entry processes. A profound understanding of the role that these factors play in the assessment process can contribute to more accurate interpretation of the results obtained and a better appreciation of changes that occur over time.

Introduction

The idea of comparing the educational outcomes of different educational systems by means of standardized tests administered to large representative samples was first introduced in the mid 1950s. In 1955, a group of educators and education researchers from several countries began to meet on a regular basis with the goal of identifying ways in which educational research can contribute to the improvement of education throughout the world. The first formal meeting of the group, later named IEA, was held in Eltham, England, in 1958. At that session, a decision was taken to conduct a pilot study in order to determine the feasibility of a multi-national study of educational performance (Wolf, 2004). The first operational administration took place in 1964. Students aged 13 and 17-18, from 12 countries, participated in what was known as the "First International Mathematics Study" (FIMS). Top experts in education, psychology, and statistics contributed to the development of test design, sampling and administration procedures, and analysis of the results. In 1966, IEA was legally incorporated under Belgian law. Over the years many other countries joined the organization, which became a leading force in international assessment.

Approximately 15 large-scale comparative studies, in various subject areas, were conducted, primarily by IEA, between 1964 and the early 1990s. Most of the twenty-or-so countries that participated in these studies were developed countries. The fairly small number of countries involved during this period, and the constant changes in their constellation, made it difficult to draw significant conclusions from the results. In addition, in spite of the tremendous effort that was invested in improving the methodological quality of the studies, their quality was often criticized and their merit was questioned (Porter and Gamoran, 2003).

The abovementioned reasons may have contributed to the fact that, although most international comparisons were endorsed and financed by governments, policy makers showed little interest in the results for the first three decades (OECD, 1992). This attitude has been changing rapidly over the past decade. Policy makers around the world are expressing growing interest in the results of such surveys, realizing – albeit somewhat belatedly – their importance for precipitating educational reform.

Several factors have contributed to rising levels of interest in the results of international comparisons of educational achievements. These factors include

technological progress, increased international trade and competition, speed of communication, as well as globalization and increasing awareness of the role that human capital plays in furthering economic development (Keys, 1997). The growth of intergovernmental organizations such as UNESCO, OECD and the World Bank, and their increased interest in the promotion of education was also a factor (Robitaille and Robeck, 1997). The heightened focusing of these organizations on education, which incurred substantial investment in this area in developing countries, created a need for sound measures whereby the impact of the investment on educational outcomes could be assessed and progress monitored over time. These enterprises may well have contributed to the involvement of new countries in international surveys.

Nevertheless, policy makers and education experts in many countries occasionally fail to interpret the results of cross-national surveys in an accurate and useful manner. One possible explanation for this phenomenon is their inadequate background in the field of measurement and lack of awareness as to the potential influence of diverse methodological factors on the results of the tests.

The present paper discusses the impact of various factors, other than achievement, on test performance and the interpretation of score results. These include sampling, administration, translation and adaptation, factors associated with the structure of the test (e.g., item format and test specifications) and lastly, response evaluation and data entry processes. A thorough understanding of the role that these factors play in the assessment process can contribute to more accurate interpretation of the results obtained and a better appreciation of changes that occur over time.

Review of international comparisons of educational achievements (IEA, ETS/NAEP, OECD/PISA)

The following is a short review of the principal international assessment enterprises, their objectives, range of participating countries and scope of testing.

IEA Studies

The International Association for the Evaluation of Educational Achievement (IEA) is the oldest and most prolific organization conducting cross-national comparative studies of educational outcomes. IEA currently has 58 members. The IEA General Assembly, the main governing board, consists largely of officials from education

ministries, and some researchers. IEA studies are predominantly curriculum based, focusing on "the output of educational systems – that is, the attitudes and educational achievements of students – and attempt[ing] to relate these outputs to those inputs that have an effect on them. The overriding goal is to learn more about factors that influence student attitudes and achievement which may be manipulated to bring about improvements in attitudes and achievement, or efficiencies in the educational enterprise" (IEA, 2004). Over the past 40 years IEA studies have covered a wide range of subjects, typically focusing on two age groups, 9-10- and 13-14-year-olds and attempting to facilitate comparisons by grade as well as age group. Early studies also included 17-18-year-olds, yet differences in graduation age, school enrollment and specialization in various school subjects largely undermined the conclusions that could be drawn from these results. Over 20 international studies have been carried out by IEA over the past 40 years. These cover a wide range of subjects including mathematics, science, reading, civics, foreign languages and information technology (see table 1 for a list of the principal studies). The most recent and influential studies are the TIMSS, which assesses performance in mathematics and science, and the PIRLS which focuses on reading literacy. The assessment cycle of these two studies was set to four years. IEA tests are typically accompanied by student, teacher and principal questionnaires which collect data on values, attitudes, social background and school features. Information generated by these questionnaires is also compared across countries and is often used to generate hypotheses regarding the causal relations between educational achievements.

ETS/IAEP Studies

Two international surveys entitled International Assessment of Educational progress (IAEP) were carried out by the Educational Testing Service in 1988 and 1991. Both studies focused on the achievements of 9-10- and 13-14-year-old students in mathematics and science (see table 2 for more details). The studies were similar in design to the National Assessment of Educational Progress (NAEP). The IAEP studies made important methodological contributions to international surveys by introducing advanced analytical techniques. These techniques, such as the use of IRT models and differential item functioning techniques, proved useful in the context of NAEP (Linn, 2003).

OECD/PISA Studies

In 1997 the Organization for Economic Co-operation and Development (OECD) launched the Program for International Student Assessment (PISA). PISA is a three-yearly survey of the knowledge and skills of 15-year-olds who are approaching the end of compulsory schooling in the principal industrialized countries. The OECD/PISA assessment takes a broad approach to assessing knowledge and skills, moving beyond common curricula towards the application of knowledge in everyday tasks and challenges (OECD, 2003). PISA assessments focus on literacy in three domains, reading, mathematics and science. Each administration in a given cycle focuses primarily on one of the above domains, with the other two domains occupying a secondary position. In addition to the assessment of performance, PISA collects rich contextual information on students, families and institutions. This is later put to use in the interpretation of educational achievements. PISA's first operational administration took place in 2000, with reading as a major subject. The 2003 administration focused on mathematics and the 2006 administration will focus on science (see table 3 for more details).

Objectives and merits

The typical outcomes of most national surveys of educational progress are: a detailed description of the knowledge and skills possessed by students of a particular age or grade level in a given domain, a further description of contextual variables believed to be related to the above, and the pattern of relationships between performance and contextual variables, as well as within each group of measures. These outcomes and further analysis of the data collected facilitate comparison of achievements with local standards, monitoring of progress over time, comparison between various population groups and delineation of possible correlates of achievement. While national assessment may provide rich and valuable information, it also raises new questions: Is the level of achievement high enough? Is it possible to set and attain higher standards? Is reasonably quick progress made over time? Are differences in achievements between groups inevitable and within an acceptable range? Do the relationships between achievement and contextual variables make sense? Might there be other educational indicators, associated with achievements, which can explain differences in achievements observed between groups?

One of the main objectives of international assessments is to answer these questions. by giving educational systems the opportunity to compare their attributes and outcomes with those of other countries. This is accomplished through the administration of an identical set of tests and questionnaires. In fact, many researchers agree that the main benefit of large-scale international surveys of education stems from the fact that "education in one country can be better understood in comparison to education in other countries" (Porter & Gamoran 2003).

International assessments proffer few merits beyond those derived directly from the declared objective, namely the opportunity to compare educational systems on a wide range of educational variables. Participating in test development for international surveys provides professionals with a unique opportunity to examine a wide variety of curricula used in other systems, as well as the corresponding achievement standards or benchmarks associated with various levels of proficiency. Furthermore, the results of the surveys can highlight the strengths and weaknesses of individual education systems and facilitate identification of alternative approaches to teaching and learning as well as potentially useful instructional models. High quality international surveys offer all participants a unique opportunity to become familiar with advanced methodologies in testing and evaluation. They also enable professionals to implement new and innovative approaches and techniques in local research projects. In addition, since national assessments are usually conducted directly by government agencies (e.g., ministries of education), or closely monitored by them, and are often perceived as reflecting government performance, they risk being influenced by policy-makers of various ranks. By contrast, international assessments, being external, are far less susceptible to such interference and thus viewed by policy makers, the press and the public as more authoritative and more likely to provide an objective and true picture of the state of the education system. As a result, international assessments are highly effective in initiating educational reforms.

Reservations and pitfalls

As experience in the field has accumulated, the methodological quality of international comparisons of educational achievements has improved considerably. The current main projects, IEA and PISA, are known to maintain the highest professional standards of test design, sampling, administration, scoring and analysis. Leading experts in the fields of education, psychology and statistics are actively

involved in all stages of the assessment processes. Moreover, all stages are conceived, planned and closely monitored by a central body. Strong quality assurance procedures are implemented in translation, sampling and data collection. The expertise of the leading teams and the rigorous control of the various processes assure maximal standardization of the assessment. Both studies use multiple test forms and attempt to achieve rich content coverage. Both are highly concerned with test fairness and particular attention is paid to cultural differences, the heterogeneity of education systems, and how they relate to the nature and structure of the assessment (keys, 1997; Linn, 2003; OEDC 2003).

In spite of significant progress made in the quality of international studies over the past 40 years, several problems remain unresolved.

While some countries use the results of international surveys to initiate curriculum reform, other education establishments are pressured to adapt their curricula in order to excel by international standards. These adaptations can be content-related or pedagogical and are considered an imposition on local culture that is not necessarily beneficial. Another detrimental effect that arises from governments' desire to demonstrate progress during their term is the phenomenon of teaching to the test. It is not uncommon to witness extensive drilling of items similar to those that appear in the test.

The third type of problem pertains to interpretation of results. Critics of international surveys often point out the lack of clarity as to what conclusions can be derived from international comparisons. They also caution against misinterpretation of results, which can be a function of insufficient background in the field of measurement. The tremendous volume of results reported in most studies, the overload of methodological details and the complex relationships among the various measures, often make results unintelligible to policy makers and experts in areas other than measurement.

The last type of problem is methodological. Policy makers in many countries still fail to interpret the results of cross-national surveys in an accurate and useful manner, partly because they are unaware of the potential influence that diverse methodological factors have on the results of the tests. In the extreme case, these factors may seriously invalidate the results and lead to erroneous conclusions.

The remainder of this paper will focus on this category of problem and review several factors, other than knowledge and skills, that can affect performance on tests and which should be taken into account in the interpretation of international survey results.

Factors affecting score results

Keys (1997) identifies three fundamental conditions which must necessarily be met if an international survey is to produce reliable information on comparative achievements:

(1) the sample of schools and students in each country must be fully representative;
(2) tests must be as fair as possible to all countries; and
(3) administration procedures must be similar in all countries.

In spite of the immense effort made by the organizers of international surveys – central control and close monitoring of all stages of assessment, painstaking training workshops and highly detailed and specific manuals – the abovementioned conditions are not always met.

These factors include sampling, administration, translation and adaptation, factors associated with the structure of the test (e.g., item format and test specifications) and lastly, response evaluation and data entry processes. A profound understanding of the role that these factors play in the assessment process can contribute to more accurate interpretation of the results obtained, and a better appreciation of changes that occur over time.

The first question that should be addressed with regard to the interpretation of achievement results is the type of information that one should look for. The most popular and accessible measure is the rank of a given country in relation to other participating countries. While this measure is easily understood by non-professionals, it is also the least informative and least appropriate measure for comparison. The number of participating countries fluctuates from one survey to another, as does their constellation. Accordingly, a change in the ranking of a given country does not necessarily indicate a true change in achievement. Moreover, higher ranking can occasionally be associated with a lower mean achievement. In addition, sizeable differences in ranking, particularly those of countries that score close to the international mean, may often reflect negligible differences between their mean

achievement scores. This is largely due to the high density of scores around the mean.

Both IEA and PISA studies report standardized proficiency mean scores in the assessed domain. Though comparisons of standardized means across countries are far more informative, they should be used with caution when making inferences regarding progress over time. This is because changes in the constellation of participants from one survey to the next may significantly affect the international mean.

One way to overcome the above problem for the sake of monitoring change over time is to compare the means of given countries with the means of "anchor" countries obtained over several successive administrations of a given survey in a particular domain. In the case of the PISA studies, anchor countries may be the 30 OECD members, all of which are expected to participate routinely. While this approach to monitoring change over time may be viewed as somewhat improved, it should be remembered that the quality of education in any given education system may change over time, hence making the anchor a somewhat unstable marker.

In addition, extra attention should be paid to the interpretation of standardized means, with careful distinction made between mean scores obtained on a given administration and mean scores calibrated to previous administrations in the same domain. While the uncalibrated means allow for comparison of achievement with those obtained by all countries participating in a given survey, the use of calibrated means allows for monitoring of changes over time.

Perhaps the most informative measure of educational outcomes is the distribution of students across proficiency levels, namely, the percentage of students scoring at each proficiency level. Comparing these percentages to those obtained for other education systems may provide detailed and useful information for examining student proficiency in a given domain.

As mentioned above, several factors should be taken into account while interpreting results obtained from a given survey and comparing results over time. These factors include: sampling issues and changes in population, test specifications, translation and adaptation, administration, grading and data entry. The following is a brief discussion of each factor.

Factors associated with sampling & population

Strict adherence to a well-defined sampling plan is a necessary condition for drawing sound conclusions regarding the comparison between countries, or comparison over time. In addition, changes in the population composition should be taken into account in the analysis of changes in performance over time in a given country.

Sampling plan and frame

The sample of schools and students in each country must be fully representative to allow for valid comparison. Furthermore, the sampling plans used in each and every participating country must be identical. A decision must be taken regarding whether the sampling is to be conducted by age or grade-level. However, this is not a sufficient condition to guarantee the validity of comparisons. If sampling is conducted by age, all else being equal, students in a country in which formal schooling starts earlier will perform better than students in other countries. However, switching to sampling by grade-level would not eliminate the problem, as older children at the same grade level have, on average, higher achievement levels (Cahan & Cohen, 1989).

Countries may also vary in the structure of the sampling framework. Two factors are of particular relevance to the definition of the sampling framework: (1) the exclusion of special populations from the sample (e.g., independent education systems, students assigned to special education) (2) school enrollment rate. An extreme case of variability in the latter was demonstrated in the IEA First International Mathematics Study (FIMS). One of the age-groups examined in this study consisted of pre-university level students. While in the US the enrollment rate at this age (the percentage of the age group that reached this particular educational stage) was 70%, it was only 11% in France and in the Federal Republic of Germany (Table 5.1, Postlethwaite, 1967). Accordingly, the correlation between enrollment rates and mean mathematics scores in this survey was -0.62. In countries with low enrollment rates, the average student performance may be affected by various demographic characteristics of the studied group (e.g., gender, ethnicity). Thus, for example, in the FIMS study of achievement at the pre-university level, three quarters of the Australian senior classes consisted of males, while males and females were equally represented in the parallel age-group in the US. Thus, if a particular gender group tends to perform better, the fact that it is not equally represented in the sample may affect the

country's mean score. IEA researchers recently decided to refrain from comparing students above ninth grade because of problems resulting from differential enrollment rates.

Sampling methods and centralized monitoring of actual sampling of students in each of the participating countries are constantly being improved. In recent international surveys, cases of flawed sampling have, for the most part, been flagged and reported. Thus, the danger of misinterpreting the results of the surveys due to sampling-related problems has been minimized.

School and student response rates

Strict adherence to the sampling plan is also a necessary condition for drawing sound conclusions regarding differences between countries or cohorts. With the accumulation of experience in international surveys, increased attention has been paid to this aspect of the survey and, as a result, stricter control measures have been adopted. Nonetheless, sampling can be subjected to various pressures within a given education system, such as schools in remote and isolated areas being replaced by schools in more central areas in order to reduce travel expenses. Also, certain schools may refuse to participate in the survey on ideological grounds (open schools, anti-testing movements, etc.). Hence, it may prove extremely difficult to obtain the full cooperation of all students, parents, teachers and local authorities. As a result, a portion of the sample may not be covered. As long as the missing portion is fairly small, its effect on the results of the survey is expected to be negligible. However, a large proportion of missing cases (schools or students within schools) may invalidate the results. The fact, that countries that do not meet the sampling requirements are flagged, and in the extreme case even excluded from the final analysis, does not prevent educators, journalists and politicians from citing their results while completely overlooking the fact that the sampling was flawed.

Changes in population

When comparing achievements over time, researchers should be cognizant of increasing migration rates worldwide. This phenomenon makes it difficult to distinguish between changes effected by teaching and learning and those that result from demographic changes. Rapid demographic changes can produce unexpected results, as the following (hypothetical) example illustrates. Suppose that a country has a minority group that is less-affluent and less-educated than the rest of the

population and hence demonstrates significantly lower educational achievements. If the growth rate in this minority group is higher than the growth rate in the rest of the population, test results collected over time may exhibit a decreased average mean for the entire population, even if the mean for both minority and majority groups has increased. This paradox – an upward trend in each and every sector, yet overall decline – is known as "the Simpson paradox".

Test specifications

Linn (2003) points to the role of test specification in the interpretation of the results of international comparisons; "The particulars of the definition of the domain can have a significant impact on the relative position of nations on the assessment". These particulars relate to the relevance and representativeness of contents and cognitive processes that constitute the table of specifications for a given assessment. Too much emphasis on certain curricular topics may give certain countries an advantage, while putting others at a disadvantage. Likewise, over- or under-emphasis of certain cognitive processes may have a similar effect. It should be also noted that curriculum representativeness may vary, depending on the degree of congruence between the intended curriculum and the implemented curriculum.

Student familiarity with item formats is yet another factor. Early versions of international studies consisted almost exclusively of multiple choice (MC) items. With the growing awareness of authenticity in testing and the desire to assess more complex higher-order cognitive processes, current international studies employ a wide range of item formats such as, MC, short constructed responses, essays and performance assessment tasks. Familiarity and previous experience with the various item formats may have a substantial effect on student performance on these items. Of particular importance is the writing load required by the assessment. Not only do such items involve an additional skill, which may obscure the interpretation of the results in any domain other than writing, they are also highly susceptible to motivational factors. Writing is a highly demanding task, thus, education systems or population groups characterized by less disciplined or less motivated students are expected to do less well on items which require extensive writing.

In order to control for differences between the representativeness of different

curricula in determining the specifications of a given domain, some surveys (e.g. TIMSS) allow each participating country to select a subset of items that best represents its curriculum. There are at least two criteria according to which items may be selected for this subset: by comparing the representativeness of items in the intended curriculum and by comparing their representativeness in the implemented curriculum. The latter represents an emerging tendency to focus comparisons of achievement on the Opportunity to Learn (OTL). While either of these practices may solve the problem of curriculum representativeness, they run the risk of creating a different problem. Any process, in the course of which certain topics within content areas (rather than complete scales) are selected for the purpose of comparison, is bound to result in the selection of the easiest items in a given content area. This in turn, may lead to a significant shrinkage of the variability of scores obtained for all participating countries, and hence obscure the differences between them. Clearly, a selection based on the intended curriculum may yield a different set of items than one based on the implemented curriculum and thus lead to different interpretations.

Whereas test specifications may vary from one assessment to the next, within the same subject area, they may account for differences in achievement over time and thus should also be taken into account in the interpretation of such differences.

Administration factors

Countries vary to a great extent in their testing culture: the frequency of testing, the impact of tests, and the degree of reliance on test results for high stakes decisions affecting students (e.g., streaming, keeping back), teachers (e.g., promotion, laying off) and schools (e.g., budgeting). One of the most significant variables affecting performance, and perhaps the most overlooked, is student motivation, namely the commitment to performing well on tests. The fact that neither individual scores nor mean class scores are reported to the participating classes, and that student achievements have no impact on their annual report card, may play a key role in the degree of student commitment, more so in some countries than others. Reduced commitment is more likely to affect scores on items requiring extended cognitive effort such as those involving higher-order cognitive processes and those requiring extended writing. The above suggests a possible interaction between item format and motivational factors. One possible indicator of motivation is the ratio of missing responses on open-ended items,

while controlling for performance on MC items. Other indicators are attitudes towards school in general and attitudes towards the assessed domain in particular.

Student motivation levels can be somewhat improved by certain preparatory activities such as conveying the importance of the assessment for accurate monitoring of the nation's education system, presenting them with results obtained from previous surveys, and carrying out essential preparation activities (e.g., familiarization with various types of items and tasks). Most of these activities are appropriate and even recommended, as long as they are kept within reasonable limits. However, significant differences between countries in the intensity of preparation for the test may become another source of variability between nations.

Other administration factors which may vary between countries include the quality of the administration process per se; the caliber of testers and the quality of their training, the meticulousness with which the administration instructions are followed (e.g., keeping to time limits) and the degree of interference in the actual testing process (e.g., assisting, prompting, etc.).

A rather new source of variation between countries is the percentage of students with learning disabilities who receive test accommodations and the nature of those accommodations. While some countries are highly aware of the need to grant accommodations to learning-disabled students, other countries may be completely oblivious to this need.

Translation and adaptation of tests

When looking at the ranking of countries, it is easy to disregard the fact that although countries are ranked on a single scale, the measurements are derived from responses to items and questions that are posed in different languages. The multiplicity of languages raises several issues. The first issue is how to translate names and concepts from one language to another. In one of the last international surveys, a reading passage told the story of the Hungarian physician Ignaz Semmelweis. While his name has meaning in Germanic languages, it sounds peculiar in other languages. Hebrew-speaking readers may not even know how to pronounce it. Languages differ in the manner in which they name concept. A scientific concept may be left Latinized in one language while it is literally translated from Latin or Greek in another language, thus making it easier to understand even if the student is unfamiliar with the concept.

The issue of language cannot be separated from issues of cultural differences. Questions about a reading passage that refer to topics which are more familiar in one country cannot automatically be considered equivalent when presented in a different language in another country. Thus, reading passages relating to climatic conditions such as ice and snow cannot be presented in countries where cold conditions are rare. In order to consider translated items even arguably equivalent, it is advisable to check, prior to the operational administration of the test, whether the items function similarly in the two languages.

Differential Item Functioning (DIF) analysis can be used to detect cases of non-equivalence. As has been empirically demonstrated (Allalouf & Sireci, 1998), translated reading comprehension items can display considerable DIF, and the magnitude of the DIF depends on the item content and type.

The source language, the language from which the tests are translated, plays a role in determining the distance between the original item and its translated version. Translating within a family of languages will tend to be more accurate than across different families; vocabulary and grammar are more similar within a language family. When possible, it is therefore advisable to have more than a single source text and more than one version of the translated text. "Back translation", the process by which the target text is translated back to the source language and then compared with the original, is a good way to prevent cases of mistranslation. In particular, it can be used to detect cases of ambiguity in the source text that escaped the eyes of the translator.

There are also differences in the way that different languages are represented in writing. Languages differ in the number of words needed to express a particular idea, and in the number of characters (or signs/pictograms) that are needed for representing a word. The same passage would be shorter in English than in Russian, and will still be shorter in Hebrew. These differences are reflected by the space needed for the written text, or by the (font) size of the letters/signs. This in turn, may have implications pertaining to the amount of time and effort needed to read a written text. More time is needed for a longer text, potentially a problem if the tests are administered under strict time constraints.

Different language versions of the test may also differ in the register of the language, richness of vocabulary and complexity of syntax. These differences can occur as the result of a personal tendency on the part of the translator, but more often they are related to the culture associated with the language. There are languages in which writing in a "lower" register is considered bad taste, or even prohibited. In Arabic, for example, spoken language cannot be transcribed to paper. The writer has to switch to different vocabulary and syntax, which are usually more difficult to understand for the average reader.

In summary, preparing a test in multiple languages is not only a matter of translation but also a process of adaptation. It requires a good model of the minds of the typical test takers in different cultures, and must be accompanied by meticulous pretesting.

Evaluation, data entry and scoring

Countries differ in the standards of data entry. Hence, if no measures of quality control are instituted by the organizing body, difference in practice can increase the error of measurement in some countries. This, in turn, leads not only to bias in the mean achievement level but also increases the total variance and thus reduces the chance of detecting significant differences. In recent years, as experience in conducting international surveys has amassed, there are stricter measures of quality control that ensure that the effect of this kind of error be minimized.

More important than data entry is the process of marking or grading student performance. Here the questions are: who are the graders, how are they selected and trained. Are the graders in different countries given the same amount of training? Do they have the opportunity to see the full range of student performance in all countries? Does the answer key have the same meaning in different languages, or alternatively – was the answer key adapted as carefully as the test items were translated and adapted? Of course, these questions gain more importance in tests that require a good deal of reading and writing and as they include more open items and tasks. All of these factors have to be taken into account when drawing conclusions from test results.

Researchers should be aware of the fact that sometimes performance on a given test item or task cannot be compared across language and culture groups.

Summary

Given the potential variability of a vast number of factors between countries, all or some of which may affect performance on a given international assessment, one might well question the merit of international comparisons of achievement. Nevertheless, there is reasonable cause to believe that the effect of the factors discussed above on performance and interpretation of results is likely to be negligible, provided they are kept within limits. The expectation is that mean performance scores will remain robust in the face of such variability. However, in countries where multiple factors are simultaneously present, their combined effect will be augmented and mean scores significantly biased.

Certain measures are recommended in order to yield a fuller and richer depiction of international educational achievements, among them multiple comparisons. These are essential, as Mislevy (1995, p. 427) asserts, "because no single index of achievement can tell the full story and each suffers its own limitations, we increase our understanding of how nations compare by increasing the breadth of vision".

One example of multiple comparisons, as suggested by Linn (2003), is comparison of score patterns, namely comparison of topic scores within content areas. Comparison of scale scores and dimensions of cognitive processes is another. A third is comparison by item format. In the particular case of monitoring changes over time, changes in sampling, test specifications, test translation and adaptation, administration and marking should be closely monitored in order to account for their effect on fluctuation in results. In order to further validate results obtained from international studies it is recommended that group differences obtained from these surveys be compared with those obtained routinely from national surveys. Significant deviation in group differences obtained on international surveys from the nationally accepted ones may indicate a problem and should be examined further.

Lastly, an interesting challenge to the organizers of international surveys would be to design studies and develop procedures that assess the extent to which these factors are present in a given assessment and estimate their potential effect on the observed achievements.

Table 1: IEA main studies

Year	Assess.	Domain	No. of Countries
		Mathematics & Science	
1964	FIMS	Math (1st)	10-13
1970-1	FISS	Science (1st)	16-18
1980-2	SIMS	Math (2nd)	13-18
1984	SISS	Science (2nd)	13-17
1995	TIMSS	Math + Science (3rd)	46
1999	TIMSS-R	Math + Science (4th)	38
2003	TIMSS	Math + Science (5th)	51
		Language	
1971-2		Reading Comprehension	15
1991	RL	Reading Literacy	32
1997		Language Education	25
2003	PIRLS	Reading Literacy	35
		Other domains	
1070-1	Six Subject Study	Science, Reading comprehension, Literature, English as a foreign language, French as a foreign language, Civic	8-15
1989		Computers in Education	22
1991		Civic Education	28
1992		Computers in Education Information technology in Education (1st)	
1996	LES	Languages in Education	
1999	SITES-M1	Information Technology in Education (2nd)	
2002	CIVED	Civic Education	
2002	SITES-M2	Information Technology in Education (2nd)	

Table 2: ETS/IAEP Studies (ages, 9 and 13)

Year	Assess.		No. of Countries	Elementary school	Inter-mediate
1988	IAEP1	Math + Science (1st)	6 12 systems	-	13
1991	IAEP2	Math + Science (2nd)	20	9	13

Table 3: OECD/PISA Studies (age, 15)

Year	Cycle	Subjects	No. of Countries
2000-2	PISA 1st	Reading Math Science	43
2003	PISA 2nd	Math Science Reading	42
2005	PISA 3rd	Science Reading Math	48 interested

Bibliography

Allalouf, Avi & Sireci, Stephen, G. (1998). *Detecting Sources of DIF in Translated Verbal Items,* Research Report no. 245. Jerusalem: NITE.

Cahan, S. & Cohen, N. (1989) Age versus schooling effects on Intelligence Development. *Child development, 60*, 1239-1249.

Chromy, J. R. (2003). Sampling issues in design, conduct and interpretation of international comparative studies of school achievement. In Porter, A. C. & Gamoran, A. (Eds.) *Methodological Advantages in cross-National Surveys of Educational Achievements*. National Research Council, National Academy Press, Washington, DC.

Hambleton R. (2003). Adapting achievement tests into multiple languages for international assessments. In Porter, A. C. & Gamoran, A. (Eds.) *Methodological Advantages in cross-National Surveys of Educational Achievements*. National Research Council, National Academy Press, Washington, DC.

Keys, W. (1997). What do international comparisons really tell us? *International Electronic Journal for Leadership in Learning. 1, (4).* http://www.ucalgary.ca/~iejll. Retrieved 2/11/2003.

Linn, R. L. (2003). The measurement of student achievement in international studies. In Porter, A. C. & Gamoran, A. (Eds.) *Methodological Advantages in cross-National Surveys of Educational Achievements*. National Research Council. National Academy Press, Washington, DC.

Mislevy, R. J. (1995). What can we learn from international assessments? *Educational Evaluation and Policy Analysis, 17*, 410-437.

Organization for Economic Co-operation and Development (XXXX). The OECD program for International students Assessment http://www.pisa.oecd.org/. Retrieved 10/16/2003.

Organization for Economic Co-operation and Development (2003). *The PISA 2003 assessment framework – mathematics, reading, science and problem solving knowledge and skills*. Paris: OECD.

Porter, A. C. & Gamoran, A. (2003). Progress and challenges for large-scale studies. In Porter, A. C. & Gamoran, A. (Eds) *Methodological Advantages in cross-National Surveys of Educational Achievements*. National Research Council, National Academy Press, Washington, DC.

Postlethwaite, N. (1967). *School organization and student achievement*. Almquist & Wiskell: Stockholm.

Robitaille, D. F. & Robeck, E. C. (1996). The character and the context of TIMSS. In Robitaille, D. F. & Garden, (Eds.) *Research Questions & Study Design* (TIMSS Monograph No. 2). Vancouver: Pacific Education Press.

Wolf, R. M. (2004). *The contribution of IEA to research and education*. Paper presented at the 1st IEA International Research Conference (IRC-2004), Lefkosia, Cyprus.