Chapter 6

A discussion of some methodological issues in international assessments

Wim J. van der Linden*

Department of Educational Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, Netherlands

Abstract

Three topics addressed in the previous chapters are identified and discussed from a somewhat different perspective from that of the chapter authors. The topics are: the level of scoring in assessment studies, translation of test items, and sampling of curriculum content. Based on the analysis of these topics, five recommendations are offered. International assessments should be scored and reported at a more specific level that is currently the practice. There is a need of sound statistical checks on the quality of item translations. Rather than sampling the curriculum only once or twice, sampling could be in real time and on a permanent basis. Tests could be administered with open books and a well-chosen time limit per item. Finally, schools could be instructed to prepare their students for the assessment. © 1998 Elsevier Science Ltd. All rights reserved.

This issue brings together a group of authors who deal with a variety of methodological problems in international assessments. The majority have been involved in the Third International Mathematics and Science Study (TIMSS) conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IAE), an organization with decades of unparalleled experience in the practice of international assessments.

The publication of this issue is a timely event. With the current trend toward internationalization of the economy, national governments as well as local organizations have come to realize that they now have to compete in a global market, not only to sell their services and products but to hire qualified personnel as well. The requirements of this new economic order have already had their impact on national educational systems. Teaching institutions in higher education have begun to prepare their students for the global job market, particularly by offering them access to international exchange programs as part of their curricula. In doing so, one of the first

*E-mail: vanderlinden@edte.utwente.nl.
questions they have faced was the one of the comparability of degrees and course credit across national borders. Also, professional organizations around the world have realized that a global job market will expose their home countries to an influx of foreign workers with unknown qualifications and that they need to increase the visibility of the qualifications of their own members abroad. In Europe, several of these organizations are now in the process of equating their certifying examinations.

An important factor in the attempts to make degrees and certificates comparable could be a systematic input from international educational assessments. We should therefore be grateful to the authors of the contributions to this special issue for their willingness to share with us their experiences in international assessments and inform us about the technical problems they have met in their work.

In this chapter, each contribution is not discussed individually; rather, three topics addressed in these contributions are identified and discussed from a somewhat different perspective from that of the chapter. The topics are the level of scoring in assessment studies (Beaton, Schmidt, Jakwerth and Mcnight, Wolf), translation of test items (Ercikan), and sampling of curricular content (Foy, Wolf). This chapter concludes with a few recommendations on the future of the methodology of educational assessments.

1. Level of scoring

A persistent feature of international assessments is the interest in overall ranking of the participating countries. Upon the release of the results of TIMSS in the Netherlands, the media highly touted the released ranking of the country and politicians, without any reservations, took the ranking as a confirmation of the quality of current educational practices in mathematics and sciences. No interest was expressed in the specific aspects of teaching or the curriculum that might have furthered this result. Neither was any interest found in relative rankings on more specific topics in these two large domains or in possible difference between relevant national subpopulations. Of course, curriculum experts will follow up with more detailed analyses of the TIMMS reports and no doubt come to interesting conclusions. But to the general public, international assessments are just about ranking of countries.

The point here is not so much that potentially useful information in assessments is lost if attention focuses only on overall rankings of countries. Rather, it is that, as already put forward by Schmidt, Jakwerth, and McKnight (Chapter 2), these rankings are superficial, the result of arbitrary definitions, and often just misleading. A few comments are necessary to substantiate this point.

First, the idea of ranking countries is based on the assumption of an underlying unidimensional variable on which the countries can be ordered. For international assessments this assumption of unidimensionality is unlikely to hold. Familiarity with comparable data sets suggests that in international assessments the full set of item responses for all countries needs to be explained by a multidimensional structure of underlying variables. On each of these variables countries show different rankings. No overall ranking can be inferred.
If the responses on items in large-scale assessments are, nevertheless, pressed into the framework of a unidimensional variable, any ranking of countries on this variable depends on the incidental selection of the items in the pool. More precisely, such rankings depend on the loadings that the items have on the set of underlying variables. As a consequence, the result is necessarily unstable with respect to item selection. A replication of the study with a different selection of items — for example, with a subset from the current item pool — is bound to yield a different ranking, even if the new pool of items measures the same underlying structure of achievement variables and the national distribution of each country over these variables has remained constant.

It is true that most large-scale assessment studies supplement overall rankings with rankings on subsets of items addressing a common topic or content area. However, this practice is either inconsistent or meaningless. If the achievements on the total set of items are truly unidimensional, the ranking of any groups of examinees on any subset of the items should be identical to their overall ranking. On the other hand, if the achievements on the total set of items are not unidimensional, overall rankings are meaningless and should not be provided.

Also, it should not be forgotten that in educational assessment distributions of achievement have to be ranked. The notion of ranking or ordering distributions is much more complicated than that for individual objects. The problem of ordering distributions is a well-known problem in statistics, where the criterion of stochastic order has been developed for this purpose. This criterion tells us that distributions on the same variable are ordered if their cumulative distribution functions do not cross. A fictitious case of two distributions meeting this condition is given in Fig. 1a.

According to this criterion, Distribution B should be ranked higher than Distribution A. Only a few families of distributions meet this condition; one of the best-known cases being the family of normal distributions with a common variance. Empirical distributions of test scores can seldom be ordered. Such distributions usually take

Fig. 1. Fictitious examples of two distributions that (a) meet the condition of stochastic order and (b) do not meet the condition of stochastic order.
a large variety of shapes — one distribution may be more skewed than another, some may be multimodal, all have different variances, and so on. As a consequence, their distribution functions cross one or more times. Two fictitious empirical distributions are given in Fig. 1b; for these distributions it is hard to tell which should be ranked higher.

Reducing empirical distributions to a single statistic and then ranking them by their value for this statistic always involves an arbitrary choice. A choice for the mean will produce another ranking than for the median or, say, the 80th percentile. None of these choices can be taken to result in a “true ranking”. For educational assessments, a fair practice would be only to report the full score distributions for all countries. The second best choice would be to provide multiple rankings for a set of statistics that together describes these distributions satisfactorily, not singling out any of them as the ranking.

Reporting assessment results requires even more nuance. In the past, school districts or municipalities in several countries have tried to rank individual schools under their jurisdiction. One of the standard objections to early studies was that ranking on the basis of achievement alone is not fair. Schools differ in the conditions under which they have to operate as well as the input of students from previous programs or local families they have to accept. The generally accepted idea is that a better practice would be to compare them only for “added value”, that is, for the knowledge and skills schools add to whatever students already know and can do when they enter their program given their individual background. The statistical technique to deal with this requirement is conditioning achievement score distributions on relevant variables. The same principle can be defended for international assessments, particularly if the comparison is between populations of students in secondary education. According to this point of view, possible rankings of countries in international assessments should be based on conditional rather than marginal distributions of students’ achievements.

2. Translation of test items

The question of whether test items remain parallel when they are translated into multiple languages is a basic concern of international assessments. In the first half of this century, item translation was an exclusive interest of psychologists, particularly of those involved in the translation and renorming of major test batteries developed in the US for use in other languages. With the emerging internationalization of education as well as the trend towards giving comparative education a more solid empirical footing, educational researchers and measurement specialists also faced the problem of item translation. In fact, they soon discovered that, due to the linguistic nature of knowledge, their problems in translating achievement items were much more complicated than those for items measuring psychological abilities or skills. The topic of test item translation is now the object of much research, and a large variety of rules for checking test translations have been documented. Worth mentioning is a recent project to formulate guidelines for test translation coordinated by the International
Test Commission (e.g., Hambleton, 1993). Several of such guidelines are reviewed in Ercikan (this issue), whereas Wolf (this issue) explains the item translation practice in TIMSS.

Even if all guidelines for item translation have been followed carefully and the results been checked meticulously by content and linguistic experts, there is no guarantee that translated items behave identically to their original versions. However, a helpful resource is statistical analyses of the distributions of the item responses in populations with different languages, with appropriate matching on achievement level. The proposal by Ercikan (this issue) to check item translation results by applying a statistical technique for differential item functioning (DIF), known as the Mantel–Haenszel (MH) test, is valuable. In fact, this might even be one of the very few cases where this technique can be applied meaningfully. Some psychometricians, the current author being one of them, do not believe that, except for incidental cases where an item contains an homonym that can be interpreted meaningfully in different subpopulations of examinees, the same physical test item could ever function differentially (see below). This question does not apply to the problem of checking item translations where we do have different physical versions of the same intended question. Then it does make much sense to test the statistical null hypothesis of no difference between these versions against the alternative hypothesis of some of the versions functioning differentially.

An important question is if the application of the MH test does have power exclusively against this specific alternative. If its statistic appears to be sensitive to other features of the response data, the result may be items that are flagged unnecessarily for wrong translations as well. The MH test does have the tendency of producing false negatives. Also, it is argued that this tendency is particularly strong in educational assessments with number-correct scoring at a high level of aggregation, such as in TIMSS.

In a recent review, O’Neil and McPeek (1993) summarized the results of a decade of empirical research on the relation between DIF and characteristics of test items, mainly for such comparisons as male–female and minority–white groups. Inspection of their results shows that, without any exception, all cases of DIF concern items that require certain other knowledge in addition to the knowledge they were intended to measure. An example is an algebra item requiring geometric insight. Or a physics item with a description of a case that, for example, assumes knowledge about automobiles. If the focal and reference groups differ in their command of this unintended knowledge, the distributions of their responses on the item will be different no matter if we condition on intended knowledge. The point here is that the reason for the result is not a differential functioning of the item — the same knowledge is required from each member of the two groups to answer the item correctly — but the multidimensional nature of this knowledge and its distribution in the two groups. A more extensive analysis of this relation between “differential item functioning” and multidimensionality is given in Ackerman (1992).

The MH test neglects the multidimensional nature of DIF. As a consequence, it can produce false negatives for the following reason. Assume that some of the items in the assessment require unintended knowledge. If we condition or match on the
number-correct scores on the full test to make the comparison between different versions of the item fair, as the MH test does, and, for example, the examinees in the focal group have a lower level of unintended knowledge, then they need a higher level of intended knowledge to produce the same number-correct score. Under these conditions, the conditional p-values of the items in the test that measure only the intended knowledge variable can be expected to be higher in the focal than in the reference group. As a consequence, if the sample is large enough, the MH test will produce a significant difference. This case is not hypothetical. The author of this chapter once produced an example of 15 items that were fully parallel for the focal and reference group. These two groups differed only in unintended knowledge on six of the items. Nevertheless, all but one item was flagged as “DIF” by the MH statistic (van der Linden, 1993).

As already argued, in international assessments the achievements are bound to represent multidimensional rather than unidimensional knowledge. Also, national populations can be expected to have different distributions on each of the dimensions; in fact, international assessments are designed just to detect such differences. All conditions for the MH test to produce false negative results are thus present.

3. Sampling of curricular content

Another key issue in international assessments is the (content) validity of the tests with respect to the intended curriculum. How TIMSS has dealt with the problem of content validity is described in Wolf (Chapter 1). The main vehicle to guarantee content validity is an elaborate test blueprint. Test blueprints are two-dimensional tables with dimensions that represent the topics and objectives in the curriculum. The cells in this table define the weights that are to be used when the curriculum is sampled for a pool of test items. However, since curricula extend over time, sampling with respect to the dimension of time is also necessary. As described in Foy (Chapter 5), sampling of the time dimensions in TIMMS was realized by the definition of the populations of students in the study. The definitions of these populations were (partly) grade level based. The reason for this choice was that grade level was considered to be an important determinant in the process of getting instruments with good curriculum coverage.

The item pool was thus created sampling topics and objectives as well as the time dimension of the curriculum to ensure content validity. The question is whether this type of sampling does produce fair comparisons and meaningful results.

In educational assessments, test administrations are typically planned to fit one or more regular class hours in schools, preferably such that they do not interrupt class management too much. To the students, however, the tests are a dramatic interruption of their flow of learning. All of a sudden they are asked to produce answers on test items addressing topics that may have been learned months or even years ago. Unlike their usual tests and quizzes, they had no opportunity to prepare for the test. They are assessed only for what they remember here and now. This practice is believed to result in a large underestimation of what students actually know and can do.
Sooner or later, everything that is learned is forgotten. For this reason, most professionals practice what they have ever learned in school with a selection of their handbooks permanently on their desks. For more mathematical subjects, such as those assessed in TIMSS, the speed of forgetting is even more dramatic. The author of this paper has taught classes on test theory, statistics, and research methods for over two decades. Nevertheless, he regularly has to go back to his textbooks to check equations and formulas if some of them are needed in his classes.

It is just not realistic to expect schools to produce students who, at any time, have immediate access to anything learned in the past. Once graduated, students are not considered to be successful because they have a large reservoir of knowledge at hand. What distinguishes successful graduates from failures is the speed at which they refamiliarize themselves with old knowledge and their high rates of success when this knowledge is applied to new problems. These factors are typically not tapped in educational assessments. These assessments therefore give an unnecessarily pessimistic impression of the level of achievement of the students in the populations assessed.

A well-known distinction in educational assessment is the one between an implemented and an “attained curriculum”. As already indicated, implemented curricula are processes evolving over extended periods of time. However, “attained curricula” are no fixed quantities either. After initial learning has taken place, “attained curricula” change as a function of the reinforcements offered by the implemented curriculum as well as individual forgetting. Current knowledge then is a product of complicated interactions between learning and retention curves; however, interest in educational assessments typically is more in the former than the latter. Beaton (Chapter 3) presents a promising method to account for national differences in “opportunity to learn” in international comparisons. By the same logic we also need a method to account for the “opportunity to forget”.

4. Some recommendations

The recommendation for scoring international assessments at a more specific level and reporting on multiple variables should not be taken as a plea for returning to the early days of assessment when the interest was only in achievement on individual items and the comparisons were made between item p-values. This practice runs the danger of capitalizing on incidental properties of individual items. Rather, it is a plea for the definition of meaningful topics or content areas that lead to homogenous sets of items and allow for unidimensional scaling and scoring. The same message has already been given by Bock et al. (1982). Such definitions are not obtained by careful curriculum analysis and test blueprint writing only. In addition, multiple rounds of item pretesting, scaling, and, possibly, revising items, and reorganizing the item pool are required until the pool contains subsets of items which provide meaningful interpretations and have proven unidimensionality. In a word, what is needed is a practice known as calibrated item banking. Several testing organizations follow this
practice in their programs. Most assessment programs, however, still follow a one-shot approach and scale their items only after the main data collection has taken place.

Item translation is the Achilles heel of international assessments. The only sufficient check on item translations is statistical analyses of the empirical response distributions on the items in populations with different languages. As already argued, the problem of item translation is an obvious case for applying a test for differential item functioning such as the Mantel–Haenszel test. However, a necessary requirement for this test is that the test scores represent a unidimensional knowledge variable. The need of sound statistical checks on the quality of item translations is thus another case for the practice of item banking with the intent to obtain meaningful unidimensional subsets of test items.

The final topic addressed in this chapter was the sampling of the time dimension of the curriculum. What could be done to make assessments of educational achievements more realistic in this respect? Three recommendations can be offered. First, rather than sampling the curriculum only once or twice, sampling could be in real time and on a permanent basis. That is, achievement could be monitored throughout the curriculum, each time scoring the results on the same scales. Of course, this option would not fit the organizational structure of current one-shot assessments. But some countries do already have student-monitoring systems that operate on calibrated item banks. Once such systems are in place, national assessment is only a matter of sampling their databases. For international assessments, the (multiple) scales in such systems would have to be coordinated and linked. The practical problems involved in doing so are no doubt huge. But the methodological problems are not different from those that should be involved in attempts to make national curricula comparable in traditional assessments.

Second, tests could be administered with open books and a well-chosen time limit per item. This format would permit students to retrieve knowledge from their textbooks. Because of the time limits, only students who are able to retrieve the required knowledge quickly and manage to apply it successfully produce correct responses to the items. Timed administration of items is easily implemented in computerized testing. For group-based testing more complicated organizational procedures are required. Also, the need to set good time limits does require pretesting of the items under a variety of limits.

Finally, schools could be instructed to prepare their students for the assessment. This practice would enable students to refresh their knowledge and would neutralize differences between countries in the amount of time elapsed since the subject matter was taught. Of course, the amount of time spent on the preparation should be well chosen and uniform across the schools in the international sample. The latter requirement may be difficult to realize. As a matter of fact, this option simulates the way students typically prepare for a comprehensive examination at the end of school. For countries that do have national school leaving examinations, a natural implementation of this option would be to run international assessments through a process of coordinating and linking their current national tests rather than the administration of one or more earlier ones.
References


**Wim J. van der Linden** is Professor of Educational Measurement and Data Analysis, Faculty of Educational Sciences and Technology, University of Twente. He specializes in test theory, optimal design, decision theory, data analysis, and research methods. His publications have appeared in all major research journals and include the recent *Handbook of Modern Item Response Theory*, New York, Springer-Verlag, edited with R.K. Hambleton. He is the current President-Elect of the Psychometric Society and the chair of the NCME International Testing Committee.