How may we use international comparative studies to inform education policy?

by

Harvey Goldstein

University of Bristol

h.goldstein@bristol.ac.uk

Abstract

This paper reviews some of the methodological issues surrounding international studies of comparative educational achievement. It is argued that many of the claims made for these studies are unjustified and that their use for guiding educational policy is limited, especially because they lack longitudinal data on students.

Keywords

International comparisons, item response models, PISA, TIMSS, PIRLS, educational policy

NOTE: This paper is published in French as follows: -

Comment peut-on utiliser les etudes comparatives internationale pour doter les politiques educatives d'informations fiables ? In 'Revue Francaise de Pedagogie, 164, Juillet-Septembre 2008, 69-76.

Introduction

Comparative studies of student performance are frequent, large and expensive programmes that are widely quoted in the media and used by policy makers. They are organised by two rival bodies; the Organisation for Economic Cooperation and Development (OECD) and the Association for the Evaluation of Educational achievement (IEA), and are funded largely by governments of the countries that take part. The best known studies are the series of PISA studies from the OECD and the TIMSS and PIRLS studies from the IEA. The PISA studies attempt to measure the achievements of 15 year olds in schools in language, mathematics and science; the TIMSS studies concentrate on measuring science and mathematics achievement among fourth and eighth grade students, and the PIRLS studies concentrate on measuring reading literacy among fourth grade students. The results of these studies are frequently used for policy purposes, both directly in terms of changes to curriculum or indirectly through the publications of results that generate pressure for change. Details of these studies and the organisations that sponsor them can be found at their respective web sites: http://www.oecd.org/home/ , http://www.iea.nl/ .

In this paper I will explore some of the scientific issues raised by the studies, the ways in which they are reported and their possible uses.

Issues

The major issues can be formulated as follows:

- How the contents of the studies should be chosen to yield useful results
- How students should be sampled to provide valid comparisons
- How to decide what are the issues when comparing across different educational systems cultures and time periods
- How the data from these studies should be analysed

In the following sections I will try to provide insights into these issues. I shall refer to a number of published accounts of the methodological basis for these studies, one particularly useful one being that by Porter and Gamoran (2002).

Choosing questions and test items

The predominant statistical model used in the process of designing and choosing those items to be included in the achievement tests is the item response model (IRM), often referred to as ‘Item Response Theory’ (IRT). The most commonly favoured such procedure is the so called ‘Rasch model’ which is in effect just a basic factor analysis model where the measurements or indicators are binary rather than continuous. According to the proponents of these models, item selection should be predominantly governed by an underlying assumption of a single `dimension' of achievement. In other words that performance in say, ‘reading comprehension’, is governed by a single underlying factor or trait within an individual. One consequence of this assumption is that all the items in a test that purports to measure such performance, are perceived to be in the same difficulty order by all students. The great advantage of this assumption, if it can

be sustained, is that it enables all kinds of straightforward statistical scaling procedures to be adopted so that countries can be ranked along a single scale. This assumption is meant to hold universally, across cultures and educational systems, with group variation exhibited only in terms of average differences along such a dimension.

Prima facie such an assumption seems unreasonable and indeed it is not difficult to demonstrate empirically that this is so. Goldstein (2004) shows how a two dimensional model, one with two factors, picks up differences between France and England in terms of familiarity with multiple-choice formats (see also Goldstein et al., 2007). The first factor is a general factor and the second tends to distinguish between the multiple choice and free response items.

Nevertheless, because it enables cross-country comparisons in a straightforward way there is a natural reluctance to move away from the 'unidimensionality' assumption, and so one finds few attempts to test this assumption rigorously. Indeed, much of the 'item analysis' activity in the test development stage is concerned with rejecting items that do not conform to this assumption, so creating a test structure in terms of dimensionality that is largely self-fulfilling.

A critique is given by Blum et al. (2001) who point out that these procedures can lead to subtle biases that will depend on the actual countries involved in the study and may `smooth out' important country differences. Thus, PISA reports make no reference to the debate over such issues, but simply remark that 'items that worked differently in some countries were suspected of cultural bias. As a result, some items were excluded' (Kirsch et al., 2002, p. 21). Such items are referred to as `dodgy' and if an item is identified as such from separate analyses in more than 8 countries, it becomes a prime candidate for exclusion. In fact, these items may be informative and reflect interesting differences in responses between countries. That such items are considered suitable for exclusion illustrates clearly the emphasis on simple comparisons across countries rather than any attempt to understand the complexity of differences.

It may sometimes be the case that for certain purposes, such as pupil certification, aggregation of scores or subscales into a single 'unidimensional' scale is needed. In general this is inappropriate for comparative international surveys. The aim of these should be to obtain understandings about underlying differences between countries and to explore the data to reveal these. If comparisons are to be made between countries, then the existence of multiple dimensions should be reflected in such comparisons.

Sampling students

Students in international comparative studies are typically sampled either in terms of their ages or in terms of their grade or school year. Each of these has drawbacks and this implies that results need to recognise this. To make matters clear, the following example relates to comparisons between France and England (Goldstein et al., 2007).

One problem that arises in comparing France and England (as in other country comparisons) is that students move through the systems in different ways. Thus PISA samples by age, for example, all children born in 1984. In England, most children start school in September of the school year in which they reach the age of 5. There is almost no repeating of years, so that a 15-year-old at the time of the PISA survey in April/May

2000, born in August 1984, will start school in September 1988 and be in Grade 11 at the time of the PISA survey and in a class where there are a number of older children (not in PISA) born in September 1983 to December 1983. However, the first year of schooling is designated as reception, so that, in fact, that child will have been in formal schooling for 12 years. A child born in September 1984 will start school 1 year later and be in Grade 10, and this latter child is about the same age as the former but has had 1 year less schooling. In France, on the other hand, children start school in September of the calendar year in which they reach 6 years. Thus, a child born in August 1984 who does not repeat a year will be in Grade 10, as will be a child born in September, and both will have received the same amount of schooling. In France, the 1st year in school is counted as Grade 1. Any child who repeats a year (approximately one third do so by the age of 15) will be in Grade 9. Because the normal transition from college to lycee occurs after Grade 9, these children will be in college along with children who have not repeated, that is, those born in 1985. Thus, for the children born between September 1984 and December 1984, the French and English students will have been in formal schooling for the same length of time in terms of grades, although if reception is counted, the English will have been in school 1 year longer. For those born between January 1984 and August 1984, the French students will have been in schooling 1 further year less than the English, whether or not they repeat. However, 100% of French children are in preschool provision (ecole maternelle) for 2 years prior to formal schooling and 94% of French children 3 years prior. In England, about 80% of 3-year-olds are in part-time nursery education. All of this makes comparisons very difficult, if not impossible. Thus, for example, unless the school system structure is taken into account, the between-school variation for 15 year olds will be inflated for France compared to other countries that do not have high retention rates.

One way to deal with this particular problem is to incorporate longitudinal information so that prior achievements can be taken into account. Cross sectional studies can say little about the effects of schooling per se. Observed differences will undoubtedly reflect differences between educational systems but they will also reflect social and other differences that cannot fully be accounted for. To make comparisons in terms of the effects of education systems it is necessary (although not sufficient) to have longitudinal data and it remains a persistent weakness of all the existing large-scale international comparative assessments that they make little effort to do this. For example, the PISA (2000) report (OECD, 2001) claims that literacy level ‘has a net direct effect on pre-tax income, on employment, on health'. Such causal effects may indeed exist but they cannot be inferred from cross-sectional studies.

Cultural Comparisons

Among the desiderata for comparative studies many advocates of international comparative studies suggest that a study should be characterised by research neutrality (Porter and Gamoran, 2002). While this sentiment seems plausible it makes the strong assumption that there is such a thing as `research neutrality' along with the possibility of making culturally unbiased judgements. Such a concept may seem desirable, but it is nevertheless highly contestable. Yet as a result of the western funding sources, western dominated psychometric modelling and the primary use of English as a medium of

communication and item development, there is a prima facie case for supposing that there would indeed be a pro-western bias (see Goldstein, 1995, for further discussion).

There have been few systematic critiques of the major international studies. One of the most sustained is the reanalysis of the International Adult Literacy Survey (IALS) which was funded by the European Commission over a period of several years and involved interactions between the original team carrying out the study for OECD and a group of external researchers. For this reason many of the insights and critiques of international comparative studies can usefully be illustrated by reference to this reanalysis. Blum et al. (2001) present many of the principal findings and the full report is also available (Carey, 2000). It was shown that all kinds of cultural factors can influence performance.

Blum et al. (2001) give an example where repetition of a term in the question and in the accompanying text adjoining the answer is more frequent in the original English-language documents and forms a source of bias. For example, in a question referring to the use of '*couches jetables'*, the phrase containing the answer uses the term '*changes complets*'. In the Anglo-Canadian questionnaire the term 'disposable diapers' is repeated, as is the term 'disposable nappies', used in the British one. The respondent is drawn in English more easily towards the phrase containing this term and therefore to the correct answer, whereas in French the reader has to understand that these terms are equivalent before being able to answer. The resultant bias considerably increases the difficulty of the questions in French.

There is often greater precision of English terms and, as a general rule, the questions are drafted in a more precise form in English. For example, one question rendered in English by 'What is the most important thing to keep in mind?' is translated into French as '*Que doit on avoir a l'esprit?*' [What must be kept in mind?]. However the sentence containing the answer reads in English 'the most important thing' and in French *'la chose la plus importante'* [the most important thing]. The link is therefore easier to establish in English. Another task is defined in English as 'List all the rates'. It is translated into French as *'Quels taux'* [What rates], omitting to state 'all the rates'. This omission frequently led French interviewees in IALS to state only one rate instead of the list required for the answer to be considered correct.

There are also translation errors, some of which may be relatively unimportant from a strictly linguistic point of view but become important in relation to comprehension. The following example is particularly characteristic. A question rendered in French as *'soulignez la phrase indiquant ce que les Australiens ont fait pour …*' [underline the sentence indicating what the Australians did to…] is linked to a text worded: '*Une commission fut re´unie en Australie'* [A commission was set up in Australia]. In English the question becomes 'What the Australians did to help decide …', the corresponding wording being 'The Australians set up a commission'. The answer is ambiguous in French because the place is given instead of the people.

These authors conclude that in fact there were important differences between the actual orders of difficulty for the questionnaire items in the supposedly 'linguistically

equivalent' countries, and give examples also of where the real life context in which an item is embedded will affect its perceived difficulty and they list the situations where such discrepancies can arise.

In similar vein, Wuttke (2007) argues strongly that PISA, and by implication similar studies, are very little concerned with understanding how students respond to the test items and how they interpret them, focussing instead on sophisticated summaries using constructed scales. He also refers to the many problems surrounding the sampling strategies used by these studies. The criteria for acceptability tend to be based upon achieving target response levels, yet what really matters is whether the non-respondents are atypical and little effort generally goes into this aspect.

These findings are supported by other commentators. Thus, Grisay and Monseur (2007) conclude that 'the equivalence of a test instrument is always lost when translating it into other languages' (p 73), and this remains the case however careful translators are.

Bonnet (2002) is critical of the quality of the background data that are obtained in PISA, especially about the socio-economic backgrounds of childrens' parents, and suggests that this casts doubt on many of the 'explanatory' analyses carried out.

One inference that can be drawn from the IALS analysis and these other critiques is that a major interest in such studies could be the investigation of culturally specific responses, rather than relying upon an assumption that valid comparisons can be drawn by careful design of the tests, translations and the use of background data.

Comparisons over time

In 1972 two researchers at the National Foundation for Educational Research carried out a study of changing reading standards from the late 1940s to the early 1960s. They used results from repeated administration of the same test over this period and pointed out that the curriculum had changed and so had the use of language over this period, and for these reasons they suggested that the test itself had become "harder", so that apparent declines in test scores could not be viewed in any sense as a decline in standards of achievement. This duality of interpretation has long been recognised: in general, without making further assumptions, we cannot know, for example, whether the individuals taking a test or examination have in some sense become "better" or whether the test has become "easier" because the social, cultural or educational context has changed. Exactly the same considerations apply to comparisons over time for international studies. Two main procedures are used for attempting to overcome this problem.

The first of these is test equating, where the basic idea is that one has two different tests administered at two different times. There are several variants, but I shall describe just two; the "common item" procedure which underlies many practical schemes, and a sampling procedure. In the first approach each test contains a small number of identical questions, say 15% of the total: a small enough number to avoid detection but large enough to carry out satisfactory equating. This is a major reason for the international

comparative studies retaining an unpublished set of items. The assumption is that these items are invariant, that is, they can be assumed to have a common meaning at both times, while the other items are allowed to reflect changes in curriculum, general environment etc. The common items are then used as a calibration set to create a common scale over all the items in the tests. This common scale is then used to report any changes. The actual procedures used to carry out the scaling vary in terms of the complexity of modelling used, but typically some kind of item response model is used. The problem, however, is twofold. First it is necessary to make the invariance assumption for the common items and this, inevitably, is a matter for judgment which may not be universally shared. Then even if such an assumption is accepted, because the non-common items are allowed to reflect background changes, the relationship between the common item set and the non-common items can be expected to vary across the tests; yet it is necessary to assume that this relationship is constant. This second assumption is therefore contestable and also a matter for judgment.

An interesting example of these problems arose with the US National Assessment of Educational Progress, where there was a very large and unexpected drop in test scores over a 2-year period in the 1980s (Beaton and Zwick, 1990). A large-scale evaluation essentially concluded that the common item equating procedure was unreliable, for a variety of reasons including their juxtaposition with different surrounding items in the two separate instruments. Similar problems of equating using common items are clearly present in international comparative studies.

In the second approach the idea is that a very large bank of items is selected and for each test a random, possibly stratified, sample is selected for use. This means that, apart from sampling error, a common scale does exist and can be used for inference. Such procedures are often referred to as “item banking”, although that term is also used in other contexts. The difficulty is that the pool has to be selected *before* any of the tests are administered and it cannot be known in advance which items may become outdated and hence become “harder” etc. Thus, again, contestable assumptions about test item behaviour have to be made.

I am not arguing that these approaches are pointless, or that test equating is not useful in other situations. Rather, I am suggesting that they are not simple objective devices for solving the problem, but in fact involve important, and crucial, value judgments that may or may not find consensus among interested parties. One of the unfortunate aspects of much of the literature on equating is that this need to exercise value judgments is rarely stated.

In fact, the situation is made worse in the international comparative studies because the countries entering for any particular study change over time, so that any scaling and subsequent comparison is with reference strictly to the group of countries taking part.

Data analysis

One of the features of more recent international comparative studies is their attempt to take account of school differences using multilevel modelling. In essence a multilevel

model seeks to represent all the sources of variation that influence a given response variable. Thus, for example, pupil attainment is assumed to depend on various student characteristics, including such factors as gender, social background and prior attainment, and in addition on features associated with the school or schools that they have attended. When fitting a statistical model that includes such factors, we will also generally include factors that are characteristics of schools, but we typically find that there remains variation between schools that is not accounted for by either the pupil or school level characteristics. Multilevel models provide an efficient and valid representation of such a situation by explicitly incorporating such `residual' variation in the model. In addition it can readily be extended to allow gender or other differences to vary across schools. For a straightforward introduction to such models see Snijders and Bosker (1999).

Studying the relative amounts of variation at school level is a second-order comparison that potentially is both more valid and more interesting than the comparisons of means (Goldstein, 1995). Thus, for example, an analysis of Geometry items in the Second International Mathematics Study (SIMS) (Goldstein, 1987, chapter 5) shows that variation between schools in Japan is much smaller than that between schools in British Columbia, Canada. It is also well established that schools differ along a number of dimensions and that the between-school variation is a function involving random coefficients of other factors such as gender, social class, etc. If only the average variation is fitted and there are sizeable random coefficients, then important information is lost. Thus, for example, the relative amounts of variation between countries may vary by social group or parental education. If the average between school variance is small, nevertheless it may well be much higher for those coming from high and/or low social groups.

Another important issue relates to the ways in which the scores on the scales that are produced by the analyses are interpreted. Typically, cut-offs determine score ranges that are then given interpretations of what individuals can and cannot do in terms of reading literacy, mathematics etc. This leads to statements about the percentages in each country with ‘poor’ or ‘excellent’ literacy skills. Again, using the IALS Blum et al show how different ways of defining such cut scores can lead to very different conclusions. Thus, for example, using the IALS measure, 65% of French respondents have a ‘prose literacy’ level of 1 or 2 (the lowest two categories); whereas with an alternative measure based on the notion of maximal performance, this proportion falls to 5%. For the UK, the proportions are, respectively, 48% at level 1 or 2 using the IALS measure, and 3% at the same level using the alternative measure. Needless to say, the latter estimates are less amenable to sensational headlines.

One of the features of all the international studies is the secrecy that surrounds the release of items used; only a selected few are released. Various reasons are given for this, most notably the supposed need to retain some items so that they can be used to ‘equate’ tests across time. The problem is that unless users can see what items are actually used, it becomes difficult, if not impossible, to judge what the tests are really measuring and whether comparisons are really valid. The user, therefore, has to rely upon the judgements made by the test developers, and this then closes down a large area of fruitful debate.

Informing educational policy

Previous sections have dwelt upon some of the limitations of international comparative studies of achievement. It is important to recognise these limitations and the constraints that they impose on the legitimate use of results, but within these there may nevertheless be some useful messages from these studies.

There are certain basic requirements that any such study should adhere to. First, it is important that cultural specificity is recognized in terms of test question development and also that this is recognized in the subsequent analysis.

Secondly, the statistical models used in the analysis should be realistically complex so that multidimensionality is incorporated and country differences retained rather than eliminated in favour of a `common scale'.

Thirdly, the multilevel nature of any comparisons needs to be stressed, and here the limited attempts to do so are a welcome start. Comparing countries on the basis of the variability exhibited by institutions and possible explanations for differences, potentially provide powerful new types of insight for cross-cultural studies.

Fourthly, it is very important that comparative studies should move towards becoming longitudinal. With only cross-sectional data it is very difficult, if not impossible, to draw satisfactory inferences about the effects of different educational systems. Even following up a sample over a one-year period would add enormously to the value of a study.

Fifthly, there needs to be much more transparency in terms of releasing all items used in the tests so that users can properly judge what is being measured. Despite the possible technical drawbacks, such transparency should become a fundamental requirement.

Finally, all of these studies should be viewed primarily not as vehicles for ranking countries in crude league tables, even along many dimensions, but rather as a way of exploring country differences in terms of cultures, curricula and school organization. To do this requires a different approach to the design of questionnaires and test items with a view to exposing diversity rather than attempting to exclude the `untypical'. This point is discussed in some detail by Langfeldt (2007). Such a viewpoint will involve a different approach to the analysis of item response patterns in the context of the questions, and the acquisition of local information about the contexts of educational institutions. The complete process, from the choice of collaborators and advisors through to the publication of all questionnaires and test items, should be transparent. The naïve interpretations often placed upon results by governments and the media are typically unfounded but, more importantly, counter-productive. They can lead and actually have led to ‘panic’ reactions where these are not only unjustified by the evidence but are also time and resource consuming at the expense of more considered approaches.

International comparative studies of achievement studies should be treated as opportunities for gaining fundamental knowledge about the reasons for differences, not as competitions to see who comes top of the league table.

References

Beaton, A. E. and R. Zwik (1990). Disentangling the NAEP 1985-86 reading anomaly. Princeton, Educational Testing Service.

Bonnet, G. (2002). "Reflections in a critical eye: on the pitfalls of international assessment.." Assessment in Education **9**: 387-400.

Carey, S. (2000). Measuring Adult Literacy - The International Adult Literacy Survey in the European Context. London, Office for National Statistics.

Goldstein, H. (1987). Multilevel models in educational and social research. , London, Griffin; New York, Oxford University Press.

Goldstein, H. (1995). Interpreting international comparisons of student achievement. Paris, UNESCO.

Goldstein, H., Bonnet, G. et al. (2007). "Multilevel structural equation models for the analysis of comparative data on educational performance." Journal of Educational and behavioural Statistics **32**: 252-286.

Grisay, A. and C. Monseur (2007). "Measuring equivalence of item difficulty in the various versions of an international test." studies in Educational Evaluation **33**: 69-86.

Langfeldt, G. (2007). PISA – Undressing the Truth or Dressing Up a Will to Govern? PISA According to PISA. S. T. Hopman, G. Brinek and M. Retzl. Wien, http://www.univie.ac.at/pisaaccordingtopisa/

OECD (2001). Knowledge and Skills for Life: first results from Programme for International Student Assessment. Paris, OECD.

Porter, A. and A. Gamoran, Eds. (2002). Methodological advances in cross-nation surveys of educational achievement. Washington DC, National Academy Press.

Snijders, T. and R. Bosker (1999). Multilevel Analysis. London, Sage.

Wuttke, J. (2007). Uncertainties and Bias in PISA. PISA According to PISA. S. T. Hopman, G. Brinek and M. Retzl. Wien http://www.univie.ac.at/pisaaccordingtopisa/ .