Goldstein, H. (2017). Measurement and evaluation issues with PISA. In The PISA effect on global educational governance, ed. Louis Volante, New York and London, Routledge.

Measurement and Evaluation Issues with PISA

by

Harvey Goldstein, University of Bristol*

Abstract

This chapter comments on the restricted nature of the data modelling and analysis that underpins PISA, and the resulting interpretations. It points to certain features that raise questions about the provenance of the data collected, the ways in which it is disseminated and the limitations these impose on the interpretations that can be made in terms of international and over-time comparisons. Specifically, it discusses the issue of item translation and selection which PISA addresses through simplistic latent variable, or item response, modelling. It also looks at the ways in which data are provided to users and the technical limitations of this, and finally issues concerned with the legitimacy of causal inferences that have driven educational policy changes.

Keywords

International comparisons, PISA, comparability, item response models, educational assessment

*Graduate School of Education

35 Berkeley Square

University of Bristol

Bristol BS8 1JA, UK

h.goldstein@bristol.ac.uk

1. Introduction

Whilst the present chapter is devoted to PISA, the same issues occur across the whole range of international comparative assessment studies such as those of the International Association for the Evaluation of Educational Achievement (IEA), and indeed more generally when international comparisons are made. All of these studies face the well-known and fundamental problem of defining and then attempting to ensuring 'comparability of meaning' for their instruments across diverse educational systems and cultures. In the following sections I shall deal in turn with the following issues, all of which are relevant to comparability.

Translation of assessment instruments into different languages or dialects lies at the heart of comparability and the focus of considerable attention by instrument constructors. Closely associated with translation is the psychometric 'modelling' involved in question design analysis and dissemination of data and the key issues will be discussed, as far as possible, in a non-technical fashion. The third major issue is that of making causal inferences from assessment results, illustrated particularly by the use of country rankings to influence curriculum and other educational policies. Finally I will attempt to situate PISA and similar studies within a more general research paradigm, according to whose criteria, I will argue, it should be judged.

1. Translating assessments

The OECD through its various international assessment instruments seeks to ensure what might be termed 'translational comparability'. This term is intended to convey an understanding that any particular item in a test instrument, for example a textual passage with questions, can be translated to be understood in the same way in different educational systems. Since there is no objective external procedure for ensuring this, certain operations are applied. A common one is to design the relevant questions in a base language, typically a dialect of English, translate into another language then back-translate to English with an acceptance criterion related to the degree of match obtained with the original. The process may be repeated with questions being modified (or discarded) until an 'acceptable' match is obtained. Alongside this, certain psychometric measures may be used, for example to assess difficulty rankings of items, and I will return to these in the next section. An early publication for OECD perhaps comes closest to embodying this requirement

`*Meaningful measurement requires the instruments to be culturally and linguistically equivalent for all participating groups' (Kirsch et al., 2002, p. 19).*

There is, of course, much ambiguity here with comparability being defined as being 'equivalent' and 'equivalence' requiring translational matching. Of particular interest is the requirement for 'cultural' equivalence, including one assumes within the same language group, for example such as French as used in various European countries and parts of North America. There is no obvious definition of what this means but nor is there any recognisable attempt by OECD to test such an assertion. There is, however, other research into this issue. For example, an attempt to understand cultural equivalence was undertaken in a reassessment of another OECD study, the International Assessment of Adult Literacy (IALS) (Blum et al, 2001). Thus, translations of the questionnaires and documents from their original English language version were done in each participant country and checked. 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 between question and answer is therefore easier to establish in English. Such embedded language differences are not only hard to predict, but also virtually impossible to eliminate.

Blum et al. quote another example where the local cultural context is important. The task required is to work out which are the comedies in a review covering four films. In two of these reviews, in both English and French, the term 'comedy' appears, which makes the question easy. In France, however, many interviewees gave as their answer a third film, which from the description is obviously not a comedy. The only possible explanation is the presence in that film of the actor Michel Blanc, who is well known in France for his roles in many comedies, but is little known abroad. Here, association predominated in the answering process to the detriment of careful reading of the reviews. Blum et al. give other examples of translational problems leading to different results existing for example between French speakers in France and Switzerland.

We see in these examples the need to recognise inherent contextual constraints to any notion of comparability and a strong requirement for those designing instruments to be used across languages and cultures, to study such effects. Indeed, one of the aims of designing such instruments could be to gain insights into the nature of cultural differences in understandings. For OECD, however, the imperative is to make comparisons, so that such nuances, however important they may be in reality, will tend to be ignored. Instead, it is often argued that while there may be nuances of these kinds, they will not influence the overall picture in terms of country comparisons. Such a stance, however, is not backed by evidence and amounts to little more than special pleading.

2. Psychometric models

A major feature of OECD and IEA studies is the effort put into the development of sophisticated models to design, analyse and present the results of their studies. When educational assessments are carried out on a large scale it is common to report results in terms of summary measures such as 'reading achievement' or 'proficiency in algebra'. This is of course useful, but depends on each constituent of such summary scales reflecting the concept described by the label. This not straightforward to achieve and is a major concern of test constructors. For established test instruments and for many routine testing and examination programmes the instruments themselves are generally publicly available so that someone analysing the data or interested in its interpretation, can form a judgement about the extent to which the instrument is effectively assessing what it claims to assess. In the case of PISA, however, only a small percentage, generally less than 15%, is released for public view so that users of the data have to rely upon the descriptions of the test instruments and their interpretations that are provided by the PISA test constructors. The stated reason for this is that the retained test items are to be used to facilitate making comparisons over time. Even if this could be achieved, the failure to allow public scrutiny of the items is simply bad practice. Moreover, there are reasons to question whether this does really allow absolute comparisons over time, as we now explain.

The basis for claiming comparability over time is the use of a technique termed test equating. The basic idea is that one has two different tests administered at two different times. The "common item" procedure retains a small number of identical questions, say 20% of the total: a small enough number to avoid detection but large enough to achieve reasonable accuracy. The key 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. It should also be noted that this assumption of a 'common meaning' is one that is used when forming a single scale across countries. 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 which we describe below. 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 at different times; yet it is necessary to assume that this relationship is constant. This second assumption is therefore contestable and very much 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.

The technical foundation for PISA and similar studies lies in the construction of statistical (psychometric) models of some considerable complexity that form the basis for test construction and the provision of scores along the various scales and subscales provided. A major problem with such (item response) models is that they can only be understood by someone having a high level of technical competence and this results in a general level of opacity that tends to preclude criticism. The assumptions that are made by these techniques are important and without going into technicalities, these can be summarised as follows.

The essence of these models is that they assume that, for any scale that is to be reported, say reading comprehension, there is a single underlying 'trait' or dimension that all the relevant test items are representing. In fact, as Goldstein (2004) showed, this assumption can readily be tested and in the particular case analysed there were as many as three separate dimensions reflected by the items. Forcing the test items to 'load' onto a single dimension means that one dimension tends to dominate and some items that have a strong relationship to a minor dimension will tend to be rejected as 'non-fitting'. In this way, the diversity of assessment tends to be narrowed. Likewise, since the modelling is applied across countries, where there is a minority of countries where different dimensions dominate or where individual items produce a different kind of response pattern, this will also tend to bias the instrument against those countries. Thus, before there can be any examination of the items in the final test instruments, subtle biases that result from the test construction process itself are likely to have been built in.

Essentially we see here a reversal of the normal scientific paradigm whereby mathematical models are developed to mirror, as far as possible, the data structure to which they are applied. In PISA and similar studies the data are forced to conform to a simple 1-dimensional structure by eliminating from consideration those parts of the data that are discrepant. Yet it is this assumption of unidimensionality that allows PISA to support its claims about comparability, and it seems that they therefore have little interest in rejecting this. The key issue is that there is no inherent 'objectivity' about the techniques. They rely upon specific assumptions that are contestable.

3. Sampling of schools and students

The normal requirement for PISA is that some 80% of schools provide responses, possibly after replacement of those that fail to take part after selection. Within schools, the standard requirement is that 50% of students respond.

While it is the case that lower response rates are associated with a tendency to biased results, even having a relatively high rate is no guarantee that results will be unbiased in the sense that those not responding can be viewed as effectively 'randomly missing'. To establish whether an obtained sample can reasonably be assumed to be representative of the population of interest, the usual procedure would be to seek to compare key characteristics of the sample with known population data. Thus, in some educational systems there are administrative databases of all school students, and there will also typically be national data from censuses that allows comparisons of socioeconomic status, birth order etc. Also, many educational systems have systematically collected data on school characteristics that would allow comparisons.

PISA appears to impose no requirement on participating countries to undertake such analyses, so that we are left being unable to decide how representative PISA samples really are. In fact, it may well be the case that some countries with relatively low participation rates may produce more representative samples than other countries with higher participation rates. Furthermore, requiring only 50% within-school participation does seem likely to lead to selection bias happening at that level. The rules that allow exclusion of certain students are very loosely specified. For example one of these reads as follows (OECD, 2012, P150):

"Students with insufficient assessment language experience are students who need to meet all of the following criteria:

i) are not native speakers of the assessment language(s

ii) have limited proficiency in the assessment language(s); and

iii) have received less than one year of instruction in the assessment language(s).

Students with insufficient assessment language experience could be excluded."

We see here that there is considerable leeway for schools and survey managers to exclude poorly performing students, this emphasises the need for rigorous comparisons of PISA samples with objective available data to ascertain the extent of biases that may be present.

4. Interpretation and analysis

While OECD itself produces reports with details of country results, it also makes available individual level data in the form of scores that allows secondary data analyses, and educational researchers have made extensive use of these. In the next section I will look at some of the technical problems associated with this, but one issue of interpretation deserves further consideration. Blum et al (2001) in their reanalysis of the IALS data raise the issue of how item responses should be combined to produce individual scores. They point out that item response models, such as the favoured Rasch model, that are in effect equivalent to simple averages of the item correct (1) or incorrect (0) responses, are not the only procedures that can be used. They consider in particular a score based upon the average difficulty level of each item. Out of all the items answered correctly the score for an individual is then given by the highest difficulty level among these items. This form of item scaling is used in other contexts and the authors show that using this measure results in quite a different picture in terms of comparisons. The point is that there is a choice of which scaling procedure to use and it would be both interesting, and arguably more honest, to provide users with such alternatives. At the very least a scientific debate about this issue is needed. Currently it would appear that decisions about scaling are regarded as purely 'technical' and effectively decided by a small group of technical experts.

PISA is intended to look at knowledge and skills for life and is not intended as a study of how far students have `mastered a specific school curriculum' (Kirsch et al., 2002, p13) and it follows that `reading proficiency among the PISA populations cannot be directly related to the reading curriculum and reading instruction' (p. 14). Despite such statements, however, PISA does claim to `monitor the development of national education systems by looking closely at outcomes over time' (Kirsch et al., 2002, p. 13). Throughout the various PISA reports, and in many commentaries, there is the clear assumption that direct comparisons of educational systems are possible. This apparent confusion of aims is further compounded since all the studies are cross-sectional and can therefore say little about the effects of schooling per se. Thus, 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. Yet, the results of these studies, encouraged by OECD itself, are interpreted in terms of simple causality and used by countries to make policy, most notably to advocate curriculum changes.

This causally oriented approach to interpretation continues when the report quotes high (0.8±0.9) correlations between reading, maths and science test scores and uses these data to suggest that `reading is a prerequisite for successful performance in any school subject' (Kirsch et al., 2002, p. 15). This may have some truth, but the existence of high simple correlations does not demonstrate this. It would be more relevant to look, for example, at how progress in maths correlates with progress in reading and we do know from other studies that such correlations are much lower. Thus, in Inner London (Goldstein et al., 1993) the simple correlation between English and Maths at 16 years is 0.62, but only 0.40 in terms of progress between 11 and 16 years.

For reasons that are not entirely clear, OECD appears to have foresworn the implementation of longitudinal follow-up studies as a routine, although individual countries have attempted to do so. Since it is generally accepted that the *minimum* requirement to be able to make any kind of causal statement is through having longitudinal data, the usefulness of PISA results for policy making remains dubious.

5. Data release and ‘plausible values’

The underlying assumption when producing estimates of student ‘abilities’ in each reported domain is that there is a single (latent) dimension of ability and that observed item responses reflect just that dimension. It follows that for each country any given set of items should preserve the same difficulty order, as the technical report (OECD. 2012, P150) puts it

“Item by country interaction: The consistency of item parameters across countries is of particular importance in an international study. If the test measures the same underlying construct (or latent trait) the item should have similar relative difficulty in each country.”

In fact, items that do not satisfy this criterion, country by country, are eliminated and referred to as ‘dodgy’ items. Thus, we see here a process that in practice allows comparisons between countries only after what may in fact be rather interesting interactions, in terms of different relative difficulty, are removed. This is done in order that the favoured Rasch based analyses can be used, and I would suggest is a prime example of the technical tail wagging the substantive dog.

Another issue that is always present when comparisons *over time* are required is that different instruments will typically be used at each time point. PISA lays great emphasis on being able to interpret trends over time and to do this ‘common items’ are used in each survey. The basic assumption is that these can be used as an ‘anchor’ to align all items on a common scale for comparison. The difficulty with this is that the relationship between the common items and the underlying concept being compared may change over time and there is ultimately no way of understanding the possible effect of this *using the test scores themselves*. Thus, despite some sophisticated manipulations carried out by PISA, any interpretations about changes over time remain highly problematical. See Goldstein () for a more detailed discussion.

Because of the way in which item booklets are randomly distributed to students, each student responds to just a subset of items. Such designs can be efficient in terms of content coverage, but they do pose problems for analysis. We can consider each student’s potential responses to the complete set of items or item booklets so that the data available is one where for each student we have some missing data – those items not responded to. Techniques for handling such missing data, where, as here, they can be regarded as missing completely at random as a result of the distribution of booklets, are readily available. PISA adopts the following procedure.

6. Conclusions

In this brief discussion I have attempted to problematise the OECD PISA and other studies in terms of their design, their technical assumptions and the use that is made of their results. The efforts and resources that go into these studies are considerable and there is a serious question about whether they are in any sense cost-effective. The acceptance of the results at face value by most olicymakers in most countries is of concern, not simply because much of the interpretation can be described as bad science, but because by relying on highly constrained evidence it Is not at all clear that rational decisions are being made. In particular, attempts to derive causal inferences from correlations between educational policy and parallel changes in test scores over time are invalid.

There is a role for attempts to make cross country and cross-cultural comparisons, but in the spirit of exploring how these arise and what the role of cultural differences may be. 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'. It 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. Such studies should be treated as opportunities for gaining fundamental knowledge about differences, not as competitions to see who comes top.

Finally I would wish to emphasise that the production of simplistic country rankings, especially when these are justified using a relatively esoteric technical narrative, does not serve the interests of knowledge acquisition and nor does it serve as a basis for educational decision-making. Unless OECD changes its focus so that its studies abide by accepted rules for scientific enquiry, it is difficult to see a good case for the continuation of such studies.

References

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

Blum, A., Goldstein, H. & Guerin-Pace, F. (2001) International adult literacy survey (IALS): an analysis of international comparisons of adult literacy, Assessment in Education, 8(2), 225± 246.

Goldstein, H., Rasbash, J., Yang, M., et al. (1993) A multilevel analysis of school examination results, Oxford Review of Education, 19(4), 425±433.

Goldstein H. 2004. International comparisons of student attainment: some issues arising from the PISA study. *Assessment in Education, Vol.11, No.3, November 2004 pp 319-330*

Kirsch, I., Long, J. D., Lafontaine, D., et al. (2002) Reading for change: performance and engagement across countries (Paris, OECD).

OECD. (2012). Technical report. Paris, OECD. Downloaded, 12/10/2016 from https://www.oecd.org/**pisa**/**pisa**products/**pisa**2012**technicalreport**.htm