

*Plenary talk to Association for Educational Assessment – Europe:
Budapest Nov 4-6, 2004.*

The Education World Cup: international comparisons of student achievement

by

Harvey Goldstein

Institute of Education, University of London

h.goldstein@ioe.ac.uk

Introduction

Brought to you these days by both the Organisation for Economic Co-operation and development (OECD) and the International Association for the Evaluation of Educational Achievement (IEA), comparative studies of student performance are frequent, large, and expensive programmes, promoted using the very best public relations techniques. Governments whose countries do relatively well are quick to claim credit for their educational policies, while the rest try either to ignore the results, or to find a class of people to blame: teachers are a favourite, and occasionally previous governments can be tarred with that responsibility. Opposition politicians, of course, can be expected to indulge in mirror-image reactions.

In this talk I will not dwell too much on the politics of international testing, partly because I am not especially well qualified to do so and partly because the reactions of most politicians seem all too predictable and depressing. Rather, I shall aim to bring together the results of work that I have been involved in since the early 1990's. I will then use this to ask questions about future directions for these studies that, despite certain important contributions they have made to understanding, are still largely seen as entertaining, but ephemeral, spectator sports that rank countries in simple league tables according to the scores of their students. By contrast, I would like to encourage efforts to turn them into instruments whose main purpose is to provide useful data for the improvement of education rather than raw material for governmental propaganda.

The next section describes the present scene in terms of the offerings of OECD and IEA.

The present scene

The OECD has invested heavily in a series of studies known as the 'Programme for international Student Assessment' (PISA). This began in 2000 in some 32 countries. The testing was carried out in the first half of 2000, and this study was intended to be the first of a series. It concentrates on reading but also has Mathematics and Science components. The second study carried out in 2003 concentrates on Mathematics and the third in 2006 will concentrate on Science.

Praise is needed where it is deserved. Considerable efforts have been made to obtain

good response rates and careful attention has been devoted to the design of instruments, and many lessons from previous studies were clearly absorbed. Among these were attempts to include items whose linguistic origin is not English, although the majority of items still do originate in English. The first comprehensive report (OECD, 2001) appeared in 2001 and an extensive (300 page) well documented technical report (Adams and Wu, 2002) provides detail about the procedures used. In addition the data themselves are available for secondary analysis from the OECD web site (www.pisa.oecd.org/pisa/outcome.htm). In addition, OECD is also setting up a new study of adult basic skills that aims to produce results before the end of the decade.

The other major player, increasingly facing stiff competition from OECD, is the IEA. Established in the late 1950s, and apart from a brief attempt to challenge its territory by Educational Testing Services (ETS) in the late 1980s with the IAEP study (Lapointe et al., 1989), it has been the dominant influence in the comparative international testing field. IEA has developed and pioneered much of the current methodology, in terms of sampling, translation, question design, scaling and reporting. Its current major studies are the Progress in International Reading Literacy Study (PIRLS; Mullis et al., 2003) and the Trends in International Mathematics and Science study (TIMSS; Mullis et al., 2001). The former, carried out in 2001 sampled about 4,000 fourth grade students in 35 countries, testing reading literacy and the second round of this study scheduled for 2006 is being planned. The latter sampled similar numbers and began in 1995 (then known as the Third International Mathematics and Science Study), and sampled also in 1999 and 2003. It appears that plans are also being prepared for further rounds of TIMSS.

In the next section I shall look at the organisation of these studies, effectively who controls them and where they get their funds from.

Organisation and funding

The first thing to note is that the documented funding sources represent only one component of the true costs. The value of the time spent by national advisory groups, schools and others is clearly large but nobody seems to have bothered to obtain reliable estimates. Both OECD and IEA expect participating countries to cover the cash costs of their own samples and any individual country analyses; in most cases £1 million would be a minimum figure for participation, but this will depend on individual circumstances. The cash costs and the hidden costs will tend to limit the participation of poorer countries, and will even give wealthier ones pause for thought in the light of competing demands. For both organisations, therefore, a key concern is to retain interest in the studies so that countries will continue to sign up to individual surveys; if too few countries are willing to take part a survey will find it difficult to be economically viable. Hence the stress on public relations and media-friendly presentation, together with a less than dispassionate discussion of methodological problems.

While the OECD, as an established organisation, provides general support for its studies, the IEA has only a very small headquarters secretariat staff with a rotating chairperson and each study needs to raise funds for general as well as specific support. Because of this the IEA has always looked for support elsewhere and receives funds from the US National Center for Educational Statistics, the US National Science Foundation, the World Bank and the United Nations development programme. These funds support individual studies but also more general

developments, for example in methodology. In addition, participating countries provide an annual amount to support the IEA infrastructure.

As I shall argue below, the methodology used is not culturally or politically neutral and we might well expect the cultural and political assumptions of such funding bodies to interact with this methodology in ways that determine its form and content. Likewise, the OECD itself embodies particular political and cultural perspectives that too will shape the methodology and also the practices of studies that it sponsors. In order fully to understand the results from all of these studies, in my view we need to appreciate where they come from, and that will then help us better to evaluate their contribution to knowledge.

For each study or sampling round a set of committees, panels and expert groups is set up to devise protocols, pilot instruments, select contractors, oversee methodology etc. Both organisations set up committees with formal overall decision-making control; for OECD the 'PISA Governing Board' and for IEA its general assembly that meets annually. In practice, however, real control is exerted by a relatively small group with executive responsibilities, whether this is for designing and finalising questions or for processing the data. This control is particularly evident in the case of the scaling and statistical methodology, largely perhaps because of its highly technical nature which is inaccessible to most others, and I shall now look further at this.

Una Tecnica Mafiosa

That subset of the psychometric profession that is concerned with testing, is very much an Anglophone group, centered largely in North America, and to a lesser extent in Australia, the Netherlands and England. Its major theoretical journal is *Psychometrika*, and the major applications journal is the *Journal of Educational Measurement*. While many of its practitioners are spread among academic institutions, much of the literature and the greater part of the practice emanates from commercial organizations. The principal globally active organizations are quite clear about their international missions. Thus we have Educational Testing Service (Princeton, USA – “Centre for Global Assessment”), University of Cambridge Local Examinations Syndicate (UCLES, Cambridge, England – “A World leader in educational assessment”), CITO (Arnhem, Netherlands – “Technical systems for international comparisons”), NFER (Slough, England – “active in international research”), and the Australian Council for Educational Research (ACER, Melbourne, Australia – “A range of services for international clients”). Given the high level of technical expertise required to develop curriculum materials, and especially to construct suitable measuring instruments, it is increasingly the case that only well provided commercial organizations will have the resources to bid for and to undertake this activity. We might also expect that this will influence the decisions that are taken which will reflect, in part, the global interests of such corporations.

While these organizations are undoubtedly in competition with each other, they also share a common approach. Most notably this is expressed in their adherence to certain procedures for designing, scaling and analyzing tests. I have already mentioned the league table approach to comparing countries and the simple-minded ways in which decision makers use these to determine educational policy. Since it is generally these same decision makers who determine the funding, it is not surprising that those responsible for delivering the goods find themselves advocating just those methodologies that appear to support simple interpretations.

Most obviously the favourite technique of this psychometric subset, a group that displays certain characteristics of a mafia¹, is something called the ‘Rasch model’, named after a Danish mathematician George Rasch. This, and certain limited generalizations of it (typically the ‘2-parameter’ model) have come to be known under the title ‘Item response Theory’ (Lord, 1980), although the term ‘theory’ is something of a conceit since it is really just a special case of a statistical model that is widely used by social and other researchers to summarise a wide variety of data. I shall refer to it simply as an item response model (IRM).

The essence of this modeling approach is that it assumes that a student’s responses to a set of test items can be summarized adequately in terms of a single underlying dimension or factor (unidimensionality). This assumption is meant to hold universally, across cultures and educational systems, with group variation exhibited only in terms of average differences along the underlying scale or subscale. The prima facie unreasonableness of this assumption is countered in many subtle ways, all of them sharing the characteristic that they are backed up by technicalities that are either too obscure to be understood by anyone outside the ‘family’, often just inadequate and sometimes both. Thus, for example, a common procedure is to ‘test’ the unidimensionality assumption by applying a technique that, given the limited nature of the data available, is unlikely to be able to refute it – i.e. achieve a statistically significant result (see e.g. Goldstein, 1980). Another favourite procedure is to examine items in a test individually and exclude those that do not seem to ‘fit’ a unidimensional scale. These items are referred to as ‘dodgy’ and happily consigned to the recycling bin. Thus, by eliminating such items the remaining ones are much more likely to conform to a unidimensional scale so that the test developers can point to how their original assumption is then satisfied (see Goldstein, 2004 for a discussion of this in the context of PISA).

When, however, a wider analysis is carried out of tests such as those of PISA we find that the unidimensionality assumption does not hold and there is much more complexity than is typically allowed for (Goldstein et al., 2005). In one of the most far reaching re-analyses of such a study, the International Adult Literacy Survey, also sponsored by OECD, Blum et al (2001) showed that there were all kinds of inherent difficulties associated with cultural differences linked to translation and methods of reporting that excluded the possibility of making definitive international comparisons. In effect, while paying lip service to ideas of cultural diversity, the practice of these studies effectively denies it by imposing a procrustean² ‘one size fits all’ model³.

To all of these issues one might also add a crucial one when comparing educational systems; that without longitudinal performance data on the same sample of students it is quite impossible to make inferences about the effects of educational systems per se that are separable from the influences of social background, economic circumstances

¹ I hasten to point out that I have absolutely no reason to believe that members of this group have any links to organised crime.

² In the Greek myth Procrustes promised his guests that a bed for the night with the extraordinary property that it would exactly match the requirements of their size. As soon as the guest lay down Procrustes set to work; if the guest was too long for the bed he chopped off part of his legs etc. In the legend, interestingly enough, the hero Theseus turned the tables on Procrustes by adjusting him to fit his own bed.

³ One of my most treasured artefacts is a baseball cap adorned with the ETS logo, and inside which the manufacturers tag proclaims that ‘one size fits all’.

etc.

The future

I have little doubt that we shall see more of the same for some time to come. Demands for simple measures to compare countries are very strong and the commercial pressures are closely aligned to such political needs. Of course, this is not to say that some of the data from these studies cannot be used positively to help us understand educational processes, but a vital question is whether the effort that goes into these studies can be justified economically in terms of the knowledge that emerges compared to the resources that are fed in. Most importantly, these well resourced studies consume the energies of many talented researchers, who otherwise might be funded to produce more socially useful products.

My own view is that, in their present incarnations and contemporary sources of support, these studies represent a very poor return on investment. Their lack of a serious longitudinal element, their obsession with simple-minded unidimensional scaling models and the concomitant disregard for contextual diversity and incompatibility, as well as their role as high profile advertising media for certain multinational corporations, make them poor vehicles for serious research. The useful research that does emerge is in spite of this rather than because of these factors.

Since this is a European conference I would like to raise briefly the issue of the future of a specifically European assessment strategy. The Lisbon European Council (2000) clearly envisaged a specifically European approach to education and the evaluation of education systems. Given the procrustean nature of both the actual techniques used as well as the administrative procedures for implementing the OECD and IEA studies, they do seem to be poor models for Europe. A different approach is needed that genuinely respects differences while also promoting convergence in terms of objectives, at whatever level such differences allow convergence to occur. It would not be appropriate or useful to try to impose conformity along the lines of the pseudo-comparability built into the existing international studies. Of course, such a different approach initially will be difficult to formulate and implement, but that simply presents an interesting problem for Europeans, which I am sure they are more than capable of dealing with.

Finally, to keep a debate going, let me issue three challenges to those involved in the OECD and IEA studies.

First, if you wish to dispute any or all of the accusations that have been levelled, I (and others) would be more than happy to have a very public discussion, but only on condition that you do not attempt to obscure the issues by retreating behind technicalities as you have so often done in the past; the crucial issues can be discussed intelligently without this. Secondly, prepare a cost-benefit analysis that weighs the inputs against the contributions to knowledge and understanding in a way that allows a comparison with typical research based studies. Thirdly, see if you can set out a future for international studies that truly learns from the mistakes of the past, reduces dependence on current political and commercial dictates, and makes a really serious attempt to reflect cultural and educational diversity rather than constraining it within the confines of a particularly rigid psychometric model.

I look forward to an interesting debate.

Acknowledgement

I am indebted to Gerard Bonnet both for pointing out factual errors on an early draft and for general suggestions about improving the clarity and content of this paper. Naturally, I take sole responsibility for any remaining inaccuracies as well as for the views contained in this paper.

References

- Adams, R. and Wu, M. (2002). *PISA 2000 technical report*. Paris, OECD.
- Blum, A., Goldstein, H. and Guerin-Pace, F. (2001). International adult literacy survey (IALS): an analysis of international comparisons of adult literacy. *Assessment in Education* **8**: 225-246.
- Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of mathematical and statistical psychology* **33**: 234-246.
- Goldstein, H., Bonnet, G. and Rocher, T. (2005). A study of procedures for the analysis of PISA 2000 reading data. *Submitted for publication*.
- Lapointe, A. E., Mead, N. A. and Phillips, G. W. (1989). *A World of Differences*. Princeton, Educational Testing Service:
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, New Jersey, Lawrence Erlbaum Associates:
- Mullis, I. V. S., Martin, M. O., Gonzalez, E. J. and Kennedy, A. M. (2003). *PIRLS 2001 International report*. Chestnut Hill, Boston College.
- Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., et al. (2001). *TIMSS assessment frameworks and specifications 2003*. Chestnut Hill, Boston College.
- OECD (2001). *Knowledge and Skills for Life: first results from Programme for International Student Assessment*. Paris, OECD.
- Yamamoto, K. and Kulick, E. (2000). Scaling methodology and procedures for the TIMSS mathematics and science scales. *TIMSS 1999 technical report*. M. O. Martin, K. D. Gregory and S. E. Stemler. Chestnut Hill, Mass., Boston College.
- Lisbon European Council (2000).
http://europa.eu.int/ISPO/docs/services/docs/2000/jan-march/doc_00_8_en.html