*Assessment in Education*
*Vol. 11, No. 2, July 2004*

Carfax Publishing
Taylor & Francis Group

## REVIEW ESSAY

# International comparative assessment: how far have we really come?

**Methodological advances in cross-nation surveys of educational achievement**
Andrew Porter and Adam Gamoran (Eds), 2002
Washington DC, National Academy Press
US$66.00 (pbk), 384 pp.
ISBN 0309083338

## Introduction

This volume arose from papers presented at a symposium held by the Board on International Comparative Studies in Education of the US National Research Council, a body set up by the US National Academy of Sciences.

According to the editors the papers in this volume 'represent the most up-to-date and comprehensive assessment of methodological strengths and weaknesses of international comparative studies of student achievement' (p. ix). The volume contains 11 chapters written from a variety of viewpoints, although, as I shall argue, they tend to share common assumptions, which, together, place certain limitations on this claim. The opening chapter by the editors sets the scene and embodies most of these assumptions and I will start by examining this in detail.

## Setting the stage

Porter and Gamoran set out what they see as the purposes of international comparative studies, largely in terms of how these can benefit the education systems of countries, especially those of the USA. They argue that comparisons between the characteristics of high- and low-achieving countries can 'suggest hypotheses of how education in low achieving countries might be improved' (p. 5). They caution that any observed associations cannot be taken to imply cause and effect, but later claim, with approval and without irony, that 'it was largely the international comparative data on student achievement, showing the United States as ranking low among other countries, upon which the (*A nation at risk*) report built its case (p. 5). They are not

ISSN 0969–594X (print)/ISSN 1465–329X (online)/04/020227-08

DOI: 10.1080/0969594042000259510

the only ones to seem to want to have their cake and to eat it (see Reynolds & Farrell, 1996, for a UK example), and one would have hoped for a more incisive critique of official responses to international comparisons. Despite this, however, there is an important recognition from the authors of the need to test any hypotheses that might be generated from comparative studies, within each individual educational system. In particular the authors point to the useful discussions generated about curriculum coverage within countries.

Among the desiderata for comparative studies the authors suggest that a study is 'characterised by research neutrality (e.g. not just a Western perspective)' (p. 7). While this sentiment has some face plausibility 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 is, of course, contestable but nowhere in this volume is it subject to scrutiny. 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).

On the technical side the authors are clearly very much in favour of item response models for producing common scales and subscales along which countries can be compared. They exhibit no doubts about the appropriateness of these methods, and certainly make no attempt to explore the assumptions that underlie them. The authors, however, do point to particular problems that need to be addressed. They raise the issues of curriculum coverage in tests, test format and the need to solve problems arising from different age/grade structures in different systems. They also emphasize the need to understand and measure the overall cultural context within which educational systems operate and suggest that much more attention needs to be paid to this in interpreting data. Likewise, the development of measures of 'opportunity to learn' has been an important feature of comparative research and the authors rightly see this as an important contribution.

One of the most important omissions from both this chapter and many others in the volume (but see Raudenbush & Kim, below) is that of any serious discussion of the need for longitudinal data in comparative research. For example, in their discussion of the need to adjust for background factors when interpreting differences, the authors fail to mention the far more important need to adjust for prior achievement. With the exception of the Second International Maths and Science study, there are few attempts to follow-up students, even over a one-year period. Difficult as this may be, without such longitudinal data the existing research literature indicates that it is impossible properly to attribute any observed differences to the effects of education per se, despite this being a major aim of comparative studies of achievement. Yet those agencies involved in carrying out these studies, principally the Organization for Economic Cooperation and Development (OECD) and the International Association for the Evaluation of Educational Achievement (IEA), continue to ignore this issue and the failure properly to address it in this volume can only serve to perpetuate this serious weakness.

Finally, reading this chapter one is continually reminded that the authors are really interested only in how comparative surveys of achievement can and should influence education in the USA, and I will return to this later.

Linn provides a wide-ranging discussion. He reviews the purposes of the various studies, their designs and their analysis procedures. The account is a careful one that shows a considerable understanding and sympathy for what has been achieved. The sections on different item format issues, such as the use of multiple choice, are succinct but capture the essence of the problems. He stresses the importance of curriculum differences and the importance of measuring curriculum coverage and opportunity to learn (OTL), although there is little discussion of how one might validate the OTL measures. Linn is also concerned with the likely pro-US bias involved that might arise from use of US tests as starting points, but makes no reference to any studies that explicitly investigate this.

When it comes to describing the field-testing, selection and analysis of the test items, Linn resorts to describing the traditional psychometric descriptions and justifications based upon item analysis techniques and item response modelling (usually referred to as Item Response Theory). According to this view item selection is predominantly governed by an underlying assumption of a single 'dimension'—the great advantage of this assumption being that it then enables all kinds of straightforward statistical scaling procedures to be adopted so that countries can be ranked along a single scale. Linn does point to a need to report results in different domains, but somewhat undermines this by talking about the 'usefulness' of single scale summaries. While Linn does bemoan the reliance on a single underlying dimension in one place, he also seems to think of it as a considerable advantage when discussing how items with 'differential functioning' or poor discrimination are eliminated.

Likewise, his description of complex test designs with rotated forms is a good one, but is again unfortunately marred by incorrectly claiming that such designs can satisfactorily be analysed only by using item response models. While he concedes that this is so only if the assumptions of those models are correct, he gives no discussion about the reasonableness of those assumptions, leaving the reader to assume that they are indeed satisfactory.

The next chapter by Hambleton looks at the issue of translational comparability and is a valuable source of information. He starts by making a useful list of the ways in which tests may fail to be comparable: the item formats and contents may be differentially familiar, the translations may be poor, etc. He also looks at some of the debates about cultural differences and political decisions about the nature and need for translation. He suggests that 'test adaptation' is a better term than 'translation' because it may be necessary to change certain aspects of the format, language or administration to make a test 'equally valid' in each country.

The central part of the chapter sets out nine steps involved in test adaptation. The first is 'construct validity' by which is meant the definition of a construct that can be assumed to have the same meaning in different countries. The second step is to decide whether an adaptation is really feasible or whether an attempt to obtain complete comparability should be abandoned. The third and fourth are concerned with employing multiple translators and back-translation to cross-check versions. The fifth is a thorough review of the adapted tests. The sixth and seventh are pilots and field-testing of the adapted test. The eighth step is to find a common scale for all the adaptations and Hambleton recommends item response models for this. The final step is to prepare good documentation. In a final section Hambleton discusses the results of a survey of adaptation procedures used in comparative assessment studies. He points to the considerable progress made over the last forty years or so and the increased understandings now available. He also points out that there is still much to learn and care is needed whenever comparative study results are interpreted.

While Hambleton's account is useful two important things are missing. The first is that he gives us no indication with real examples of where existing comparisons made with the IEA or OECD studies may have to be revised because of adaptational problems. For example, he makes no reference to the evaluation of the International Adult Literacy Survey (IALS), which uncovered a wealth of evidence about the problems of translation and indicated where conclusions were unsafe (Blum *et al.*, 2001). The second issue is that of using the *results* of a study to understand adaptational problems further. As the IALS evaluation found, it is only when real large-scale data become available that some problems emerge. The sense of Hambleton's chapter is that if we can get the prior testing right, including any caveats, then results can be reported without further attention to this issue. Not only is this misleading, it also ignores the important insights into cultural and other differences in understanding that may emerge from the analysis of comparative data.

James Chromy looks at sampling issues and how they arise in the design and analysis of comparative studies and he presents a useful summary of the issues. He reviews the available guidelines for drawing acceptable samples and goes on to a detailed review of the sampling procedures used in some 15 major comparative studies. He traces and summarizes the various critiques made of the sampling designs from the early 1980s and pays particular attention to the recurring issues of how to sample students with respect to age and/or grade of schooling, noting that the different organization of school systems makes this a key issue.

Chromy concludes with a careful discussion of what he sees as outstanding issues. The first is that of population definition, in terms of ages and grades and also exclusions from schooling and home schooling. The second issue is that of obtaining an accurate sampling frame of schools, including the need to update old lists. The third issue is that of sample precision and the fourth that of implementing the sampling. Finally he looks at understanding and dealing with non-response and non-sampling errors.

Bembechat, Jimenez and Boulay explore cultural and cognitive issues in comparative studies. They argue strongly in favour of understanding the local culture and context when carrying out comparative studies, especially that there will be variation within individual nations. Here, context includes belief systems, for example, about the nature of learning and assessment. They contrast Japan and the USA in some detail, citing existing studies by psychologists and anthropologists which extend existing views about relative perceptions of 'effort' versus 'innate ability'. They also point to the increasing concern of comparative studies with non-academic measures, such as student motivations, obtained through questionnaires and interviews, but say little about the results from such studies. They discuss the problems of imposing American and European views about learning on other cultures and stress the importance of understanding differing cultural perspectives when designing and using the results of comparative studies. They argue for more qualitative studies of student, parent and teacher beliefs and for combining quantitative and qualitative data in future studies. They are justifiably critical of the 'horse race' reporting of comparative studies and their chapter is a welcome reminder of what needs to be done in this area.

Buchmann looks at problems of measuring family background. She discusses the importance of measuring family background and the processes that operate to ensure that this influences children's schooling, and presents a history of some of the studies in different countries that have explored the relationships between family background and educational performance. She discusses the measurement of social prestige, family structure, family capital, parental educational background and wealth, with suitable references. She also reviews briefly, but usefully, the literature on the relative importance of family and school factors on educational achievement, pointing out particularly the problems of comparable international measurement of social background and the need to use appropriate multilevel modelling techniques. She goes on to report and analyse the ways in which some of the major comparative studies have measured family background and advocates the use of household possessions as a suitable comparative measure. She argues strongly for the analysis of social differences in educational performance and in particular that this can provide more useful information than simply comparing country means. She concludes with a set of recommendations for the future collection of family data in comparative studies.

LeTendre looks at the issue of measuring and analysing 'cultural' effects. He starts from a viewpoint that educational achievement has to take account of culture, that countries themselves exhibit a range of cultures and that these cultures will generally have different views about what constitutes achievement. He argues strongly in favour of studies that integrate quantitative and qualitative data in order to gain a valid understanding of cultural effects. He goes on to describe qualitative studies of student aspirations and emotions but seems to view quantitative analyses as reductionist and unilluminating.

LeTendre suggests that we need to find a way of jointly analysing qualitative and quantitative information. He is critical of the IEA Third International Mathematics and Science Survey (TIMSS) study which gathered both kinds of information, suggesting that there was in fact little integration—for example, the case studies failed to influence the design of questionnaires. He also suggests that the rich qualitative data collected in TIMSS has largely gone untapped. By contrast the IEA Civic Education Study did promote the interchange of ideas among all the researchers involved. He promotes the idea of triangulation to use each type of data to formulate hypotheses that can be tested on the other. He suggests an iteration between the two kinds of data during the course of any study and believes that researchers from different backgrounds need to find ways of working in genuine partnership. That said, however, the practice of integrating quantitative and qualitative data presents large problems, which LeTendre does not touch on. Thus, while practitioners of each type of study have striven to develop acceptable methodologies we have little idea of how we should combine these because there is no overarching common set of practices and criteria for acceptability. It is questionable whether, in particular situations, we should even attempt to force a combination. Certainly, it is not at all clear that large-scale comparative studies are a good place to start.

Floden looks at the history of attempts to measure 'opportunity to learn' (OTL), or curriculum exposure. From the early days of the IEA surveys there has been an increasing interest in using measures of student exposure to the topics in the tests both to adjust national differences and as of interest in their own right as descriptions of curricula. He discusses the difficulties of capturing exposure, using teacher questionnaires, logs of topics covered and web-based systems. Whatever method is used there remains often considerable unreliability in the measures and this is a particular problem for large-scale comparative surveys that are unable to take detailed records. He makes the important distinction between the intended and delivered curriculum and points to the further difficulty that practice may vary considerably within countries.

The evidence is that OTL is correlated with performance, but attempts to use OTL measures to adjust or explain country differences seem to lead to inconclusive results. One possible explanation, not discussed by Floden, is the low reliability of OTL measures, and another, briefly touched on, is the absence of longitudinal data since one would expect *progress* to be more strongly influenced than single time measures of performance. Another problem, but not discussed by Floden, is that of endogeneity: the exposure that students have may be determined by their prior achievement, so that the relationships observed may partly be explained by students' prior achievement. Having longitudinal data would help to untangle this. These problems with OTL have resulted in relatively few good uses of such data, and until improvements in measurement and study design are made it seems they are really of limited usefulness.

Raudenbush and Kim address some of the statistical analysis issues that arise in comparative studies. They argue that the distributions of performance measures within countries should be studied as well as mean values, and that measures of uncertainty, such as confidence intervals, should be given for all comparisons. They discuss the population definition problem, including the complexity of age and grade distributions in different countries. They usefully supply graphs to illustrate these issues and the discussion is at a technical level that should make it widely accessible. They give an extended discussion of the kinds of inferences it may be legitimate to make by comparing studies (cohorts of students) at a given age and comparing differences by age group, and point out the difficulty of assigning any kind of common scale over time or age. They go on to rehearse arguments about drawing of causal inferences from non-randomized studies. While all of this discussion is useful and does provide a helpful introduction to these topics, the overarching problem remains that of lack of longitudinal data; without such data causal inferences about the effects of educational systems on performance remain extremely limited. The authors do recognize this, although rather in passing, and quite rightly point out that even with longitudinal data some problems do remain. Finally, they suggest that, limited as they are, cross-sectional studies are useful in suggesting *hypotheses* about causal factors. This does, however, seem somewhat optimistic. The history of such studies is that the policy-makers have often made causal inferences that have been counterproductive in terms of misdirecting national policies. Users often do not read the small print, or if they do, they choose to ignore the caveats that are contained there. It would be good if these studies were indeed treated as tentative statements about possible causalities, but that sadly is not the reality.

Smith, himself an ex-government civil servant, directly addresses this issue of making inferences for policy and he is concerned with the USA. He points out that it was not until the 1980s that US policy-makers began to take an interest in comparative studies. Then, the apparent poor performance of US students encouraged emulation of educational practices in countries with higher performances, especially those of the Pacific Rim. He discusses the increasing federal control of education and how this has made it easier to implement general changes. He argues that strong causal inferences are not valid using comparative studies, although he suggests 'weak' inferences are possible. He also suggests that the synthesis of evidence, from comparative studies and other sources can provide useful policy guidelines, but it is not at all clear what this means in reality or how it is to be done in other than a highly subjective manner. Smith goes on to discuss four examples of TIMSS data that have been used in policy-making. He discusses the role of 'best exemplars' as having special appeal, but the discussion is parochial and difficult to follow for anyone without a sound knowledge of US educational politics. The same parochialism is apparent in his discussion of relative grade differences for US students and OTL data. The final section makes some practical suggestions for ways in which policy-makers could be educated to make sensible judgements about results from these studies and offers some views about which of the findings from TIMSS are the most interesting.

The final chapter, by Rowan, attempts to summarize the current situation and suggest future directions. He rehearses many of the points made by the other authors and takes the view that, despite their shortcomings and misuses these studies have been extremely useful for the USA in terms of stimulating debate and have provided scholarly insight. While this may be true, as I have already suggested, there is some doubt about whether these studies overall have had a positive effect: unfortunately a careful discussion of this point is not to be found anywhere in this volume.

Rowan gives examples of where he believes the comparative studies of IEA have provided important insights, for example in terms of the relative strengths of the relationship between socio-economic status (SES) and performance. While this is interesting one might ask whether such studies are the best way to obtain such data. At the very least a case could be made for carrying out studies in different countries, which did not require the complex apparatus aimed at achieving absolute comparability of measures, especially in view of the difficulties of doing this.

Differing, but locally relevant, measures of both SES and performance might provide more useful comparisons of such relationships.

Reading this volume I have certain recurring concerns. The first is the almost exclusive focus on what the comparative studies can do to improve US education. While this in some ways seems natural in a volume sponsored by the US National Research Council, it implies that the contributors think nothing is to be learned from how other countries have used these studies—if indeed they have taken much notice of them at all. By the same token, while particular US experiences may be of some interest elsewhere, I doubt whether much of the writing here will be of real interest in countries with very different cultures and educational systems.

Another concern is with the tendency to accept the status quo of these studies: I have already mentioned a failure to question the implicit assumption of the possibility of 'cultural neutrality' as a goal. While there is discussion of how to improve the sampling, response rates, translations etc., there is little radical thought into how the studies might be improved by introducing substantial longitudinal elements. Likewise there is no serious questioning of the over-simple and potentially misleading item response models that are used in the design and analysis of these studies (Goldstein, 1995), nor is there any real discussion of how to use multilevel modelling techniques creatively in the analysis. There is no breakdown of the costs involved, real and hidden, and whether these can be justified. Similarly, even in a volume ostensibly concerned with 'methodological advances', it is a shame that there is no analysis of the politics of these studies: who funds them, who sets the agenda and especially, how does the dominance exerted by the US and certain other industrialized countries, in terms of technical expertise and other resources, influence the outcomes.

Overall, despite the claim that this is the most 'up-to-date and comprehensive assessment' of the methodology of international comparative studies, I find it disappointing. There are some useful discussions and descriptions, but it has few deep insights and one comes away wondering whether we really do need yet more millions poured into these exercises. Possibly we do, but the present volume doesn't make the case.

**Harvey Goldstein**, Institute of Education, University of London, 20 Bedford Way, London, WC1H 0AL, UK. Email: h.goldstein@ioe.ac.uk

## References

Blum, A., Goldstein, H. & Guérin-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. (1995) Interpreting international comparisons of student achievement (Paris, UNESCO).

Reynolds, D. & Farrell, S. (1996) *Worlds apart?* (London, OFSTED).