Assessment in Education
Vol. 11, No. 3, November 2004

Carfax Publishing
Taylor & Francis Group

International comparisons of student attainment: some issues arising from the PISA study

Harvey Goldstein*
University of London, UK

This paper raises some methodological concerns about the conduct, analysis and interpretation of results from the Programme for International Student Assessment (PISA) study. While in many respects PISA represents an advance on previous attempts at international comparative assessment studies, it retains certain problematic aspects. The article comments on the restricted nature of the data modelling and analysis, and the resulting interpretations. It points to certain features of the results that raise questions about the adequacy of the data and it stresses the failure to introduce a longitudinal component. The paper makes suggestions for ways in which such studies can be improved.

Introduction

Two major organizations are involved in international comparative surveys of achievement. The first is the International Association for the Evaluation of Educational Achievement (IEA), one of whose best-known recent studies is the Third International Mathematics and Science Study (TIMSS). The second is the Organization for Economic Co-operation and Development (OECD), which, for example, carried out the International Adult Literacy Survey (IALS), and the Programme for International Student Assessment (PISA).

All of these studies face the well-known and fundamental problem of ensuring 'comparability' of meaning for test scores across diverse educational systems and cultures. The first step in tackling this issue is usually through attention to issues of translation. From early, somewhat naive, views about the possibility of providing satisfactory translational equivalences, there is now a recognition of the tentative and approximate nature of translated materials and the need to recognize contextually appropriate interpretations. The other major step, with which this paper is mainly

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

ISSN 0969–594X (print)/ISSN 1465–329X (online)/04/030319-12

DOI: 10.1080/0969594042000304618

concerned, is to formulate a psychometric 'model'. Such a model presupposes that differences between systems can be ascribed to variation along 'scales' defined by a technique known as Item Response Theory (IRT)—better described as Item Response Modelling (IRM) since it contains little in the way of substantive *theory*.

A key feature of the IEA and OECD surveys is the lack of any systematic, theoretically sound procedure for evaluating the assumptions made by IRM. I will look at ways of evaluating the adequacy of existing IRM techniques using a new set of more general statistical procedures known as multilevel structural equation modelling (MSEM). These include existing IRM techniques as a special case and allow us to establish whether the data in fact have a structure that is more complex than that allowed for by IRM. The extra complexity comes from two sources. The first is the ability to model more than one dimension of underlying ability or attainment so that individuals (and countries) may vary simultaneously across several dimensions. The second is the recognition that students are grouped within schools and that a proper analysis of school-based data has to involve a multilevel approach to achieve a full understanding and description of student performance and its correlates. These issues will be illustrated through an analysis of the PISA study mathematics items (OECD, 2001).

The PISA study represents a very ambitious and wide-ranging attempt to measure and compare performance of 15-year-olds in 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. The sampling design selected schools as first stage units and sampled 15-year-old pupils within schools with a maximum of 35 students within each school. Extensive piloting of test items and general procedures, including translations, was carried out.

Considerable efforts were made to obtain good response rates and careful attention was devoted to the design of instruments, and many lessons from previous studies were clearly absorbed. The first comprehensive report (OECD, 2001) appeared in 2001 and an extensive (300 page) technical report (Adams & Wu, 2002) provides considerable detail about the procedures used. In addition the data are available for secondary analysis from the OECD web site (www.pisa.oecd.org/pisa/outcome.htm).

The present paper will look also at a recent report (Kirsch *et al.*, 2002) that deals with reading literacy in detail. In particular, the paper identifies problems with the study in terms of its conception, the data analysis, and the interpretations; these aspects will be explored below. In a final section we will look at a re-analysis of the PISA Mathematics data that raises questions about the validity of existing analyses and interpretations.

Issues concerning the main aims of PISA

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,

p. 13) and it follows that 'reading proficiency among the PISA population cannot be so 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 the study is cross-sectional and can therefore 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 report 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.

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.

The core of PISA is an attempt to provide comparability across countries. 'Meaningful measurement requires the instruments to be culturally and linguistically equivalent for all participating groups' (Kirsch *et al.*, 2002, p. 19). In fact PISA utilized item response model analyses to score items and pupils on a series of underlying factors or scales; there were three principal subscales (retrieving information, interpretation and reflection) and an overall proficiency scale, and tables for all are presented. For each scale, as with the International Adult Literacy Survey (IALS) cut points are identified and verbal descriptions of proficiencies associated with these scores are derived. We shall look at the scaling issue in some more detail in a later section when we discuss dimensionality.

In attempting to achieve comparability PISA went to considerable lengths, for example, in terms of starting test item developments in both French and English, as well as the usual procedures involving careful translation and back-translation of questions and test items. In this sense it represents some advance on previous studies, although it makes no reference to some of the more subtle issues that were raised from the re-analysis of IALS (Blum *et al.*, 2001) which demonstrated that in many cases lack of comparability could only be judged after the study had been done in terms of analysing response patterns to reveal cultural specificities. Thus, for example, items translate differently in terms of relative difficulty because of the different cultural contexts, and this is extremely difficult to allow for. PISA makes no real attempt to confront this issue. Again, 'PISA focussed more on the construct of reading literacy and less on the issues of content and familiarity' (Kirsch *et al.*, 2002, p. 20). Yet it is precisely the specificity of familiarity and content that makes comparability so problematic, and this is just not seen as an issue.

There is an interesting comparison between IALS and PISA. The latter took 15 items from the IALS prose literacy scale and incorporated them into its tests. Countries were then compared for these items. The French results are of particular interest since France has a mean score on these items above average and similar to the UK, Switzerland and Sweden. In the original IALS study France in fact withdrew after it emerged that its results were so extremely low that inferences were likely to be unsound. This created considerable controversy and led to a re-analysis of the IALS data (Blum *et al.*, 2001), which confirmed the doubts about the validity of the original IALS results. The PISA data would seem to support this judgement.

Several existing critiques of PISA already exist. Some of these question the nature of the test questions. Thus, for example, Prais (2003) points out that one explanation for the differences between PISA and the IEA studies is that the PISA questions were not designed to reflect curriculum content. As mentioned above, the PISA reporting of results is equivocal about this. Bonnet (2002) also makes this point and discusses the difficulties with translation across diverse systems. A particular concern among some Francophone commentators on PISA (e.g., Romainville, 2002) is the bias that may be induced by the Anglo-Saxon composition of the research, the technical advisors and the origins of the test materials. Sampling problems are also a concern, especially where response rates were low, as in England (Prais, 2003), and the absence of longitudinal data is also commented upon (Fertig, 2003).

Dimensionality and item response scaling

The procedure used for constructing each subscale makes some basic assumptions. PISA relied on the IRM approach to decide whether items 'fitted' their unidimensional models well. A critique of this is given by Blum *et al.* (2001) who point out that this can lead to subtle biases that will depend on the actual countries involved in the study and may 'smooth out' important country differences. PISA makes no reference to the debate over such issues, but simply remarks 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). In the PISA technical report 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.

One implication of this unidimensionality assumption is that even if a second dimension is present but only expresses itself in terms of a few items, such items are more likely to be regarded as 'dodgy' when only a single dimension is allowed in the analysis (Blum *et al.*, 2001). Thus, there is a strong tendency for scales to be constructed that are indeed one dimensional, but possibly at the cost of excluding certain kinds of potentially important information.

While it may be the case that for certain purposes, such as pupil certification, aggregation of scores or subscales into a single scale may be 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.

In PISA the 141 reading items were divided into 3 groups representing the 3 subscales and for each group a one-dimensional item response model (IRM) was fitted. An overall reading proficiency scale was derived using the total set of items in a similar way. The average correlation among the subscales was 0.93, which is very high and makes any separate interpretations difficult. For example, 11 countries share the top 9 places on all the scales, with Mexico and Luxembourg always at the bottom. Also, we are not provided with estimates of these correlations after adjusting for other factors such as school or social background, and in fact no multivariate analyses are reported that would allow such estimates to be made. The method of constructing the subscales may well be crucial here. Items were allocated to subscales by the PISA team and experts with each item belonging to only one subscale.

The standard procedure for fitting several underlying dimensions to a set of measurements is a factor analysis where each measurement, or item in this case, 'loads' on each of the underlying factors and the numerical loadings are measures of the strength of association between the factor and the measurement or item response. Thus the response for each item is 'predicted' by a (weighted) combination of underlying factors. Such analyses are not without their interpretational problems, but they do offer a more realistic structure than the one used by PISA, which is just a factor analysis with a single underlying factor for each subscale, with no exploration of the dimensionality of the full set of items. In a subsequent section we describe a re-analysis that has been carried out for the PISA mathematics items in order to show how such factor-analytic models can provide insight into the data.

Multilevel data analysis

One of the features of PISA is its 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 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).

The PISA study carried out simple two level variance components modelling. A basic variance components model can be written as

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_j + e_{ij}$$
$$u_j \sim N(0, \sigma_u^2), \quad e_{ij} \sim N(0, \sigma_e^2)$$

where the *i* subscript indexes the pupil and the *j* subscript the school. The intercept term is β_0 and β_1 is the coefficient of the predictor variable x_1, where this might be, for example, social class, gender, etc. In general, as in the standard regression model, we may have several such predictors. The term u_j is the 'residual' for school *j* and is assumed to have a Normal distribution with zero mean and variance $\sigma^2{}_e$, and e_{ij} is the 'residual for student *i* in school *j*. The percentage of variance $\sigma^2{}_u/(\sigma^2{}_u + \sigma^2{}_e)$ between schools for each country is quoted (Kirsch *et al.*, 2002, chapter 7). Further details of such models can be found in Goldstein (2003). While the use of multilevel modelling is a welcome innovation in comparative studies, there are several difficulties with this analysis that suggest some general problems with country comparisons.

First of all, 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 now well established (Goldstein, 1997) 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, as quoted in the case of Iceland, nevertheless it may well be much higher for those coming from high and/or low social groups.

The second problem arises when one looks at the results for each country. The percentage of the total variation in test scores explained by schools varies from 8% for Iceland to 67% for Hungary, with many countries at 50% or more. Such high values would normally be associated with highly selective school systems. Yet, these figures are much higher than those suggested by most of the existing literature, and in particular in comparison with the IEA studies for some of the same countries. Thus, for example PISA finds 50% for France, whereas the IEA Third International Maths and Science Study (TIMSS) estimates 25% and only 18% for Hungary. For some countries the figures are comparable, such as the UK being 22% and 19% respectively and this is also consistent with other studies. The PISA reading report attempts to explain this largely by arguing that IEA only sampled, at random, one class per school whereas PISA took pupils from all classes in the relevant year group. This would mean, however, that PISA is estimating a *smaller* quantity than IEA since the IEA estimate combines the between-school and the between-classroom variance components. The report, however, concentrates on comparisons with the IEA Reading Literacy Survey where some countries do have larger IEA values than PISA, but largely ignores the TIMSS comparisons. It should be noted further that the percentage of variation between schools is typically *greater* for Science and Mathematics than for Reading. It is not clear why PISA obtains such unusually large values but it does suggest that there may be serious problems with the data, perhaps connected with the ways in which the tests were administered within countries in different schools. In France, for example, TIMSS sampled only students in 'college' whereas PISA sampled from both 'college' and 'lycée', where the latter tend to have higher attainment levels. Thus one would expect to find greater between-school variation for PISA.

A two-dimensional multilevel binary factor model

For present purposes we have chosen the Mathematics test items for two countries, France and England. Out of a total of some 200 items there are 32 Maths items. The basic model factor model for student responses can be written as follows:

$$\begin{aligned} &f(\pi_{ij}) = \beta_{0i} + \sum_{h=1}^{q} \beta_{hi}\theta_{hj}, \quad \theta \sim MVN(0, \Omega) \\ &y_{ij} \overset{iid}{\sim} Bin\,(1, \pi_{ij}) \end{aligned}$$

Here, π_{ij} is the probability that student j responds correctly to question or item i, and y_{ij} is the actual response (1=correct, 0=incorrect). We assume that the y_{ij} have a binomial distribution with mean π_{ij} and denominator 1. The probability π_{ij} is related to the factor structure via a suitable 'link function' $f(\pi_{ij})$. Item response models generally use the logit link function but here we use the probit link function, which gives very similar results and has certain computational advantages (Goldstein & Browne, 2004). The factor structure, as defined in the second line of (2), specifies that for the j-th individual $f(\pi_{ij})$ is a linear function of q factors with values θ_{hj} (h=1,..., q) and β_{hi} and is the coefficient (or loading) of the h-th factor for the i-th item. We can include further predictors such as gender, country, etc. Finally, the factors are assumed to follow a multivariate Normal distribution.

In PISA several questions are grouped in that they all relate to the same problem. For example, one problem described a pattern of trees planted as a set of squares of different sizes, and associated with this problem there were three separate questions. It is doubtful whether the independence (*iid*) assumption would hold for those composite questions and for simplicity we have selected 15 items, each of which is a response to a different problem and dichotomized into correct/incorrect, treating part-correct answers as correct. Model (2) is also extended to the multilevel case

Table 1. Separate country analyses with probit link function model for each item

Columns show English mean scores on the probit scale and the French—English difference between means. Standard errors in brackets. 10,000 MCMC iterations with default priors. The type of item is shown by each item name (MC=multiple choice; FR=free response)

Item	England	France-England
Student level		
33q01 (MC)	0.80	−0.06 (0.05)
34q01 (FR)	−0.25	0.03 (0.06)
37q01 (FR)	0.65	−0.11 (0.07)
124q01 (FR)	0.01	−0.18 (0.07)
136q01 (FR)	−0.23	0.69 (0.05)
144q01 (FR)	0.16	0.40 (0.05)
145q01 (FR)	0.65	−0.13 (0.06)
150q01 (FR)	0.78	−0.35 (0.06)
155q01 (FR)	0.54	0.27 (0.06)
159q01 (MC)	0.89	−0.24 (0.06)
161q01 (MC)	0.96	−0.70 (0.06)
179q01 (FR)	−0.11	0.64 (0.06)
192q01 (MC)	−0.28	0.07 (0.06)
266q01 (MC)	−0.75	−0.26 (0.06)
273q01 (MC)	−0.04	0.03 (0.06)

where we allow a factor structure at the school level also. Full details of the model and analysis are given by Goldstein and Browne (2004).

Table 1 shows the results of a separate probit model fitted for each item for each country. The probit scale means are given for England and for the France-England difference.

Of the 10 statistically significant item differences, France does better on 4 (all free response items) and worse on 6 (3 free response and 3 multiple choice items) than England. One interpretation of the probit function is that it predicts a value from an underlying standard Normal distribution with mean zero and standard deviation 1. This can be converted to a probability using the cumulative density function of the standard Normal distribution. Thus, for example, the French students are, on average, 0.7 standard deviations ahead of the English for item 136Q01 (a free response Geometry item) but 0.7 standard deviations behind on item 161Q01 (a multiple choice Geometry item). This suggests that the item format may be an important feature of country differences related to curriculum and teaching, a feature which should not therefore be ignored. This is reinforced in Table 2, which fits a two factor model at the student level and a single school level factor.

The first student level factor is a general factor and the second tends to distinguish between the multiple choice and free response items. The school level factor is a general factor.

Essentially the two-factor model can be regarded as summarizing the interaction between country and item type. In terms of comparing countries we have two

Table 2. Loadings for two orthogonal factors at level 1 and one factor at level 2

First loading of factor 2 constrained to zero. Variances constrained to one.

Item	Factor 1	Factor 2
Student level		
33q01	0.51 (0.06)	0
34q01	0.67 (0.05)	0.22 (0.09)
37q01	0.81 (0.10)	0.42 (0.14)
124q01	0.56 (0.11)	0.80 (0.21)
136q01	0.60 (0.09)	0.47 (0.12)
144q01	0.58 (0.10)	0.08 (0.10)
145q01	0.57 (0.06)	0.19 (0.12)
150q01	0.72 (0.10)	−0.07 (0.18)
155q01	0.44 (0.06)	0.28 (0.10)
159q01	0.50 (0.06)	−0.04 (0.12)
161q01	0.43 (0.07)	−0.27 (0.14)
179q01	0.46 (0.08)	0.46 (0.17)
192q01	0.62 (0.06)	0.28 (0.10)
266q01	0.41 (0.06)	−0.10 (0.09)
273q01	0.42 (0.06)	0.21 (0.12)
School level		
33q01	0.27 (0.03)	
34q01	0.39 (0.04)	
37q01	0.76 (0.06)	
124q01	0.82 (0.10)	
136q01	0.52 (0.05)	
144q01	0.32 (0.04)	
145q01	0.47 (0.04)	
150q01	0.45 (0.05)	
155q01	0.31 (0.04)	
159q01	0.36 (0.04)	
161q01	0.25 (0.04)	
179q01	0.46 (0.05)	
192q01	0.44 (0.04)	
266q01	0.34 (0.04)	
273q01	0.36 (0.03)	

dimensions along which this needs to be reported. In our model these two factors are uncorrelated, although it is possible to fit a model that allows a non-zero correlation. By contrast PISA uses the equivalent of the first factor only in a single level model and for comparison Table 3 shows the loadings from fitting this model.

Note that the ordering of loadings, and hence interpretation, changes when compared with the first factor for Table 2. Thus, for example, free response item 124q01 has a high loading (0.94) for the 1-factor model but only a moderate loading (0.56) for the first general factor in the 2-factor model. By contrast, for the full model

Table 3. Loadings for 1 factor fitting a single level probit link factor model

Item	Loading
33q01 (MC)	0.53
34q01 (FR)	0.79
37q01 (FR)	1.15
124q01 (FR)	0.94
136q01 (FR)	0.73
144q01 (FR)	0.60
145q01 (FR)	0.77
150q01 (FR)	0.66
155q01 (FR)	0.57
159q01 (MC)	0.57
161q01 (MC)	0.35
179q01 (FR)	0.65
192q01 (MC)	0.80
266q01 (MC)	0.47
273q01 (MC)	0.60

it has a high second factor loading and a high school level loading. Thus it seems difficult to justify the use of such a single factor model, even as a first approximation to the more complex 2-level 2-factor structure.

We have presented this analysis to illustrate the restrictive nature of the PISA model rather than to come to any substantive conclusions on the basis of a limited number of items. Nevertheless, our model does show the importance and usefulness of a more realistic set of assumptions about how pupils respond to test items.

Conclusions

The results from the present PISA (2000) survey do represent an advance in many ways on what has gone before, especially in terms of incorporating a wide range of views about how to assess reading and certain aspects of translation, and this is welcome. There is also a considerable concern with sampling procedures and quality control. Yet there remain serious reservations about PISA.

Perhaps the major one centres around the narrowness of its focus, which remains concerned, even fixated, with the psychometric properties of a restricted class of conceptually simplistic models. There is almost no reference to debates about the appropriateness of these models, nor is there reference to methodological and substantive critiques such as those of IALS. It needs to be recognized that the reality of comparing countries is a complex multidimensional issue, well beyond the somewhat ineffectual attempt by PISA to produce subscales. With such a recognition, however, it becomes difficult to promote the simple country rankings which appear to be what are demanded by policy-makers. As we have illustrated from our re-analysis of the Mathematics items, the pattern of item responses varies across (our two) countries and this essentially precludes any comparison based upon a single scale.

This is especially striking when one recalls that the process of item elimination in PISA will have tended to produce a unidimensional scale. Yet, so long as simple, unidimensional, comparisons continue to be seen as the principal product of these surveys, so their usefulness must remain in doubt and their value for money somewhat questionable.

All of this raises the question of how international comparative surveys should be conducted, analysed and interpreted. This is a complex issue and beyond the remit of the present paper to discuss in detail. Nevertheless, there are certain basic requirements that any such survey should adhere to. First, it is important that cultural specificity is recognized in terms of test question development and 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 by PISA 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.

Finally, any such survey should be viewed primarily not as a vehicle for ranking countries, 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'. 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.

Acknowledgements

I am most grateful to Gerard Bonnet, Sig Prais, Thierry Rocher, Jason Tarsh and Alison Wolf for helpful comments.

Notes on contributor

Harvey Goldstein is Professor of Statistical Methods at the Institute of Education, University of London. His principal scholarly interests are in the methodology of multilevel modelling and educational assessment.

References

Adams, R. & Wu, M. (2002) *PISA 2000 technical report* (Paris, OECD).

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.

Bonnet, G. (2002) Reflections in a critical eye: on the pitfalls of international assessment, *Assessment in Education,* 9(3), 387–400.

Fertig, M. (2003) *Who is to blame? The determinants of German students' achievement in the PISA 2000 study*. Bonn, Institute for the Study of Labour Available online at: http://www.iza.org/publications/dps/

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. (1997) Methods in school effectiveness research, *School Effectiveness and School Improvement,* 8, 369–395.

Goldstein, H. (2003) *Multilevel statistical models* (London, Arnold).

Goldstein, H. & Browne, W. (2004) Multilevel factor analysis models for continuous and discrete data, in A. Olivares (Ed.), *Psychometrics: a festschrift for Roderick P. McDonald* (Mahwah, NJ, Lawrence Erlbaum).

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

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

OECD (2001) *Knowledge and skills for life: first results from Programme for International Student Assessment* (Paris, OECD).

Prais, S. (2003) Cautions on OECD's recent educational survey (PISA), *Oxford Review of Education,* 29(2), 139–163.

Romainville, M. (2002) On the appropriate use of PISA, *La Revue Nouvelle, March–April.*

Snijders, T. & Bosker, R. (1999) *Multilevel analysis* (London, Sage).