Interpreting international comparisons of student achievement

Harvey Goldstein

UNESCO Publishing

The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

Published in 1995 by the United Nations Educational,
Scientific and Cultural Organization,
7 Place de Fontenoy, 75352 Paris 07 SP
Typeset and printed by UNESCO

ISBN 92-3-103082-5
French edition: 92-3-203082-9

Printed in France

Preface

This publication contains a report that was originally commisioned by UNESCO from Professor Harvey Goldstein of the Institute of Education, University of London, as a background paper for the *World Education Report 1993*. It also contains a critique of Professor Goldstein's report submitted by Dr Geoff Masters, chairman of the Technical Advisory Committee of the International Association for the Evaluation of Educational Achievement (IEA), as well as a rejoinder by Professor Goldstein to that critique.

Professor Goldstein was commissioned to report on the technical isues involved in interpreting international comparisons of students' learning achievement. His report was circulated to specialists for comment. Dr Masters submitted a paper embodying his comments and those of fellow members of the technical advisory committee of the IEA. In publishing Professor Goldstein's report, Dr Masters' paper and Professor Goldstein's rejoinder, the Organization considers that the issues raised merit the attention of a wider audience interested in national and international comparisons of students' learning achievement. UNESCO is grateful to Professor Goldstein and to Dr Masters and his colleagues in IEA for their cooperation.

It should be noted that the opinions expressed in this publication are those of the authors and do not necessarily represent those of UNESCO.

Contents

Acknowledgements

I am extremely grateful to the following individuals who have helped by supplying material and commenting on drafts of this paper. Peter Allerup, Alister Cumming, Derek Foxman, Gilla Hanna, Keith Lewin, Angela Little, Bill Loxley, Les McLean, Desmond Nuttall, Neville Postlethwaite, Andreas Schleicher, Elana Shohamy, John Smyth, Merrill Swain, Alison Wolf and Richard Wolfe.

Harvey Goldstein

Interpreting international comparisons of student achievement

Harvey Goldstein

Introduction

This report discusses some of the key technical procedures which have underpinned international comparisons of educational achievement, namely those concerned with sampling and population definition, translation, scaling and statistical modelling. The report is not intended to provide a detailed summary of the findings of comparative studies. It is concerned with the ways in which any such findings can be interpreted and will attempt to draw lessons from existing studies in order to make recommendations for the future.

It is clear that there are political constraints on international comparative studies. Such constraints are a source both of strength and weakness. They are useful in so far as governmental funding and support for these studies tends to ensure ready access to educational institutions and policy discussions; they are a drawback when they dictate a narrow view about which comparisons are important and how findings should be presented. This report does not deal directly with such political issues, but it should be appreciated that many of its conclusions may have political as well as scientific implications for the conduct of these studies.

The first three sections present a brief history of international comparisons and the organizations involved, and a summary of the measurements which have been made.

Historical background

There have been a number of small-scale, limited and usually informally structured comparisons of achievement among countries. However, there are only a few studies which merit serious attention, namely those carried out under the auspices of the International Association for the Evaluation of Educational Achievement (IEA) and the International Assessment of Educational Progress (IAEP). These are summarized in Table 1. IEA is an international non-governmental organization; the IAEP studies are basically international replications of the United States' National Assessment of Educational Progress (NAEP) programme.

TABLE 1. Major international comparative studies of educational achievement

Year	Sponsor collection	Age of pupils	Curriculum topic (No. of countries)
1960	IEA	13	Mathematics, science, reading comprehension, geography, non-verbal reasoning (12)
1960	IEA	13, FS	Mathematics (12)
1970–72	IEA	10, 14, FS	Science (19), reading comprehension (15), literature (10), foreign languages (French and English) (18), civic education(10)
1981–84	IEA	9–15	Classroom environment (mathematics, science, history) (10)
1982–83	IEA	13, FS	Mathematics (20)
1984	IEA	10, 14, FS	Science (24)
1984–85	IEA	10, 14–16, FS	Written composition (14)
1988	IAEP	13	Mathematics (6), science (6)
1988–92	IEA	10, 13	Computers in education (23)
1988–95	IEA	3–5	Pre-primary education (14)
1990–91	IAEP	9,13	Mathematics (20), science (20)
1991	IEA	9,14	Reading literacy (31)
1993–98	IEA	9, 13, FS	Mathematics, science (40–50)
1995	IEA		Second language

FS: Final year of secondary education (differs among countries).

It is apparent from Table 1 that there has been an increasing country participation rate from the first study in 1960, coupled with an increasing frequency of studies in the late 1980s and early 1990s. It is also clear that the most common curriculum areas covered are science and mathematics. The most popular ages surveyed are 9–10 years and 13–14 years. Besides testing learning achievement, most of the studies have gathered information on curriculum, organization, teacher experience, and selected characteristics of students and schools. With some exceptions, such as the IEA second mathematics study, there is little reliable background extra-institutional information about the characteristics of students' parents, home amenities, etc. This limits the kinds of causal explanations which can be offered.

Organizations

The organization which dominates these international comparative studies is the International Association for the Evaluation of Educational Achievement. IEA consists of a set of member institutions (upwards of fifty), usually one from each education system,[1] which send representatives to an annual general assembly, the decision-making body in which every system has a single vote. It has a small permanent secretariat, based in The Hague, an elected chairperson and a standing (executive) committee. In addition, it has a technical advisory committee whose members do not necessarily belong to the member institutions. For each project (study) an international steering committee is set up together with an international co-ordinating centre whose head is executive director of the project and is responsible for ensuring that the data collection instruments are prepared and the data processed into a form suitable for international and national analyses. In each participating country there is a national project co-ordinator who usually belongs to that country's member institution.

Each member institution of IEA is expected to have links with the country's policy-makers and to have research expertise and access to schools. The funding to enable a country to participate in a study almost always comes from central government sources. It is principally in this way that governments make their interests known. Very often, however, the costs of funding pilot or feasibility projects is found from private foundations or particular governments, typically from Western Europe, Australia, North America or Japan. A description of the IEA structure and studies is given by Hayes (1991). The importance of government funding means that some of the poorest

[1] In some countries, such as Belgium and Canada, two separate education systems are operated along cultural lines.

countries are unable to participate because government resources are too limited.

The International Assessment of Educational Progress (IAEP) has been organized by Educational Testing Service (ETS), the major testing service in the United States, building upon its experience of managing the National Assessment of Educational Progress (NAEP) programme. There have been just two IAEP surveys, almost entirely concerned with science and mathematics. Unlike IEA which is fundamentally a democratically structured organization, the principal decision-making functions of IAEP are located within ETS, which also assumes responsibility for the major analyses, although some participating countries carry out their own. There has been no announcement of plans for future studies.

Instrument-development processes

The process of developing instruments for the collection of achievement data and other characteristics of students, teachers and schools is clearly of the greatest importance. I shall describe in turn the various types of instruments and the manner of their development. This section will mainly draw on IEA experience since this is far more extensive and better documented than that of IAEP. Further details of some of the technical procedures can be found in Keeves (1992*c*).

Student and parental characteristics

Much of the information about students is obtained from questionnaires which, for example, ask about amount of homework undertaken or attitudes of parents to the student's studying. There are standard procedures for the general quality control of questionnaires as follows.

First, it is necessary to formulate objectives and then to start the process of translating these into questions or scale items. In an international study this should involve representatives of several countries in the pilot stage, with approval and modifications being sought as to the general applicability of the tasks to be placed before students. Fundamental to this process is that of language translation, discussed later (see p. 16). Nevertheless, even though questions may be correctly translated, their interpretation may differ. Even apparently 'hard' data such as the number of older children in a student's family will depend on the interpretation of 'family', for example whether it includes older 'adult' children not living with the student or children of a previous parental partnership. Such problems will tend to assume greater significance in a self-completion questionnaire addressed to students than in the more common interview survey. Often, it will be at the analysis stage that such problems are revealed. It would assist interpretations of findings if these problems were documented so that an attempt could be made to understand the problems of comparability in questionnaire information.

Although the emphasis in international studies is on comparability of information so that a common interpretation can be made, there are some variables where this may not be possible, or desirable. Perhaps the most important case is that of social status, where a common measure applicable to all kinds of economic systems and cultures will be unsatisfactory so that each country will be obliged to form its own most appropriate one – although there is a great deal of debate within countries about how this might be done. The important issue here, in terms of data analysis, is the extent to which an appropriate measure of social status is associated with achievement. Comparisons between countries might then be made in terms of the relative strength of such relationships.

Where international surveys are part of a sequence, the need for maintaining comparability of questions over time is important, but particular difficulties when new countries join a sequence of surveys may make modifications necessary.

The emphasis on comparability across countries and time is one that seems to underlie almost all the activity of international surveys, and is derived largely from the early desire to make unambiguous comparisons. I shall be returning to this point later in the discussion of interpretation, but for now simply note that in the design of questionnaires, strict comparability often may be unattainable; however, this does not rule out the possibility of useful analyses.

School, classroom and curriculum measures

The IEA classroom environment study (1981–84) was unique in attempting to measure classroom processes directly by observing and coding student and teacher activities over time. Such information can provide very valuable contextual data for comparative analyses, but requires careful training of observers and is relatively expensive.

More usually, information about the schools, teachers and curriculum policies is obtained through questionnaires. In addition, there will be information about school type, size, etc., and in some systems with prescribed curricula the 'intended' curriculum structure.

IEA has always shown interest in descriptions of the curriculum in different countries both for their intrinsic interest and to help in contextualizing achievement results (see for example, Travers and Westbury, 1989; Finegold and Mackeracher, 1986). The United States Government is funding a Survey of Mathematical and Science Opportunity (SMSO) to attempt to develop data collection strategies for obtaining 'Opportunity to Learn' (OTL) data, that is, information on the extent to which the IEA assessment item topics have been covered in classrooms.

In evaluating the performance of any group of students, it is important to know what their curriculum exposure has been, and this will be discussed in a later section. In addition, information about the extent to which different groups of students experience different topics and how this relates to overall curriculum goals is of great interest. The current concern of IEA with this issue, therefore, is welcome. Nevertheless, there are several outstanding problems to be tackled. For example, the reliability of information from questionnaires is not only likely to be low, but also to vary from education system to education system. What works reasonably well in systems with clearly described curricula may work badly in systems with more decentralized and informal curricula. More seriously, OTL data are normally collected on a class or group basis, whereas there may be variation in exposure from student to student within a group.

In the important area of curriculum description there is interest in information on the intended curriculum, namely that which is described in official, local or central documents. The implemented curriculum is that which is provided in schools by teachers and texts which interpret or modify the intended curriculum. The relationship between these two is in general not well understood and OTL data in the IEA surveys provide an important opportunity for investigating it. Finally, there is the attained curriculum, which can be loosely described as that which has been absorbed by students. It is this which the assessments themselves are attempting to measure.

In approaching the comparisons of curricula, there are considerable complexities, since each education system embodies its own cultural assumptions which interact with documentary descriptions and classroom practice. To understand curriculum differences, it is also necessary to understand the cultural contexts. Leung (1992), in a detailed comparison of the mathematics curricula in China, Hong Kong and England, describes in detail how cultural assumptions, transmitted via teachers and others, can affect the implementation of a curriculum.

From teachers, information can be obtained about qualifications, experience and attitudes. From schools, information can be obtained about organization, student grouping, resources available, relationships with parents, staffing, etc. Naturally, much of the information required about the class and school contexts is difficult to measure precisely using questionnaires, especially since much of the information is retrospective. Furthermore, as with curriculum information, this information needs to be contextualized within cultural settings.

Sampling procedures and population definitions

From Table 1 it is clear that there are certain favoured ages. There appears to be a lack of studies focusing on the early years of schooling, between the ages of 5 and 8, and likewise for ages 11 and 12 which correspond to institutional transition ages in many countries. In fact, the populations are defined in terms either of school year or grade. This creates difficulties for comparisons, since school and country policies vary with respect to grade or year promotions; in some countries the whole year group moves together, whereas in others some students repeat years or grades. This problem will be discussed in more detail later (see p. 16).

Most of the existing studies have concentrated upon cross-sectional comparisons, that is, studying students at a single time. There is a deficit of long-term longitudinal studies which could shed important light on the factors associated with student progress and possible causal mechanisms. IEA, for example, in the second mathematics study followed up a subsample of students over a one-year period in a small number of countries, but such short-term studies are of limited value when the curriculum, and programmes of study in general, are designed to cover longer periods.

Within individual countries (see Mednick and Baert, 1981) successful longitudinal studies have been carried out and there would seem to be good reasons for IEA to attempt the same, although such studies do require considerable resources. While successive surveys of literacy or science may provide interesting snapshots, the scope for making causal inferences is severely limited. Longitudinal studies can begin to answer questions about student mobility and its causes, changing performance differences between groups as they progress through the system, and many other issues of considerable significance for educational policy and theory.

The sampling procedures adopted by IEA and IAEP involve, on the whole, standard applications of sample survey methodology. The primary sampling unit usually is the school, and schools typically are stratified, for example by type, region and size. In international studies it is often difficult to ensure uniformity of sample design across countries, so different weighting procedures may be necessary prior to comparative analyses.

Particular problems can arise when sampling students in the final year of secondary education. If the population of interest is those students in school (or in other educational institutions), then there are no novel problems. For some purposes, however, the population definition will be wider, for example to include young people in training activities or even a whole cohort of a particular age. The sampling then has to encompass these groups outside institutions, and becomes more difficult and expensive. One solution to this problem is to define an age cohort of interest and to identify the sample individuals while they are still at institutions, that is prior to the age at which compulsory education ends. Such a sample would then be followed-up to the age of interest. In addition to the sampling issue, one advantage of such a procedure is that longitudinal information becomes available which might be expected to be of considerable interest. The principal disadvantage is that it requires a longer time-span. Such a scheme does not seem to have been adopted for international comparisons, but seems well worth exploring.

Response rates

The response rates – for both schools and students – vary from country to country and age to age. For example, in the second IAEP mathematics study, at age 9 the overall student response rate varied from 53 per cent in England to 99 per cent in Taiwan and at 13 years from 47 per cent in England to 98 per cent in Taiwan. In the second IEA science study the student response rate for 10-year-olds varied from 54 per cent in Norway to 99 per cent in Japan and the Republic of Korea, and for 14-year-olds from 53 per cent in England to 100 per cent in the Republic of Korea. Because lower response rates are generally associated with increased bias, it is important that any comparisons between countries with such different response rates are treated cautiously. This is underlined in the IAEP surveys where countries with low response rates or with restricted sampling frames, such as those covering largely urban areas, are listed separately. The IEA summary science report, however (Keeves, 1992*a*), pays scant regard to this issue, although it mentions that there are comparability problems 'which make it difficult to compare the performance of students'. For example, in comparisons between the first and second IEA science studies, the response rate among 10-year-olds rose from 49 per cent to 84 per cent in Italy, and the response rate for 14-year-olds in Sweden dropped from 91 per cent to 50 per cent. In spite of this, the report goes on to make comparisons without attempting to allow for these problems.

The problem of non-response in surveys is a difficult one, and although it is recognized by those designing comparative studies, it is still an area where such studies are weak. In some cases it may be possible to measure certain characteristics of the non-responding schools and students, and to use these as a check on the obtained sample. The use of callbacks with requests for basic information from schools is worth while, as is the use of nationally available information on pupil/teacher ratios, teacher qualifications, examination results, etc. Alongside other estimates of statistical uncertainty (see below) comparative tables should provide estimates of possible non-response biases.

Age and grade sampling

The IAEP samples were defined by the year of birth of the students. Thus, for example, for the 1990–91 survey the target population was all children born in 1977 and for most countries these were measured in March 1991. This yielded an age-range of one year with a similar distribution of ages within countries. For a small number of analyses, results have been reported separately for the two principal grades into which the students fell. In the second IEA science study the actual mean ages of the country samples for the 14-year-old students ranged from 13.9 years to 15.1 years (with an outlier at 16.0 years). In this study, most of the country samples were from Grade 8 or 9, and in some cases the mean age of some Grade 8 students was higher than the mean age of some Grade 9 students.

Both length of time in school, which is what grade level is intended to measure, and age itself will influence achievement, attitudes, etc. Furthermore, different systems have different policies about whether weak students should repeat grades. Comparative analyses need to take careful account of these problems, ideally adjusting results for age, grade level and the extent to which students have experienced grade repetition or promotion. Information on grade-allocation policies is of interest as a possible explanation for country differences, and this also raises a number of interesting issues about 'compositional' effects, namely how individual achievement is affected by the characteristics of the other students in the same class. This in turn requires the use of 'multilevel' statistical modelling which is discussed below (see p. 25).

To date, little attempt seems to have been made to develop reliable procedures for simultaneous age and grade standardization of information prior to reporting, and most published reports, unfortunately, do not appear to regard it as a serious problem. There are, of course, difficulties in making proper adjustments for both age and grade level (McDonald, 1992) and this is an important area for further research. A recent report on the reading literacy study (Elley, 1992, Appendix E) did carry out some limited age adjustments and showed how this affected some of the country comparisons. Unfortunately, age adjustments were not used for the comparisons in the body of the report. Fortunately, the third international mathematics and science study has proposed that students in two adjacent grades are sampled, so allowing exploration of combined age and grade effects.

Translation procedures

In IEA as well as IAEP, the first versions of all instruments are usually in English, although not necessarily devised solely by native English speakers. Once the materials have been piloted in each country, this acts as a further check on accuracy as well as acceptability and relevance.

The problems of translation of questionnaires and tests have been studied and discussed by a number of researchers, some in the context of the IEA studies (see for example, Brislin, 1970; Little, 1978; Purves, 1992). A number of guidelines have been evolved, along the following lines.

A basic requirement when translating from a source to a target language is to back-translate the text from the target to the source language and to compare the original with the back-translated version. Bilingual translators are commonly used to do this and in some cases complex experimental designs with several translators have been utilized to study the effects of factors such as textual content, translators' experience and familiarity with subject-matter. It seems clear that all these factors can influence the quality of a translation. Moreover, even where there is a good match between the original and back-translated version, the target version will not necessarily be an appropriate translation. This might occur, for example, because a single source word can have several translations in the target language, each of which would be back-translated into the original source word, yet each target language word can nevertheless have a somewhat different meaning.

An interesting example occurs with Japanese which has context-specific number systems. A number from an English designed test could have various target translations depending on the context, yet all be translated back into the same English number. Another example is given by Little (1978) and concerns the use of the word 'expect' when asking students about their future careers. In English there is a difference between 'expect' in the sense of 'wish to' and 'expect' in the sense of 'predict'; that is between hopes and predictions. Some other languages do not distinguish these meanings, partly it seems because the social and cultural conditions make such a distinction unnecessary when ambitions are strongly determined by practical realities. In both these examples, although the goal of exact translatability is unreachable the language differences lead to substantively interesting questions, and I shall return to this point shortly. It is worth mentioning, however, that even where unique one-to-one equivalences between all translated words or phrases is feasible this may be achievable only at the expense of eliminating useful test items or questions.

It is generally agreed that passive constructions, pronouns and complex structures should be avoided. There seems to be little research, however, on the effect of such an injunction upon meaning in languages other than English. For example, simple structures in English do not necessarily carry over into simple structures in other languages, especially pictographic languages such as Chinese. Translators are also well aware that within countries, there are usually dialects, some of which may be unfamiliar yet important if representative population groups are to be sampled. It is also difficult to equate levels of concreteness and abstraction in two different languages. Les McLean (personal communication) quotes the example of a French translation of a mathematics test which satisfied strict quality controls, but was unable to deal with items which were judged to be more abstract than their English counterparts. Hanna (1993) describes a study using six bilingual French/English educators who performed a content analysis of 174 items in the second IEA international mathematics study, all of which had been back-translated. They reported that seventy of the items were found to differ in significant aspects in the two languages. The examples she quotes suggest that many of the differences are potentially avoidable, but to do this would require considerable resources to implement on a large scale.

In the practical situation of a study, operating under constraints of time and resources, it is difficult to take account of all the problems associated with translations. In some cases (Rosier, 1987) lack of resources has prevented some countries even producing back-translations. Among other things, such practical issues imply that the analysis of comparative studies needs to be sensitive to potential translation biases. This will be especially so where large or unexpected differences occur and translation problems need to be eliminated as explanations. For example, it is conceivable that the context-specific nature of a Japanese translation of numerical information may contain information which facilitates a correct response. This strengthens the case for complete documentation of all the study materials, administration instructions, etc., and public access to these.

The position of English as the source language for most comparative studies raises some special issues. One of these, the assumption about simple structures, has already been mentioned. There also are other concerns. Not only is English the main source language for the instruments, it is also the common language of discourse among those jointly designing, discussing and analysing the studies. In IAEP, the fact that it was organized by ETS implies an inevitable dominance of the concerns and cultural values of particular groups in one country. Yet even in IEA, with its more democratically multinational structures, the requirement for country representatives to have a working knowledge of English in order to take part in joint discussions necessarily implies a similar, if not so pronounced, bias.

It is, of course, difficult to quantify the extent of such biases. The English-speaking psychometric tradition is so universally dominant that its assumption as a starting-point for discussions about educational measurement is usually simply taken for granted. Nevertheless, it is possible to carry out research which would throw light on this issue. Languages other than English could be chosen as starting-points for test and questionnaire development, and the resulting instruments used alongside the English-originated ones.

Psychometric approaches to translation

Recently, there have been suggestions that there are psychometric 'solutions' to judging the effectiveness of translations (Hulin, 1987; Hambleton, 1992). In essence these authors propose the following basic psychometric model.

The test item patterns from random samples in the source and target populations are compared to see if they are similar. For example, if the items are ranked in order of difficulty, based upon the proportion of students answering them correctly, then one criterion would be based upon discrepancies in the rank orderings. Similarly, if the (biserial) correlation between an item response and total score differed in the two populations, this would be viewed as evidence for possible translation problems. Variations upon such criteria are often used, for example based upon non-linear weighted functions of item responses, but the principle is the same.

This technique is akin to procedures which have been suggested for detecting 'biased' items when comparing subpopulations, for example defined by gender or ethnicity. The difficulty is that there is no way of knowing whether a few 'aberrant' items present translation or other problems or whether they are in fact valid achievement indicators measuring real population differences. The usual psychometric procedure for resolving this dilemma is to make the assumption that the set of item responses can be modelled in terms of a single unidimensional student 'ability' or 'trait', in a sense which is discussed more fully below (see p. 22). Such an assumption, unfortunately, merely restates the dilemma, this time in terms of whether an aberrant item should be regarded as problematic or whether the item set is legitimately viewed as spanning at least two dimensions. This is not to say that such analyses cannot be used to provide suggestions about interesting population differences, but rather that they cannot properly be described as tests for translational validity. A more detailed discussion of such psychometric tautologies is given by Goldstein and Wood (1989).

Finally, it does seem reasonable to ask whether in all cases perfect or near-perfect translation is worth aiming for. The inherent variation in language structures in some cases seems to preclude this anyway and in other cases the practical difficulties deny full knowledge of whether the goal has been achieved. Instead, we should perhaps regard the translation issue as belonging, at least partly, to the stage of data interpretation. The goal of trying to render tests and questions equivalent is a sensible one, so long as it is recognized that subsequent analysis may provide further insights and understandings about linguistic and cultural differences.

Data-processing technicalities

Since the early days of international studies, the computer revolution has transformed the data processing and analysis of large-scale surveys. Data transfer from test booklets and questionnaires, and other instruments can be carried out rapidly, and the process of cleaning data prior to analysis likewise has been speeded up. This has been demonstrated in the reports of the first and second IAEP mathematics and science surveys (Lapointe et al., 1989, 1992) where initial analyses were published little over one year from the start of the survey. These studies utilized computers at most stages of piloting, administration and analysis within participating countries as well as centrally.

In the past the IEA publication time scale has been longer. For example, the second international mathematics study began to produce fully comprehensive country comparisons some three years after data collection started (Robitaille and Taylor, 1986). More recently, however, the IEA computers in education study (Pelgrum and Plomp, 1991) produced a comprehensive report within two years of starting to collect data which included some quite complex statistical modelling. Likewise the IEA reading literacy study collected data in 1991 and produced a first summary report in mid-1992. It seems not unreasonable to expect future surveys to produce useful summary reports no more than a year after data collection ends and to make data available for secondary analysis shortly afterwards.

Data scaling and data interpretation

A prevailing assumption behind all international comparative studies has been that they exist principally, if not entirely, in order to describe country differences. The desire to explain differences, for example in terms of curriculum exposure, teacher attitudes or cultural expectations, has always been of concern and a relatively recent development has been the use of powerful statistical modelling for this purpose (see, for example, Pelgrum and Plomp, 1991; Keeves, 1992*a*). Without an attempt to provide such explanation the descriptive statistics have little real use, other than as propaganda. This is perhaps most evident in the analyses produced by the IAEP for science and mathematics achievement (Lapointe et al., 1989, 1992).

The first international report on the IAEP assessment of science and mathematics in 1988 was based upon NAEP, carried out under the auspices of ETS. In a slim but well-presented and speedily published booklet, ETS presented comparative information about the average performances of each of five countries under various topic headings. For example, in mathematics the percentage of items correct for each country is reported for the topics of 'number', 'relations', 'geometry', 'measurement', 'data organization' and 'problem-solving'. In addition, the average total number of items correct is reported. As well as this, there are tables comparing the reported frequency of classroom mathematics and science activities, amounts of homework, and attitudes of students towards mathematics and science. There also are some comparisons based upon 'opportunity to learn', that is an average measure of the students' exposure to the topics being tested.

Except for a couple of instances, there is no attempt to interrelate factors. For example, it is extremely difficult to establish the fact that there is an association between opportunity to learn and performance on each topic (Wolfe, 1989). While the report does present results for separate topics, its main emphasis is on the overall science and mathematics 'proficiencies'. These are simply (weighted) averages of the subtopic scores with the weights approximately reflecting the number of items in each subtopic scale. Thus, in mathematics, since there are twenty-four number items out of the total of sixty-two, and only eight problem-solving items, the proficiency scale is much more heavily weighted towards the former. The report itself fails to comment on the implications of this. Rather, it seeks an interpretation of the proficiency scale by adding verbal descriptions to it, corresponding to particular scores based upon the observed performances of individuals achieving those scores. Thus, a score of 300 is said to correspond to students (at Grade 8) who 'can add two-digit numbers without regrouping and solve simple number sentences involving these operations'. The report claims that these descriptions can inform the reader about what children at that score point 'know or can do'.

Despite a caveat in the introductory section, there is little in this same report which tries to convey the tentative nature of international comparisons, and the problems of translation and interpretation which are well recognized by those responsible for designing and analysing the assessments. I have already examined the translation issue; I now turn to some of the interpretation issues raised by this IAEP report, bringing in IEA material also.

Opportunity to learn (OTL)

Comparing educational performance among population groups is a somewhat pointless exercise unless it can be contextualized by measuring the exposure students have had to relevant learning experiences. Clearly there are many influences on performance, but if education has any effect the exposure to a topic should be associated with performance on that topic. Thus, the information that the United Kingdom does well in problem-solving should be read in conjunction with the relatively high exposure that British children receive. Indeed, such exposure information may be of more use for many purposes than the performance data. In fact, from IEA surveys, although OTL is associated with achievement, the relationship does not always appear to be very strong (Goldstein, 1987, Chapter 5). Moreover, the relationship seems to vary across countries.

The principal difficulty with existing measures of OTL is that they tend to be rather coarse, measured for a group of students rather than each one individually and based upon retrospective data, namely the responses of teachers. It is to be expected that these circumstances will underestimate markedly any relationships which exist. In view of the importance of measuring OTL, one would hope that future resources will be directed at obtaining reliable individual student-level data; the SMSO study, mentioned earlier (p. 14), promises to be a useful starting-point.

Aggregated scales

One of the more misleading presentations of results of comparative studies is the emphasis given to aggregate scores of 'mathematics' or 'science'. Such scale scores typically are formed by averaging the responses for all the items in a subject area. This has two principal drawbacks. The first is that much of the real interest lies in individual topic areas and the second is that this reflects the weightings of topic items chosen by the test constructors. It has already been pointed out how in the first IEAP study, the implicit definition of 'mathematics' was weighted by number items. This has been a persistent problem in the reporting of IEA results.

The choice of items to be used in assessing, say mathematics, is the result of a negotiation among the participants in a study. The 'core' set of 'consensus' items agreed upon as common to all countries are those upon which international comparisons will be made. Yet because these represent a compromise, they may not be representative of any single country's overall intended or implemented curriculum. Nevertheless, they will be more representative for some countries than others. Thus, for example, in the first IAEP mathematics survey already discussed, those countries where the curriculum emphasizes numerical competencies as distinct from problem-solving will be relatively advantaged in comparisons of overall mathematics scores. As Wolfe (1989) points out, if different weighting systems are used for the components of mathematics, the relative position of countries will change, and he quotes the example of England and Wales which moves up the country rank order if an equal weighting is applied to topic areas. Westbury (1992) further discusses this issue and compares student achievement in terms of the curriculum coverage of second international mathematics study items in Japan and the United States. Unfortunately, his conclusions need to be treated with caution since his analysis does not properly adjust for topic-selection factors at the student level, and it also confines itself to adjusting for pre-existing achievement at the class rather than student level.

It seems clear that the very notion of reporting comparisons in terms of a single scale, for example of 'mathematics' or 'science', is misleading. Purves (1992) makes this point strongly with respect to writing proficiency, where he suggests that at least three separate dimensions are present and that student responses have to be interpreted in the light of cultural differences and expectations. He also emphasizes the subjective nature of choice of items in any test and his reservations about interpretations can be made for the other subject areas. Likewise, Swain (1990) points to the contextual influences of item characteristics on student responses and the complex multidimensional structures involved in second-language testing.

It appears to be somewhat pointless to devise separate test forms for components of science or maths or language if reporting is then undertaken principally in terms of a single scale. One possible alternative is to report several scales, each using a different weighting, but while this seems worth investigating, it may be somewhat confusing for most readers. What then is the appropriate level at which results should be reported? At one extreme it is possible to report on each assessment item separately. This has certain merits and there is a strong case for item-level analyses to be available. Yet, again, typically there are natural groupings of items covering specified aspects of the curricula which can form meaningful reporting levels. If this is to be done, then it is also important that readers of reports have easy access to all the constituent items, in the relevant translations, and not merely a sample set.

If the analysis of subscales of achievement is to be pursued, then it will be fruitful also to study the interrelationships between scales, that is, the extent to which performance on say, problem-solving in mathematics, is correlated with data-analysis proficiency, and whether these relationships differ from country to country.

Statistical scaling

Despite the argument outlined above in favour of disaggregated reporting, a number of data analysts claim to have developed single scales for 'science' or 'mathematics' or 'language' which would allow valid comparisons in terms of a single-scale value, irrespective of which subset of items from a larger collection was used in the assessment or what relative weightings were used for different components. Thus, the first IAEP science and mathematics surveys use so-called 'item response' scaling to produce single proficiency scales in mathematics and science; and Keeves (1992*b*, 1992*c*) argues for such scales and presents one such scale for science achievement in the first and second IEA science studies. Among other things, it is claimed that such scales allow comparisons of national achievement over time, independent of curriculum or cultural changes. In essence, the argument is as follows.

In order to illustrate the procedures, the scale developed for IEA science studies will be used. The assumption is first made that all the items under consideration are reflecting a single underlying 'trait' or 'dimension'. The general procedure is to ignore prima facie evidence for separate scales but rather to see whether, after constructing the scale, the data themselves provide evidence for rejecting a single scale. In the first and second IEA science studies, the scale was developed from the science achievement test items for 14-year-olds which were common to both surveys (Keeves, 1992*a*).

For each of these common items the basic assumption is made that any change in the proportion of correct responses over time is a reflection of changes in the population rather than, in an alternative sense, changes in the facility of the item. Thus, if the correct response rate for a physics item increased significantly, from 40 to 50 per cent, between the first and second science surveys, this would be interpreted as an increase in student achievement in this area of the physics science curriculum. At this preliminary stage, some items may appear 'anomalous', for example those for which the population response remains unchanged rather than increases as for the remaining items. A common procedure would be to eliminate such items as 'non-fitting' so that the scaling is then carried out on the remainder.

It is unnecessary to go into the details of the procedures by which final scale scores are produced, typically using time-consuming statistical modelling. In essence, however, for each survey one can think of making an estimate of the underlying trait of 'science' by calculating the average item score for each student – which is the average proportion correct if the items are simple pass/fail ones. A slightly more refined method uses a weighted mean where the weights are determined by the intercorrelations of the items. For one of the surveys, say the second science one, these student scores are then simply scaled so that they have a designated mean value (say 500) and a spread (say 0 to 1,000). Once the equivalence between such a 'convenience' scale and the 'raw' student scores has been established, all the scores can be given a scale value.

Having established this scale, it can be extended to include new items, so long as they are assumed to belong to the same 'trait'. This is done by comparing student responses on the new items to student responses on the existing scale items so that each new item can be assigned a 'difficulty' value (and, if the more refined method is used, other characteristics such as 'discrimination') alongside the difficulties of the existing items. With this information the new, more extensive instrument can be used to assign scale values to students. The 'linking' of tests in this way can be carried on for more stages, and in the IEA science study the tests for 10- and 14-year-olds for each survey were finally linked into a common scale. The results of such a procedure will sometimes be incorporated into a calibrated or scaled 'item' bank. From such a bank subsets of items can then be selected to form tests whose overall difficulties and other scale properties are regarded as known.

A technical description and evaluation of these scale-creation procedures has been given by Goldstein and Wood (1989). Before going on to look at how these scales have been used to interpret achievement, their limitations need be discussed. It should also be pointed out that IEA science researchers are not alone in preferring such scales; they have been used extensively by ETS in IAEP and the United States national achievement survey, NAEP. It has also been proposed that such scales be used in the third international mathematics and science study (TIMSS/TAC, 1993).

Limitations of item response scaling

A crucial assumption used in item response scaling is that of unidimensionality, defined as follows. If, for a set of test items or questions, the responses of a group of students are determined by a single 'trait' value, then that set of items is said to be unidimensional. In other words, the responses reflect the operation of one and only one underlying factor, be it 'reading ability', 'abstract reasoning' or whatever.

First of all, it should be noted that such a definition has to be population-dependent. Thus a test may be approximately unidimensional in one group of students but clearly not in another. This is especially relevant in international studies where very different systems and cultures are operating. Because a set of relationships holds in one or more countries, this cannot guarantee that it will do so elsewhere. In practice, of course, no set of items is perfectly unidimensional, so some statistical procedure has to be used for decidin whether a set of items 'approximates' unidimensionality and thus involves subjective judgement about what constitutes an adequate approximation. In order to achieve a scale that 'approximates' unidimensionality those items representing 'minority' dimensions will have to be removed or suitably modified until they conform. This of course will tend to increase the unidimensionality of a test, but not necessarily its 'validity' or fitness for purpose, that is, its capacity to measure what is intended. This is easily seen in the following simplified example.

Suppose we have two sets of truly unidimensional items representing respectively dimension A and dimension B. A test constructor chooses a fifty-item test using forty items from A and ten items from B and then carries out the standard 'item analysis' or 'item response theory' (IRT) procedures and discovers, unsurprisingly, that the ten B items seem discrepant, that is, they do not exhibit the behaviour of the majority. In accordance with common practice and in order to obtain a unidimensional test these B items are omitted. A unidimensional test is obtained. But of course, it merely reflects the decision taken by the test constructor originally to weight the test with the A items. A unidimensional test would also have been obtained if the roles of the A and B items had been reversed. In that case, however, the test would represent something quite different, for example ranking students differently and altering comparisons between population groups. This example is a simple one, but Goldstein and Wood (1989) show how the same principle applies quite generally. Merrill Swain (personal communication) gives an example from language testing where proficiency in 'ability to communicate' and proficiency in 'grammatical accuracy' differ markedly between French immersion students in Canada and students studying English in China. A test which was reduced to items reflecting just one of these proficiencies would thus disproportionately favour one group over the other.

In addition to these fundamental difficulties, the statistical procedures themselves are far from satisfactory. General so-called 'goodness of fit' tests provide weak evidence for confirming unidimensionality unless they are concerned with contrasting a unidimensional structure with a specific multidimensional structure. Thus, in the above example, if we had information to suggest that the ten B items belonged to a separate dimension, then a powerful statistical test could be devised for this hypothesis. Usually, however, such information is unavailable and a wide variety of possible alternative structures will have to be allowed for in the statistical test procedure. As a result, such 'non-specific' tests will often fail to detect a real multidimensional structure.

There is a further problem with almost all attempts to produce unidimensional scales. This derives from the fact that the data samples used tend to come from very heterogeneous groups. This means that where there are high intercorrelations among items, some of this will be due to other factors such as family background and especially curricula differences. There are hardly any studies which have seriously attempted to study this issue by 'partialling out' such factors before reaching conclusions about dimensionality. The IEA studies in fact have relatively good data for this and the international facet would make such analyses particularly valuable.

We see therefore that the assumption of unidimensionality should be handled with care and that, despite a high level of statistical sophistication, both the objective and subjective intentions of the test constructors remain paramount. Claims for having established unidimensionality should be treated cautiously.

In the light of the discussion of scale construction and as with the case of mathematics discussed earlier, any overall scale is best viewed as a particular weighted average of its separate components with no other special meaning. Furthermore, there may be a serious conflict between claims for a single unidimensional scale while at the same time reporting separate components. This is illustrated, for example, by Keeves (1992*a*) who presents results comparing countries on a single science scale as well as in terms of separate components such as 'reasoning' and 'investigation'. On the results for the separate components the countries involved have differing rank orders, which suggests that a single international scale is highly implausible.

Time trends

While most attempts to scale test items across time have been concerned with producing general unidimensional scales and so left themselves open to the criticisms outlined above, it would in principle be possible to confine such attempts to narrowly defined, and hence perhaps truly unidimensional traits. Unfortunately this too runs up against logical difficulties, as follows (see also Goldstein, 1983).

Returning to the item whose facility rises from 40 to 50 per cent from the first to the second occasion, how do we interpret this? It has been pointed out that item response models interpret this as a shift in the population's propensity to achieve success on the item. Yet just as easily we might assume that the population had not changed in any way, but that the item had just become 'easier'. It is possible to imagine situations where this might occur, such as the recent incorporation into common language of words used in the test. The reverse situation can also occur where an item can become more difficult because, say, the school curriculum has changed. These considerations will tend to apply only over long-term periods; over short-term periods it may be reasonable to suppose that such factors are relatively unimportant. The problem, however, is that it is the longer-term periods which are usually of most interest. Even over short-term periods, however, serious problems can arise and these will be discussed below.

The point is that it is impossible to resolve the issue of whether, in some absolute sense, an item retains its characteristics and the population changes or vice versa, or perhaps a mixture of both. In some circumstances it may be possible to reach a measure of agreement about the interpretation of any changes, but there can be no purely technical solutions to this duality of interpretation. What can be said is that on a chosen set of test items – those that happen to be common on both occasions – achievement has changed in particular directions. Simple interpretations of such changes (Keeves 1992*a*, 1992*c*) are therefore uninformative. Among other factors which need to be studied is that of the continuing relevance of each of the common items to each country's curriculum. In addition, we would need to understand why the particular common items were chosen by the test constructors and whether the mechanisms of choice could have led to items 'biased' in one particular direction.

It seems that this duality problem is not well understood. For example, Keeves (1992*a*, p. 265) points out that 'to rely on the items that are common to the two occasions for any comparisons made, must likewise be considered to lead to an incomplete and inadequate assessment of change'. Yet he also claims that item response models allow the construction of a scale that is 'valid across countries and over time'.

Finally, recent empirical investigations have thrown some light on another aspect of item response-scale construction techniques, namely that of item parameter invariance. The assumption underlying most of the psychometric models for test item responses is that the characteristics of a test item, for example its difficulty or discrimination, are constant and uninfluenced by different contexts. Thus, the ordering of items in a unidimensional test will not change the item parameter values, nor will the incorporation of new items in a test. (Note that this is not the same as the assumption of item response independence which states that for a given individual and a given test, the probability of a 'correct' response to an item in the test is independent of responses to any other items.)

This assumption is extremely important for international comparisons where tests are often augmented by locally introduced items and most importantly where comparisons across time are attempted using a set of items common to tests at each occasion, but coexisting with different 'other items' at each occasion.

The recent evidence which casts doubt upon this assumption is that from NAEP, administered by ETS. It was found, upon comparing the results from 1984 and 1986 on the basis of a set of common items, that there were dramatic falls in performance for 9-year-olds and 17-year-olds. Because this was regarded as extremely unlikely, an extensive investigation was undertaken to study the reasons (Beaton and Zwick, 1990). The conclusions were that the students' performances on the bridge items changed according to the context in which the items were administered, that is how and where they appeared in the test booklets. In brief, 'when measuring change, do not change the measure'. The report urges considerable caution when contemplating the measurement of change over time and clearly recognizes that no satisfactory procedure for so doing is available.

Statistical modelling

There have been some attempts by IEA to use elaborated statistical models to explore the data. The study on the use of computers (Pelgrum and Plomp, 1991) uses structural equation models to explore the structure of data related to the implementation of computer education, and there are examples of path models and some use of OTL information. The majority of analyses, however, concentrate on country comparisons and simple group differences such as those between males and females. The IAEP analyses (for example, Lapointe et al., 1992) present aggregate-level comparisons between test scores and other variables such as hours spent on homework, and also present crude indications of the strengths of relationships within countries, between students. These analyses look at no more than two factors at a time and so are extremely limited in terms of explaining country differences.

To date there has been very little attempt to fully model the within-country variation in test scores and other variables. While the analyses generally have been careful to take account of the complex sample designs, they have not explicitly attempted to study the way in which achievement, attitudes, etc., vary from school to school, or area to area.

It is particularly important to develop explanatory models which attempt to explain, statistically, observed relationships. Such relationships might be those between, say, OTL and achievement or between achievement and reported hours of homework. What is of real interest is to explore reasons for such associations in terms of other measured characteristics. These might be the experience of the teachers, curriculum variables or the home background of the students. It is also important to study whether explanations for these relationships differ among countries and then attempt to understand why.

Clearly much of this kind of analysis could be carried out by researchers not involved in the original studies. To make this feasible, however, requires easy accessibility not only to the data files but also to the original test forms and questionnaires, and implies a high level of data organization with properly structured codebooks, sample descriptions, etc. IEA has devoted much effort to setting up suitably resourced and elaborate data archives for this purpose, and intends to provide archives for all its studies.

Multilevel models

In recent years statisticians have developed powerful tools, known as hierarchical or multilevel models, for studying simultaneously the between-school and between-student variation (Paterson and Goldstein, 1991; Bryk and Raudenbush, 1992). These models and associated software packages are now used extensively in so-called 'school effectiveness' studies, and in a wide variety of applications in the social and medical sciences. The scope and importance of these models can be summarized as follows.

It is well known that when carrying out statistical tests or calculating confidence intervals, account needs to be taken of the clustering of the data, which generally implies that students within a school are more alike in their test scores than students chosen from different schools. In the major international studies, this is usually done by calculating 'design effects' based upon preliminary analyses of the sample data. These design effects are estimates of the extent to which the precisions of various statistics, such as means or proportions, are inflated by such clustering. They can be used to make suitable adjustments to the statistical procedures (Skinner et al., 1989). Multilevel modelling takes an essentially different approach. It attempts directly to model the hierarchical structure of the data. That is, it recognizes that a test score can be considered as the sum of contributions from each school and from each student within a school. In the simplest model, each school will have its own 'mean' score and each student a contribution to be added or subtracted from this. In more complex models, parameters such as the slope of a regression line or the difference between males and females can be allowed to vary from school to school.

Such direct modelling has several benefits. It automatically deals with the problem of accounting for the clustering of data while also providing information about differences between schools. In some cases it is the latter which is the most important focus of the analysis. For example, in a study of progress made by secondary school students in Inner London (Nuttall et al., 1989) the between-school variation for initially high-achieving students was found to be much greater than that for initially low-achieving students. This 'differential school effectiveness' is important when attempting to compare schools and leads to a further set of research questions. These models can easily incorporate covariates which may be defined at the level of the student, such as attitude or time spent on homework, or at the level of the school or class, such as the average ability in the class or the characteristics of the teachers.

IEA data structures often involve complex 'matrix' designs. For example in the second science study, there was a core test of thirty items and several 'rotated' forms in the separate areas of biology, chemistry, etc. Each student would take the core plus one or two rotated forms. While such a design reduces the burden on individual students, who no longer have to complete all forms, it has meant that the resulting analysis has been difficult to carry out. Using a multilevel modelling approach, such data can be handled efficiently, and the approach allows full modelling of the core and all rotated forms. An example of such an analysis is given by Goldstein (1987, Chapter 5). It also obviates any practical need to collapse the separate topic areas into a single scale.

The ability to focus an analysis on studying variation between schools also raises another interesting possibility, namely that comparative analyses can report differences in the extent of between-institution variation and the factors which appear to 'explain' it, rather than just differences in mean scores. Such analyses can yield valuable insights into educational structures while at the same time being less compromised by problems of translation and sampling.

The problem of comparability

The continuing emphasis in international studies has been on attempts to ensure that the 'same' questions are being asked and the 'same' achievements are being measured. The difficulties associated with this have already been discussed; here I want to question whether, at least to some extent, these difficulties can be avoided by posing different prior questions.

Defining equivalence

In a strict sense the issue of whether a test, say of science knowledge, measures the same thing in two different cultures is irresoluble because there is no other external criterion which can be used to judge the issue. Rather, the problem is one of definition whereby the particular set of versions of a test have to be defined as equivalent by those responsible for producing them. Such a judgement will normally be provisional and typically contingent on satisfying reasonable criteria, for example concerning translation. Once agreement can be reached, the problem resolves itself into an empirical one of attempting to validate or dismiss the judgement. This still allows comparisons to be made, but subjects their interpretation to the general caveat that further study may cause such interpretations to be modified.

An important component of the continuing empirical validation of such judgements of equivalence will lie in the fitting of explanatory models which attempt to account for observed differences. For example, the familiarity of students with the particular type of test format used may account for some differences. It could then be argued that particular test-item formats may lead to lack of real equivalence, but that this can be taken into account by adjusting for student 'exposures'. In other words, equivalence is still possible, but its definition has to be extended in an empirical fashion. One might term this 'statistical equivalence' because it is defined within the framework of a statistical model which explicitly attempts to adjust for 'nuisance' factors which are unmistakably present but which are not the focus of interest. With such a modified definition the researcher would then wish to go on to look at other factors which were associated with remaining differences.

Of course, it may not always be possible to follow such a line of reasoning. If there are countries where there is little or no variation in familiarity with different item formats, then there may be no basis for carrying out an adjustment procedure. There is also the serious problem of deciding which factors are legitimately those which can be used for adjustment. While item format familiarity might be suitable, it is not so clear that a measure such as teaching method should be used. The latter would normally be thought of as a factor of interest in its own right for explaining differences rather than in helping to define what is a fair comparison.

Second-order comparisons

When modelling hierarchical structures, in addition to schools there will usually be higher level administrative or geographical units within which schools themselves are nested with significant between-unit variation. In addition some types of schools, for example those in urban areas, may exhibit more variation than others.

Using data from the IEA second mathematics study, Goldstein (1987) found that the percentage of the total variation between schools in Japan (4 per cent) was very much smaller than that in British Columbia (11 per cent). This may reflect greater homogeneity of curriculum or intake achievement in Japan, but might also be a consequence of the tests used. It is strictly unnecessary to have 'equivalent' tests when carrying out second-order comparisons of this kind since interest lies in the relative homogeneity of systems rather than in their absolute relationship to each other. This reinforces the argument for studying the separate components of achievement. Second-order comparisons of different components could yield interesting insights into the priorities within different systems and the extent to which institutional variation was accounted for by factors such as OTL, teacher experience and so forth.

Prior achievement

In nearly all cases large-scale international comparative studies have collected cross-sectional data, that is information on students at one point in their careers. IEA, in the second mathematics study, collected limited longitudinal information, measuring students up to nine months apart in age. While such information can be used to estimate changes during a single school year, it is of little use for making comparisons between schools in terms of their overall contribution to students' progress. There is now a considerable literature on 'school effectiveness' studies (see for example Bryk and Raudenbush, 1989) which emphasizes the importance of taking account of intake achievements when students start school in order to make fair comparisons between schools. It is generally agreed by researchers that such 'value added' estimates of each school's 'effect', together with caveats about measurement relevance and reliability, are the only sound basis for comparing schools.

We can apply similar reasoning to comparisons between countries, where the purpose is to assess the relative effectiveness of education systems. Thus, for example, comparing the achievements of 10-year-olds in mathematics may partly or even largely reflect pre-existing differences present when the students started school. Influences such as social background, health, parental education, etc. may all be influential. In order to isolate the contribution of the education systems during any particular stage of education, suitable measures of student achievement prior to entry to that stage are essential. Of course, it is no easy matter to carry out long-term longitudinal studies of cohorts which would allow such analyses to be carried out, but it is difficult to see how, without such studies, any definitive conclusions can be reached. Thus, by and large, country comparisons tend to place the poorest countries behind the richer ones and this should occasion no surprise. What would be very interesting would be to know how those comparisons appeared once the initial achievements had been allowed for. Some of those countries with the lowest achieving intakes may well have secured more progress for their children. If comparisons are to be interpreted in the light of other measurements such as curriculum content or school organization, then it is the value added by the schools which is the key measure.

If this argument is accepted, it raises a serious problem for the usefulness of existing studies. While cross-sectional information on achievement is useful in providing a baseline from which to begin to draw inferences, it is only possible to begin to draw sound inferences about the impacts of education systems and institutions from long-term longitudinal data. Similar reservations apply to other kinds of student data such as motivation and attitudes. Of course, such kinds of data are not the only kinds collected by international studies, and timely data on organization, qualifications of staff, school resources, etc., are also valuable.

Conclusions

I shall attempt to summarize some conclusions and suggest directions in which I believe international comparative studies profitably could develop.

First, it seems very clear that there is an important role for such studies and that IEA is currently the most suitable vehicle to pursue them. IEA has acted as a key forum for debating many of the relevant issues and we may expect this to continue. One of the most valuable outcomes of existing studies has been the accumulated experience gained by educationists worldwide in the construction, analysis and interpretation of comparative data. Nevertheless, it is important that improvements be made in certain areas.

In my view, there has been an unfortunate reliance upon uni-dimensional summaries of achievement test scores. In addition, the use of sophisticated statistical item response models to carry this out is an unwelcome development because it obscures too easily the true nature of what is occurring. The 'International Reading Scale' in the reading literacy study is a striking example. The discussion of these techniques is an attempt to make their essential properties better understood so that informed decisions can be taken by those responsible for designing studies. Most importantly, these issues need to be well understood by governments and policy-makers who are the principal providers of funds and important users of results. It seems that much of the pressure to produce simple summary comparisons has come from these latter groups and it is therefore extremely important that the issues are clear and that policy-makers understand the implications of their demands. This is most appropriately done by those closely involved in international studies, and IEA in particular could give an important lead in this. One implication is that, instead of study reports highlighting overall comparisons, they should concentrate on differential performance, properly contextualized, with discussions of any policy implications. Indeed, there is a strong case for refusing to report any comparisons in simple unidimensional summary terms such as 'mathematics', 'science' or 'language'.

The problems of interpretation of purely cross-sectional achievement scores are legion. Future studies should begin to plan on the basis of long-term longitudinal studies encompassing, as far as possible, whole stages of education such as the elementary or secondary periods. Despite their difficulties, these provide the only secure paths to proper understanding of the role of education. If this is not attempted, it will become very difficult to justify large-scale comparative studies if they remain solely cross-sectional.

I have referred to a number of more technical issues associated with interpretation. These are to do with equivalence across languages and cultures, the difficulties of interpreting trends over time, the problems of properly standardizing for age and grade, the need to model properly the hierarchical structure of educational data and the importance of carrying out second order comparisons based upon the modelling of between-institutional variation. In all of these areas I believe that there is important methodological work to be done which would have an importance wider than comparative studies alone. With easily available and powerful computing facilities there are no serious technical barriers to such developments.

Finally, I am convinced that the existence of an organization such as IEA with its democratic structures and enthusiastic supporters is essential. It provides the only sensible approach to making comparisons and it has often shown itself capable of responding to new issues and able to tackle difficult problems. Above all, it provides an important counterweight to the only other sources of comparative information which are based, at one extreme, upon poorly designed one-off comparisons and, at the other, on official government statistics. The former suffer from problems of poor controls and lack of experience while the latter suffer from distortions related to inadequate coverage, varying definitions and selective reporting.

References

BEATON, A. E.; ZWICK, R. 1990. *Disentangling the NAEP 1985–1986.* (Mimeo.)

BRISLIN, R. W. 1970. Back-translation for Cross Cultural Research. *Journal of Cross Cultural Psychology,* Vol. 1, pp. 185–216.

BRYK, A. S.; RAUDENBUSH, S. W. 1989. Toward a More Appropriate Conceptualisation of Research on School Effects: A Three Level Hierarchical Linear Model. In: R. D. Bock (ed.), *Multilevel Analysis of Educational Data.* New York, Academic Press.

——. 1992. *Hierarchical Linear Models.* Newbury Park (Calif.), Sage.

ELLEY, W. B. 1992. *How in the World Do Students Read?* The Hague, IEA.

FINEGOLD, M.; MACKERACHER, D. 1986. Meaning from Curriculum Analysis. *Journal of Research in Science Teaching,* Vol. 23, pp. 353–64.

GOLDSTEIN, H. 1983. Measuring Changes in Educational Attainment over Time: Problems and Possibilities. *Journal of Educational Measurement,* Vol. 20, pp. 369–77.

——. 1987. *Multilevel Models in Educational and Social Research.* London/New York, Griffin/Oxford University Press.

GOLDSTEIN, H.; WOOD, R. 1989. Five Decades of Item Response Modelling. *British Journal of Mathematical and Statistical Psychology,* Vol. 42, pp. 139–67.

HAMBLETON, R. 1992. *Translation of Achievement Tests for Use in Cross-national Studies.* Vancouver, IEA International Coordinating Centre. (Mimeo.)

HANNA, G. 1993. The Validity of International Performance Comparisons. In: M. Niss (ed.), *Investigations into Assessment in Mathematics Education.* Amsterdam, Kluwer.

HAYES, W. A. 1991. *IEA Guidebook 1991: Activities, Institutions and People.* The Hague, IEA.

HULIN, C. 1987. A Psychometric Theory of Evaluations of Items and Scale Translations. *Journal of Cross Cultural Psychology,* Vol. 18, pp. 115–42.

KEEVES, J. P. 1992*a*. *The IEA Study in Science III: Changes in Science Education and Achievement, 1970 to 1984.* Oxford, Pergamon Press.

——. 1992*b*. *Learning Science in a Changing World.* The Hague, IEA.

——. 1992*c*. Scaling Achievement Test Scores. In: J. P. Keeves (ed.), *Methodology and Measurement in International Educational Surveys.* The Hague, IEA.

LAPOINTE, A. E.; MEAD, N. A.; ASKEW, J. M. 1992. Learning Mathematics. Princeton (N.J.), Educational Testing Service.

LAPOINTE, A. E.; MEAD, N. A.; PHILLIPS, G. W. 1989. *A World of Differences.* Princeton (N.J.), Educational Testing Service.

LEUNG, F. K. S. 1992. *A Comparison of the Intended Mathematics Curriculum in China, Hong Kong and England and the Implementation in Beijing, Hong Kong and London.* London, University of London. (PhD Thesis, mimeo.)

LITTLE, A. 1978. *The Occupational and Educational Expectations of Students in Developed and Developing Countries.* Brighton (U.K.), Institute for Development Studies.

MCDONALD, G. 1992. 'Henry and Iain . . .: A Comment on a Response. *New Zealand Journal of Educational Studies,* Vol. 27, pp. 103–6.

MEDNICK, S. A.; BAERT, A. E. 1981. *Prospective Longitudinal Research.* Oxford, Oxford University Press.

NUTTALL, D. L.; GOLDSTEIN, H.; PROSSER, R.; RASBASH, J. 1989. Differential School Effectiveness. *International Journal of Educational Research,* Vol. 13, pp.769–76.

PATERSON, L.; GOLDSTEIN, H. 1991. New Statistical Methods for Analysing Social Structures: An Introduction to Multilevel Models. *British Educational Research Journal,* Vol. 17, pp. 387–94.

PELGRUM, W. J.; PLOMP, T. 1991. *The Use of Computers in Education Worldwide.* Oxford, Pergamon Press.

PURVES, A. C. 1992. Reflections on Research and Assessment in Written Composition. *Research in the Teaching of English,* Vol. 26, pp. 108–22.

ROBITAILLE, D. F.; TAYLOR, A. R. 1986. *A Comparative Review of Students' Achievement in the First and Second Mathematics Studies.* Washington, D.C., National Center for Education Statistics.

ROSIER, M. J. 1987. The Second International Science Study. *Comparative Education Review,* Vol. 31, pp. 106–28.

SKINNER, C. J.; HOLT, D.; SMITH, T. M. F. 1989. *Analysis of Complex Surveys.* Chichester (U.K.), Wiley.

SWAIN, M. 1990. Second Language Testing and Second Language Acquisition: Is There a Conflict with Traditional Psychometrics? In: J. Alatis (ed.), *Georgetown University Round Table on Languages and Linguistics.* Washington, D.C., Georgetown University Press.

TIMSS/TAC (Third International Mathematics and Science Study/Technical Advisory Committee). 1993. *Summary of Third International Mathematics and Science Study/Technical Advisory Committee Meeting, Vancouver, May 10–13, 1993.* The Hague, IEA. (Mimeo.)

TRAVERS, K. J.; WESTBURY, I. 1989. THE IEA STUDY OF MATHEMATICS I: ANALYSIS OF MATHEMATICS CURRICULA. Oxford, Pergamon Press.

WESTBURY, I. 1992. Comparing American and Japanese Achievement: Is the U.S. Really a Low Achiever? *Educational Researcher,* Vol. 21, No. 5, pp. 18–24.

WOLFE, R. G. 1989. *An Indifference to Differences: Problems with the IAEP-88 Study.* (Mimeo.)

Scaling and aggregation in IEA studies: critique of Professor Goldstein's paper

Geoff N. Masters

Introduction

This document has been prepared at the request of the Technical Advisory Committee of the International Association for the Evaluation of Educational Achievement (IEA) following a meeting of the Committee at El Escorial, Spain, on 12 September 1993. At that meeting the Committee considered the paper entitled *Interpreting International Comparisons of Student Achievement,* which had been prepared for UNESCO by Professor Harvey Goldstein of the University of London. The Committee was concerned at the contents of the paper, in particular at what the Committee considered to be the paper's inaccurate description of procedures used in IEA research and inadequate discussion of issues in scaling and aggregating achievement data. The Chair of the Technical Advisory Committee was asked to prepare the following written summary of the Committee's concerns for forwarding to UNESCO.

This document addresses a number of misunderstandings in the Goldstein paper. Some of these misunderstandings are simply incorrect understandings of IEA processes; others are more fundamental misconceptions concerning the purposes and nature of measurement in IEA research.

The role of measurement in IEA studies

Measures of student achievement are fundamental to IEA's evaluation studies. International comparative research into the effectiveness of different educational arrangements and curricula depends on valid, meaningful *measures* of what is being achieved in different countries. Measures of student achievement are also essential to attempts to conduct international longitudinal studies and to monitor changes in educational achievement over time.

Goldstein recognizes the necessity of valid, reliable measurement in international research. Indeed, measures of student achievement are essential to the methods he himself proposes. In his discussion of longitudinal studies to establish the 'value added' by different education systems, for example, he stresses the need for suitable *measures of achievement* at the point of commencing a particular stage of education:

> In order to isolate the contribution of the education systems during any particular stage of education, suitable measures of student achievement prior to entry to that stage are essential.

The construction of measures is a major task in IEA studies. Only after the measurement problem has been solved is it possible to begin to apply other statistical methods such as multilevel modelling, linear structural equation modelling, and so on, to investigate factors influencing levels of achievement in different countries. This is a well-understood principle in scientific research; before investigating complex environmental influences on animal growth, for example, it is first necessary to construct reliable instruments for measuring height and weight (the relevant dependent variables). In general, the construction of measures requires (a) a clear understanding of the variable to be measured and (b) a set of procedures for monitoring the functioning of a measuring instrument.

Each instrument used in IEA studies is constructed with a specific variable (aspect of achievement) in mind. The construction of an instrument begins with a proposition that a variable (e.g. 'narrative reading ability') can be operationalized and measured through a set of relevant questions. Whether such a variable can be usefully operationalized and measured is an empirical question that can be answered only through careful analysis of how the intended instrument works in practice.

Importantly, the construction of a measuring instrument is a deliberate decision to restrict and focus attention on one dimension of achievement at a time. The decision to attempt to develop measures of 'narrative reading ability', for example, deliberately focuses attention on one of many dimensions of student achievement. Again, this is a generally understood principle of measurement. Measures of temperature, height and blood pressure, and even global measures of 'health' and 'physical fitness' are quite deliberately one-dimensional descriptions of individuals.

Assumptions versus intentions

Many of Goldstein's misunderstandings arise from an inadequate appreciation of this fundamental principle of measurement. He argues, for example, that in international studies 'there has been an unfortunate reliance upon unidimensional summaries of achievement test scores'.

In his Conclusions Goldstein identifies this as a first area in which 'it is important that improvements be made'. At best, this reflects an idiosyncratic understanding of measurement. In educational measurement, tests are constructed to provide measures of *one* aspect (dimension) of a student's achievement at a time. This is not 'unfortunate', it is the *intention;* it is the reason for constructing an instrument in the first place.

Throughout Goldstein's report there is a failure to distinguish between assumption and intention, and an inadequate appreciation of the purposeful, constructive nature of educational measurement. For Goldstein, unidimensional measures are the result of questionable 'assumptions' that performances on a test can be explained in terms of some single underlying psychological 'trait'. But he fails to appreciate that in IEA research, each test begins with an *intention* to assemble a set of items that will work together as indicators of a particular dimension of achievement. Unidimensionality is not an assumption in IEA studies; it is an intention.

In IEA research, test items are developed as opportunities to collect evidence about students' levels of achievement on some particular variable. This means that the measurement of achievement in IEA studies is not so much a process of developing one-dimensional 'summaries' of complex data as a process of inferring students' levels of achievement on a variable using items intended to work together as indicators of that variable.

It is always possible to see differences among items that may cause them *not* to work together as indicators of the variable of interest. The task confronting the IEA researcher is to establish empirically whether items work together well enough to be treated as indicators of the intended variable. Goldstein is also critical of this feature of standard psychometric practice:

> The general procedure is to ignore prima facie evidence for separate scales but rather to see whether, after constructing the scale, the data themselves provide evidence for rejecting a single scale.

Again, this reflects a misunderstanding of *intention*. IEA items are assembled with the intention that they will work together as indicators of the variable to be measured. This intention (or hypothesis) must be put to the test in every case.

The same misunderstanding leads to another 'fundamental difficulty' for Goldstein: to construct an instrument that measures only one variable, it is necessary to remove or modify items that are not helpful in defining that variable.

> In order to achieve a scale that 'approximates' unidimensionality, those items representing 'minority' dimensions will have to be removed or suitably modified until they conform.

Although this may be a 'fundamental difficulty' for Goldstein, it comes as no surprise to any instrument developer. If, on a biology test, an item proves to be a better test of reading comprehension than of biology, then that item should be removed from the biology test. If, on a geometry test, an item proves to be a better test of arithmetic manipulation than of geometry, then that item also should be removed or rewritten. The removal and revision of test items not helpful in measuring the variable of interest is simply good measurement practice.

Goldstein makes his argument against this common-sense practice by describing a procedure that no test constructor would use deliberately, namely, trying to measure two clearly different variables with the same instrument. He points out, correctly, that if a test developer were to do this and then use an item response model to analyse the resulting data, he or she would be led to the conclusion that one of the two subsets of items (A or B) should be removed so as to provide an instrument measuring only *one* variable. (In practice, a test developer might well decide to develop *two* instruments for the two different variables involved.) Goldstein then points out that if the minority subset of items was simply discarded, then the variable measured by the resulting instrument would depend on which subset was in the minority:

> A unidimensional test would also have been obtained if the roles of the A and B items had been reversed. In that case, however, the test would represent something quite different, for example ranking students differently and altering comparisons between population groups.

This is exactly what one would hope to see happen and is a convincing argument for using an item response model to deal with the problem that Goldstein concocts.

Much of Goldstein's paper reflects the perspective of a data analyst with limited control over the data presented. Goldstein worries about making unwarranted 'assumptions' of data; he worries about the adequacy of 'unidimensional summaries'; he worries about 'prima facie' evidence that items might not be measuring a common dimension; and he worries about throwing away items. IEA instrument-developers come to the measurement problem with a different perspective. They do not come with 'assumptions', they come with *intentions*. They are not trying to summarize complex data, they are trying to construct measures of variables. They recognize that there will be differences among items but understand that the important question will be whether their items work together empirically to operationalize and measure the variable in which they are interested. And they know that to develop useful instruments it often will be necessary to revise and even discard items.

From the perspective of test developers, individual items are relatively unimportant. Their focus is on the variable they are attempting to measure. From this perspective, individual items are a means to an end: potentially useful but ultimately expendable examples from a virtually infinite number ef possibilities.

In contrast, Goldstein's paper reflects a general nervousness about focusing on variables. Much of his argument is, in his own words, an 'argument in favour of disaggregated reporting'. Even where he hesitantly acknowledges the possibility of some aggregation, he is reluctant to treat items merely as examples:

> What then is the appropriate level at which results should be reported? At one extreme it is possible to report on each assessment item separately. This has certain merits and there is a strong case for item level analyses to be available. Yet, again, typically there are natural groupings of items covering specific aspects of the curricula which can form meaningful reporting levels. If this is to be done, then it is also important that readers of reports have easy access to all the constituent items, in the relevant translations, and not merely a sample set.

Once again this reflects the perspective of a data analyst with limited control over the situation: the best that can be hoped for is that, within the complex presenting data, it may be possible to identify 'natural groupings' of items. This approach to the analysis of test data is a poor alternative to an approach designed to assist instrument-makers in their efforts to construct measures on clearly conceptualized achievement variables.

The supervision of measurement

If an instrument is to provide measures that can be compared from occasion to occasion and from place to place, then that instrument's construction and use must be closely supervised with the help of an appropriate statistical method. In IEA studies, this supervision is especially important because of the desire to use the same instrument to measure and compare levels of achievement in different education systems and languages, and over time. Comparisons of this kind are possible only if it can be demonstrated that an instrument operates in essentially the same way in the different contexts in which it is used.

A variety of statistical procedures have been developed to help supervise the construction and use of measuring instruments. Early IEA studies relied on traditional test statistics. Recent IEA studies have used both traditional test statistics and modern measurement models (item response models) developed for this purpose..

Interestingly, Goldstein does not propose a statistical method for the construction of measures. Nor does he demonstrate an understanding of the need for such a method. Indeed, his expressed views on aggregation are so confounded by his opposition to total subject scores that it is difficult to establish whether he is in favour of constructing measurement scales at all. His criticism of item response models as 'time-consuming' and his claim that they 'obscure too easily the true nature of what is occurring' suggest limited experience in using these methods in practice.

Measurement challenges in IEA studies

Attempts to make international comparisons of achievement pose a number of special measurement challenges. These include the challenge of making meaningful comparisons of achievement levels over time, across countries with different educational arrangements and in different languages. Communicating results to policy-makers in an easily understood form presents a further challenge to IEA researchers.

Measuring change over time

IEA studies make comparisons over time by including some of the same items in tests used at different times. These items provide a common point of reference for 'equating' different test forms. Goldstein describes the equating of tests in IEA's first and second science studies, although his description of this process is not accurate.

Before comparisons over time can be made it is necessary to verify that the set of common items function in the same way on both occasions (that is, their relative difficulties are statistically equivalent on the two occasions). Once this is confirmed, students' success rates on the common items are compared between the first and second occasions. If students do better on the common items on the second occasion, then that is interpreted as an indication that levels of achievement have increased. Goldstein also sees this as problematic:

> For each of these common items the basic assumption is made that any change in the proportion of correct responses over time is a reflection of changes in the population rather than, in an alternative sense, changes in the facility of the item.

In fact, this is a meaningless distinction because item response models used in IEA studies deal only in relativities. It makes no difference under the model whether improvement is expressed as an increase in student achievement or a decrease in item difficulty. However, since the items themselves are unchanged from one occasion to the next, it is convenient to assume that they have not become 'easier' in any absolute sense and to attribute the increase in success rate to an increase in students' achievement levels. Certainly policy-makers are likely to be more interested in having changes interpreted in terms of characteristics of student populations than in terms of properties of IEA items.

Understanding curricular effects

A second challenge confronting IEA studies is that of comparing measures of achievement across education systems. Such comparisons can be made meaningfully only if an instrument functions in essentially the same way in the different systems in which it is used. In recent IEA studies, item response models have been used as part of the process of checking on how instruments function in practice. Here again, Goldstein confuses assumption with intention:

> The assumption underlying most of the psychometric models for test item responses is that the characteristics of a test item, for example its difficulty or discrimination, are constant and uninfluenced by different contexts.

Item response models do not 'assume' items operate in the same way in different contexts. In fact, it is precisely because items may not operate in the same way that a model is required. The model specifies the requirements items must satisfy if they are to form an instrument to provide measures that can be validly compared from context to context.

In IEA studies, no attempt is made to provide comprehensive coverage of any one country's intended or implemented curriculum. Rather, an attempt is made to assess an agreed core of learning, common across countries. This means that IEA measures of, say, physics achievement may not adequately represent everything that students in a particular country learn in physics. Instead, IEA physics achievement is defined in terms of a central corpus of knowledge and skills agreed on by participating countries. A question can then be asked about how well this agreed content is being taught and learned in different countries.

Goldstein claims that a problem with almost all attempts to produce unidimensional scales is that 'curriculum factors' can cause items to be highly intercorrelated and so appear consistent with the intention of unidimensionality. He argues for 'partialling out' curriculum factors before drawing conclusions about dimensionality.

Scales for measuring educational achievement inevitably (and appropriately) reflect curriculum influences. In IEA studies, each achievement scale is in part a reflection of the way in which the relevant curriculum typically operates in participating countries. It would defeat the purpose of the measures to remove the influence of curriculum from an *achievement* scale. And, while it is possible for differences in curricula to raise artificially intercorrelations among items (Masters, 1988), the more usual consequence is to reduce intercorrelations.

Reporting achievement measures

A third challenge is that of deciding how broadly IEA measurement variables are to be defined. Policy-makers are often interested in broad achievement measures. They may be interested in how students in different countries perform in 'mathematics' or 'science', for example, but be less interested in the details of performances in such areas as 'algebra', geometry', 'numbers', 'chemistry', 'physics' or 'biology', and even less interested in more narrowly defined areas of achievement.

Attempts to construct and measure broadly defined achievement variables such as 'reading' must be evaluated empirically. But there is no reason in principle why valid measures of 'reading' might not be possible. The question is whether the items developed to measure this variable work together well enough to provide meaningful measures. If they do not, then it may be necessary to construct measures of different kinds of reading. Goldstein is incorrect when he says that

> the very notion of reporting comparisons in terms of a single scale, for example of 'mathematics' or 'science', is misleading.

There is nothing 'misleading' about the notion of a single 'mathematics' or 'science' scale. The relevant question is whether instruments can be constructed to provide useful measures on such broadly defined variables.

In the IEA reading literacy study (Elley, 1992) a decision was made *not* to construct measures of overall reading achievement, but to construct three separate measures: narrative reading, expository reading and document reading. It is possible now to ask whether the items developed for those scales do define different dimensions, or whether they in fact work together well enough to support the notion of a single 'reading' variable. Some recent work by Ingrid Munck suggests that while items developed for narrative and expository reading appear *not* to define different dimensions of achievement, document reading is probably usefully reported as a separate dimension.

This is an example of IEA's use of specific tests of dimensionality. Here again, Goldstein is wrong in suggesting that only global 'goodness of fit' tests are used to investigate the dimensionality of IEA instruments and that

> the statistical procedures themselves are far from satisfactory. General so-called 'goodness of fit' tests provide weak evidence for confirming unidimensionality unless they are concerned with contrasting a unidimensional structure with a specific multidimensional structure [...] 'non-specific' tests will often fail to detect a real multidimensional structure.

Where an explicit hypothesis exists concerning the dimensionality of a collection of items, that hypothesis can be, and is, tested explicitly in IEA research.

Interpreting achievement measures

A fourth challenge is that of deciding on the extent to which IEA achievement measures can be generalized beyond the particular sets of items used in a study. One of the advantages of using an item response model in the construction of a measuring instrument is that when the conditions specified by the model can be met, measures made with the instrument do not depend on the specifics of its items. Nor do they depend on the mix of item types within the instrument. Thus if a test contains three superficially different types of items, A, B and C, but students' performances on all items conform to the model, it does not matter what the proportions of items A, B and C are. (There may be other reasons – such as face validity – to include a particular mix of A, B and C items, but this will have no influence on the measures themselves.) In this sense, the measurement scale has a special meaning: it is freed of the irrelevant details of individual items and of the relative proportions of item types that make it up.

Goldstein despairs of the possibility of constructing measurement scales with this general meaning and states that

> 'any overall scale is best viewed as a particular weighted average of its separate components with no other special meaning.'

If IEA achievement measures depended on the particular mix of items used in an instrument, then there could be as many different measurement scales (and, potentially, orderings of students and countries) as there are mixes of items possible in an instrument. Only by ensuring that measures made with an instrument do *not* depend on the mix of its items can measurement scales with any general meaning be constructed. Item response models specify conditions which, when met, make such scales possible. This was educational measurement's major breakthrough in the twentieth century – a breakthrough that Harvey Goldstein seems not to understand.

References

ELLEY, W. B. 1992. *How in the World do Students Read?* Hamburg, IEA.

MASTERS, G. N. 1988. Item Discrimination: When More is Worse. *Journal of Educational Measurement,* Vol. 24, pp. 15–29.

Reply to the critique

Harvey Goldstein

Geoff Masters has responded to those parts of my paper dealing with the use of item response models for test construction and analysis. Essentially, he uses the standard justifications for such procedures in terms of their supposedly desirable properties. Unfortunately, he has failed to address my general concern that a heavy reliance of such models may distort and undermine comparative studies of educational achievement.

Much of Masters' argument is centred on the concept of 'unidimensionality', which is used in two distinct senses. First, it refers to the summarizing of different aspects of achievement in a single score or rating. I refer to this as a unidimensional summary in my Conclusions, and I am concerned not with the measurement process itself, but simply with the inherent loss of information associated with averaging. I devote the section on 'Aggregated Scales' (p. 20) to pointing out why it is inefficient and misleading to average in this way. Masters has misunderstood my point here and fails to address this issue.

The second meaning of dimensionality is in terms of the response which students make to the items or questions in a test. Broadly speaking, if the responses to the items correlate very highly we can adequately summarize these responses by a single 'factor' score for each student. It is *as if* students are responding along a single underlying dimension. There are several difficulties with such a summary measure, however. I point out in the section on 'Limitations of Item Response Scaling' (p. 22) that it is population-dependent so that, for example, what is unidimensional in one population may be multidimensional in another. I also point out that increasing the unidimensionality of a test may be at the expense of test validity.

Masters' argument that IEA *intends* to create unidimensional tests implies, as he admits, that they wish to have tests where there are very high correlations between items. I do not dispute, of course, that one can create tests with this property; my concern is with whether the results are necessarily useful. Masters gives an example for a biology test where a particular item is a better test of reading comprehension than of biology and points out that such an item should be removed. Naturally, if criteria could be established for deciding how well an item reflects achievement in biology as opposed to reading, then this is quite acceptable, but has nothing particularly to do with dimensionality. To illustrate the problem, suppose we have two kinds of biology items – A and B – where the B items are in a minority. Our reading item might in fact correlate more highly with the A items (e.g. these items might involve understanding written descriptions) than any of the B items correlate with the A items. An Item

Response Model analysis would tend to link the reading item to the A items because it will reflect the pattern of relatively high intercorrelations among these items. The resulting scale would tend to exclude the minority set of B items because these do not correlate highly with either the A items or the reading item. Thus, a rigid application of Item Response Modeling will produce an inappropriate result. I think we would both agree that educational judgements should be given precedence when studying item suitability; my argument is that this may be thwarted by reliance on Item Response Model analysis. Masters has also misunderstood the argument in this section where I point out that strict application of such procedures to produce a unidimensional test will, by and large, reflect the balance of items chosen initially by the test constructor.

I am a little puzzled by Masters' apparent hostility to presenting results at a disaggregated level. It is only if there really does exist a single unidimensional scale that disaggregated reporting becomes unnecessary. In view of all the problems with such a concept, it would seem unwise to rely on this.

Masters has failed to understand my discussion of the problems of measuring trends over time. The basic situation is really quite simple: if you have a common set of 'anchor' items in two tests at different times, then there is no way of knowing whether changes in the anchor item difficulties over time really reflect changes in the population of students, with the item characteristics in some sense remaining invariant, or whether the student populations are equivalent and the items have changed, for example by becoming less relevant to the achieved curriculum in schools. This is a purely *logical* difficulty which Item Response Modeling cannot cope with. The same logical difficulty is inherent with comparisons over systems, as I point out in detail in the section on 'Time Trends' (p. 23).

Towards the end of his critique Masters reiterates a standard claim on behalf of Item Response Models: 'when the *conditions specified* by the model can be met, measures made with the instrument do not depend on the specifics of its items. Nor do they depend on the mix of item types within the instrument' (my italics). My argument is that these conditions typically are not met and that if you think that they are or you attempt to impose them, then you may be seriously misled. It is just such a belief that leads Masters to claim that 'the measurement scale has a special meaning: it is freed of the irrelevant details of individual items and of the relative proportions of item types that make it up.' This statement ignores all the research about the importance of item context, wording, ordering, etc. and to which I refer under 'Time Trends'.

Finally, I am concerned that, as an influential voice in IEA, Masters' views about assessment may distort the traditionally eclectic approach of that organization to the process of constructing measurements and interpreting results. I sincerely hope that IEA does not allow itself to be side-tracked by simplistic views of what educational measurement is all about.