Oxford Review of Education, Vol. 29, No. 2, 2003

Cautions on OECD'S Recent Educational Survey (PISA)

S. J. PRAIS

ABSTRACT *A new survey of the educational attainments of 15-year-olds was undertaken by OECD in Spring 2000 (the 'PISA survey'). Surprisingly, British pupils appeared to perform in mathematics much better than in an IEA survey carried out only one year previously. This paper examines four main differences in the objectives and methods adopted in the two surveys. (a) Questions in the previous IEA survey were directed to the mastery of the school syllabus by the relevant age-groups, whereas PISA was ostensibly directed to so-called 'everyday life' problems—which provides less guidance for policy on schooling. (b) The IEA survey was based on samples of whole classes including, for example, older pupils who had entered school late, or had repeated a class: PISA excluded the latter pupils as it was based strictly on a 12-months' period of birth; issues of variability of pupils' attainments within a class—important for a class's teachability—cannot therefore be examined in this OECD survey. (c) England's response rate for schools was particularly low (60%, compared with 95% in leading European countries), raising serious doubts as to the inclusion of low-attaining schools. (d) The response rate of pupils (within participating schools in the PISA survey) was lower in England than in any other country, and lower than in the previous IEA survey, suggesting a greater upward bias in reported average scores. The paper concludes that it is difficult to draw valid conclusions for Britain from this survey and planned repeats should be postponed until the underlying methodological problems have been resolved.*

At the beginning of December 2001 some surprising 'first results' were published from a new sample survey of the educational attainments of 15-year-olds in some 30 mainly OECD countries, including the United Kingdom; the tests, carried out as recently as spring of the previous year, covered literacy, mathematics and science. Immense resources had been invested in carrying out and analysing the results of this survey—but not, in my view, in fully thinking through its purpose and design. The first path-breaking international educational surveys were carried out by the International Association for the Evaluation of Educational Achievement (IEA) in the 1960s in a dozen countries, and thereafter at about ten-year intervals in an increasing number of countries—most recently in 1995 and 1999 in nearly 40 countries. This new OECD-sponsored survey—the Programme for International Student Assessment, or PISA for short—included countries ranging from those still developing socially and economically (Mexico, Brazil) to the most economically advanced (such as the United States and Switzerland).

From the UK's point of view, the results of this latest survey appeared gratifying, indeed remarkably so; previous international surveys of schooling had fairly consistently

ISSN 0305-4985 print; ISSN 1465-3915 online/03/020139-25 © 2003 Taylor & Francis Ltd
DOI: 10.1080/0305498032000080657

pointed to low average attainments by UK pupils, combined with a wide dispersion—particularly, a long tail of low-achievers—when compared with leading European industrial countries. From this latest survey it appeared that average pupils' mathematical attainments in the UK had caught up even with Switzerland, hitherto seen as a sort of model educational system within the European context, to which England might aspire only distantly; and had overtaken France, Germany and Hungary—but not Japan or Korea which remained well ahead. Further, the variation among UK pupils' attainments in this new survey was narrower—we now apparently have a smaller proportion of low-attainers than Switzerland. Similar findings were reported in PISA for questions in reading-literacy and science [1].

The previous IEA survey was carried out in 1999—only a year before the OECD survey; further, the IEA survey of 1999 was based on 14-year-olds, while the OECD survey of 2000 was based on 15-year-olds. Both surveys were thus based on sampling from much the same cohort of school pupils, those born in 1984. That such contradictory results should be found in relation to the UK raises serious questions: which survey, if any, are we to believe?

In our attempts below to trace the sources of these apparently rapid changes in favour of the UK, it will be helpful to have in mind a measure of the size of the change. Broadly speaking, there has been a net shift of some 60 points: whereas the IEA surveys of 1995/9 showed the UK at some 40 points in average scores *behind* Switzerland, France, (Flemish-speaking) Belgium, the Czech Republic and Hungary, in the PISA study of 2000 the UK was some 20 points *ahead* of those countries on average (on the standardised measuring rod used in these surveys, 100 points corresponds to one standard deviation of the scores of individual pupils) [2]. Most of that apparent improvement took place in a single year, between 1999 and 2000!

If these findings can be accepted, then successive governments' educational policies in this country—from the imposition of a National Curriculum in 1988, via the publication of competitive 'league tables' of schools' results in public examinations since the mid-1990s, to the shift in emphasis towards 'basics' culminating in the National Numeracy and Literacy Strategies starting three years ago—must be judged as having achieved remarkable success; and in a remarkably short time. Educational policy makers in this country can therefore now both celebrate and relax; and overall policy priorities might properly be shifted (less of tax-payers' money to education …).

Fuller 'technical reports' on the conduct of the PISA survey were published some 6–9 months later (Summer/Autumn 2002), and we have now reached the stage where it is possible, indeed necessary, to delineate certain worrying areas of doubt attached to the results [3]. The doubts arise from, first, differences in the precise *objectives* of the surveys, namely, (a) the nature of the *questions* asked of the pupils in the PISA survey, in contrast to previous surveys; and (b) differences in the intended *age-group* covered in the surveys. A second set of doubts arises from apparently imperfect *execution* of the PISA survey and the consequent degree of reliability of the sample results, namely, (c) the representativeness of the English *schools* agreeing to participate; and (d) the representativeness of the *pupils* actually taking the test within each participating English school. A third set of doubts—explained in an Annex—relates to the 'black box' complex computerised arithmetical processing of the results of questioning, with perplexing differences in the ranking of countries according to whether we look at the percentage of questions answered correctly or at what has come out of the 'black box'.

While the stimulus for the present paper came largely from the need to examine the UK's anomalous apparent rapid rise in educational success in PISA, we shall see that

the worries raised are of wider application, and point to the need for wider basic research into the methods used in such surveys before they are repeated.

(a) NATURE OF THE QUESTIONS ASKED

We concentrate here on the mathematics questions in the survey, partly out of regard for readers' patience, and partly because previous international comparisons of English pupils' attainments have been particularly clear in mathematics, and seriously worrying for this country in relation to its workforce's technical skills. A smaller weight was given to mathematics in this PISA survey as a whole (as said, it also covered reading and scientific literacy); a subsequent PISA survey planned for 2003 is to give greater weight to mathematics—but OECD apparently thought that sufficient mathematics questions were asked in this round for it to compile an international 'league table' of pupils' performance on its 'mathematical literacy scale' [4].

Previous surveys of school pupils' mathematical attainments by IEA (1964, 1981, 1995, 1999; as also by IAEP in 1991) devoted much discussion to the school syllabuses of pupils at various ages: there was much discussion of differences between syllabuses as *intended* (centrally specified), *implemented* (as actually taught) and *attained* (as learnt); and much discussion of what might be lost in such international testing by concentrating on topics common to the syllabuses of most countries participating in the survey, while under-emphasising any concentration on specialist topics in some countries (in the mathematics syllabuses, say, on three-dimensional conceptualisation at primary-school ages, solid geometry and trigonometry at secondary-school ages). This latest PISA survey decided to side-step such issues by focusing on pupils' so-called 'Knowledge and Skills for Everyday Life', or so-called 'mathematics literacy': the stated focus was ostensibly distinct from details of the school curriculum, and was intended to elucidate how pupils might cope in real life with the help of what they have learnt [5]. In practice, as we shall see, this was not really so.

The approach adopted in PISA was similar to that of the OECD's *International Adult Literacy Survey* which was published the previous year and was based on a US precedent. Without wishing to go here into limitations of that survey (too long a questionnaire, poor response rate, …), it has to be said that what may have seemed a worthwhile objective for a survey of *adults*, who had been out of school on average for some 25 years, is not necessarily a worthwhile objective for a survey of *school* pupils, where curricular and educational policy-making are very much at stake [6].

An indication of the seriousness of the issues associated with the central objectives of the survey is to be inferred from what happened in Germany. In that country, in addition to the internationally specified questions set to their international mathematics sample of 2500 pupils, an 'extended' *(erweitete)* twenty-times larger sample (the 'PISA-E' sample of 48,000 pupils) was tested; it was so very much larger to permit more reliable comparisons between different parts of the country—some *Länder* having gone much further than others in the last generation or two in schooling reforms, such as introducing comprehensive secondary schools and more 'child-centred' teaching methods. For that extended sample of pupils—and this is the important aspect for our immediate concerns here—entirely different mathematical questions were set to reflect better (a) the actual *school curriculum* in Germany and (b) the *spread* of attainments of German pupils. In contrast to a total of 31 questions in mathematics in the international inquiry, the German national extended inquiry had an additional 86 mathematical questions with a generally greater computational constituent (of which a

quarter were of a direct arithmetical, 'algorithmic' or 'technical' type, instead of only 3% of such questions in the international PISA inquiry) [7].

The German schooling system had long emphasised accuracy and speed in arithmetical (and mental arithmetical) tasks, especially for its less academic tiers of secondary school pupils—corresponding roughly to those pupils in England taking the lowest of the three tiers of the GCSE mathematics tests. Since PISA had under-emphasised such questions in its *international* tests, German teachers felt that the true mathematical achievements of large tranches of their pupils would be significantly understated: comparisons between various *Länder* would not be soundly based, and of little validity [8].

Two important inferences can already at this stage be drawn in assessing the significance of the PISA inquiry. First, in comparisons provided by the international PISA inquiry, we must not be too surprised if German pupils—like those of other countries with similar emphases on computational competency in their mathematical curricular objectives, such as Switzerland—did not do as well in relation to Britain as in previous international educational surveys (such as the immediately preceding TIMSS surveys). Secondly, we have to take rather seriously PISA's stated objective that they were hardly at all concerned with testing mastery of the school curriculum, but with how successfully pupils might cope with 'everyday life' post-school situations. The difficulty with considering the latter to be a worthwhile objective in surveys of 15-year-olds is precisely that pupils are still in course of being *prepared* for after-school life—and that must depend to a varying extent on the expected age of completion of schooling. Even for the two-thirds of all pupils who leave full-time schooling in Germany at 15/16 to enter an apprenticeship (and similarly in Switzerland, Austria; and the somewhat lower proportion in France, The Netherlands, etc.), part-time attendance at mathematics courses remains obligatory at ages 16–18. But this is not so in Britain, where mathematics is obligatory only till 16 (the 'GCSE year'). The consequence is that the final years of obligatory *full-time* schooling in the Continental countries mentioned can concentrate more on deepening the foundations of mathematical knowledge, leaving 'everyday life' application to subsequent apprenticeship years; while in Britain those final years of full-time schooling, at 15–16, may be expected to deal less with deepening foundations than with extending applications. Similar considerations apply to US results for schooling attainments measured at age 15, when schooling generally continues there for, say, three years longer than in Europe. An analogy: two building sites are compared, both the same length of time after starting; one shows no building above ground-level and only a great hole for foundations, while the other shows three storeys plus a roof nearly completed. The deeper foundations of the former will enable it ultimately to be higher. Simply comparing heights at an early stage provides little guidance to ultimate attainments.

Let us now look briefly at some illustrative questions that have been released:

Example A A pizzeria serves two round pizzas of the same thickness in different sizes. The smaller one has a diameter of 30cm and costs 30 zeds. The larger one has a diameter of 40cm and costs 40 zeds. *Question*: Which is better value for money? Show your reasoning.

This seems close to 'everyday life' but, even so, there are limitations because of the unrealistic simplicity of the numbers [9].

Let us now look at two further published illustrative examples:

Example B The approximate relation between the diameter (d, measured in

mm) of a small plant (called a lichen) and its age (t, measured in years) is $d = 7\sqrt{(t - 12)}$. Ann found a lichen of 35 mm diameter. *Question:* What is its approximate age?

Example C A seal typically comes to the surface of the water to breathe; it then dives to the bottom and rests there for 3 minutes; then slowly floats to the surface in 8 minutes; takes a breath, and repeats the cycle. *Question:* Where was the seal 60 minutes after having come to the top to breathe?

Questions B and C do not seem at all close to ordinary problems of living; nor is it obvious in what sense they test 'mathematical literacy'—which the authors of PISA set as a prime objective (as distinct from testing mastery of 'school mathematics')—though without providing adequate justification for that distinct objective [10].

Other specimen questions that have been released attach much importance to *reading* values from a graph (not *constructing* a graph); this is presumably seen as an important part of 'mathematical literacy'—but the importance of complex graph-reading deserves debate. Let us look at a relatively difficult graphical question:

Example D A graph shows the fluctuating 'speed of a racing car along a flat 3 km racing track (second lap)' against the distance covered along the track. Pupils were assumed to know that the track was some form of a closed loop; they were also assumed to know that normally there are various bends along the track, and that speed had to be reduced before entering each bend. Pupils were required to read from the graph the 'approximate distance from the starting time to the beginning of the longest straight section of the track', and similar matters; they then had to match the speed-distance graph with five possible track-circuit diagrams.

Perhaps these are perfectly reasonable assumptions and reasonable questions in relation to Australian or German boys—but for girls in rural Greece or Portugal [11]? Answering such questions correctly may be more a test of 'common sense', or of 'IQ', than the results of mathematical schooling at this age.

It seems all too clear that many of the PISA questions in mathematics were not of the stated 'everyday life' type; indeed as much as anything they do little more than confirm that there are great difficulties in framing 'real life' questions equally appropriate for candidates from such widely disparate social backgrounds. Were there then some other guiding principles? The most readily available official reports—namely, the OECD summary report on all participating countries, and the special report with fuller results for England—provide little hint on any further general principles. However, someone persistent enough to examine the very extensive German reports (pp. 548 + 254!) will be rewarded by a discussion of the approach to teaching 'realistic mathematics', associated with the name of the late Professor Hans Freudenthal of the University of Ultrecht, which is said to have heavily influenced the kind of question asked [12]. The teaching of school-level mathematics starting from realistic contexts is of course normal in many countries (whether England, Japan or Switzerland), though there is considerable variation in the range of contexts, and in the degree to which the problems reflect true 'real life' complexity [13]. Associated with the 'realistic mathematics' approach there has more recently developed an emphasis on 'mathematical conceptualisation', or 'mathematisation', as the essence of sound mathematics teaching; that is to say, that the recognition of the 'mathematical structure' of a newly-met

'realistic context' is of the essence of mathematical mastery, rather than, say, demonstrating computational competence in a familiar application.

It is this emphasis on testing 'conceptualisation' abilities that seems to have played an unduly large role in the choice of PISA test questions; the Germans conceded that such questions would be suitable for their top-tier (*Gymnasium*, or 'Grammar School') pupils, but regarded them as largely unsuitable for the majority of their pupils (particularly their lower-tier, *Hauptschule* pupils). Consequently they developed a separate set of questions for their tests within Germany of their larger ('extended') sample of pupils. That reservation—as to the kind of mathematics question properly to be asked of the full ability-range of pupils of that age—must, to varying extents, have applied in other countries as well.

The whole of this broad issue of principle—of the proper prime objective of such a survey, and especially its change from previous international surveys—deserved wider discussion at the outset. Previous IEA surveys of 9- and 13–14-year-olds included some very basic questions, bearing directly on the school syllabus, such as

4000–2369.

This was a multiple-choice question with four possible answers, one of which was to be ticked. The results are worth repeating here: at secondary-school level at age 13 (in the IEA tests of 1995) some 90% of pupils in five Western European countries were able to answer the above subtraction sum correctly; but only under half of all pupils in England! At age 9 the same question was asked in a parallel IEA primary school survey, with only 15% of English pupils answering correctly, compared with 90% in neighbouring European countries [14]. The English problem with arithmetic thus starts at primary schooling, and is not resolved during secondary schooling.

At age 14, in the 1999 IEA tests in mathematics, a similar question was set:

7003–4078.

English pupils performed worse than in 1995, with only a third answering correctly (after adjustment for guessing) [15]. If anything, some pupils had forgotten at age 14 what they knew at age 13! Of the 38 countries participating in those tests, England's score on that question was only one place above the weakest country.

Britain's extraordinarily poor performance in such basic arithmetical questions—demonstrated in those surveys at both primary and secondary school ages—undoubtedly contributed to the recent great changes to Britain's schooling in this area of the curriculum, with subsequent greater emphasis on numeracy and mental calculation, especially in primary schooling. It might have been hoped that policy guidance related to more advanced levels of school mathematics would emerge from questioning 15–16-year-olds in the PISA survey; but this seems unlikely, judging from the stated objectives of the PISA survey.

If we are to summarise the foregoing reservations based on *differences in the kind of question* asked in this inquiry—and ignore for the moment other limitations of the survey to be discussed in the remainder of this paper—perhaps we may do so in these terms: the kind of mathematics questions asked in PISA were deliberately different from those in earlier surveys, and were ostensibly *not* intended to test mastery of the school curriculum; they can, perhaps, be said to be nearer to tests of common sense (or of 'IQ'). No one has ever doubted the common sense of the British people, nor of British youngsters: the issue has been whether they are as well served as they might be by their schooling system. For that purpose, the verdict of the previous IEA inquiries—

which were focussed on the school curriculum—still needs to be accepted; and the results of the new PISA tests in no way modify that verdict [16].

(b) WAS THE AGE-GROUP COVERED EQUALLY IN ALL COUNTRIES?

Surprising as it may at first seem, the very definition of the *age-group* of pupils covered by the PISA survey presents problems of international comparability—unfortunately barely discussed in the survey's international report, nor in the British national report (these issues are discussed in the German and Swiss national reports—but, as said, are not easily accessible to English readers). Underlying those problems of comparability of the survey's age-groups is a school-organisational issue dividing British schooling from that of most other European countries, namely, whether pupils enter school according to a strictly defined twelve months' period in which they were born; or whether there is flexibility in that placement, according to a child's rate of maturation (say, four months on either side of a 'normal' yearly period), with subsequent possibilities for class-repetition, class-skipping, etc. A first set of problems is related to sampling based on that organisational distinction. A second set of age-related problems arises from the older age-group chosen for the PISA survey—age 15–16 rather than 13–14 as in preceding IEA surveys. Greater problems of pupils' participation ensue from that older age since it is closer to the end of compulsory schooling, with some pupils already having left school, and others too busy to participate in the survey because of imminent school-leaving examinations (in the case of England, GCSE). We also discuss, at the end of this section, another definitional problem of the population to be surveyed: the inclusion or exclusion of special schools.

Age-Grouping by Schools

Previous international surveys of pupils' attainments generally sampled *whole classes* of pupils, for example, Year 9 classes, which in England consisted of those aged 14 + , while on the Continent the correspondingly-aged class ('international eighth grade') would generally *include*—in addition to those of that age—some older pupils lagging in their schooling for whatever reason, and *exclude* some lagging pupils aged 14 now in a lower class. Sampling a whole class obviously causes less disturbance to a school's time-table—and hence better participation of sampled schools—than choosing pupils from several classes on the basis that they were born in a precisely-defined twelve-months period. Sampling by classes also permits examination of the degree of variation of attainments *within* a class—a crucial factor in ease of teaching and learning. It was the basis of the earlier IEA surveys (as also of the IAEP survey of 1991) [17]. The new PISA survey, in contrast, defined the group to be sampled as those born in a particular twelve-month period (calendar-year 1984), irrespective of which class they were in at the date of the survey.

For England, of course, the difference is of no consequence since school-classes are tightly based on date of birth. To labour the point, since an important pedagogical policy principle is at issue, and not merely a statistical sampling technicality: born on 31 August, a child is required to enter school in England the day following his or her fifth birthday, and then rises from class to class 'automatically', year by year; born one *day* later, the child waits another *year* before entry. Not so in most Continental schools, where a slow-developing child born on the former date might enter school a year later than the 'normal' defined school-year, while a fast-maturing child could enter a year

earlier (limits to such exceptions vary from country to country and need not be pursued here). For countries where the date of entry is flexibly dependent on a child's maturity, etc., there is a clear difference between the population of pupils intended to be covered in the PISA approach, and that covered in previous international educational surveys. The Germans, as also the Swiss, thought it worth taking a supplementary sample so as to be able to report on the attainments of a single school grade (their ninth grade pupils). In Germany, of 15-year-olds sampled for the international PISA survey, only 64% were in their 'normal' school year; and in Switzerland, 75% [18].

At first thought it might seem that the PISA approach (sampling 15-year-olds rather than ninth graders) provides a more rigorous international comparison, in that participants are then all of closely comparable age; but to accept that view would ignore the pedagogic motivation of those countries which group pupils into classes having regard to their maturation—whether intellectual, educational, emotional, physical (in reality, some judgmental combination of all these aspects). To put it another way: since pupils vary in their rates of maturation and rates of learning, it may well be educationally both more equitable and more efficient for some children to spend more time on their schooling than others (or to spend the same standard number of years at school but aged one year older, or younger, than the majority). To choose an arbitrary single *age* for all pupils during their schooling—as here, an age at which most pupils in Continental classes are in a single grade, but others of that age are one or even two grades lower, and then publish a single average score for all pupils of that age in a country, would do an injustice to that country's schooling processes considered as a whole. It would ignore entirely any gains in attainment by, for example, slower-maturing pupils in later school-years.

As it happens, for Belgium (Flemish) PISA published average scores in their tests of reading literacy for 15-year-olds according to which school *grade* they attended, showing very substantial variations: tenth grade, average score of 564; ninth grade, 455; eighth grade, 364 (the proportion of all 15-year-olds in each of these grades was 72, 23 and 2.5% respectively; the balance of 2.5% was in other grades, presumably including those who had skipped a grade) [19]. The weighted average for *all* 15-year-olds (whichever grade they were in) was 532, that is to say, about 30 points (or a third of a standard deviation) below the average score of 564 for pupils of that age in their 'normal' class. In estimating the likely average score for a class chosen on the IEA 'whole class' basis, we need also to make some allowance for those older pupils who have repeated a year (entered school late, etc.). On the one hand, such pupils showed themselves unusually weak in earlier years and might be expected to continue to be weak; on the other hand, they have had an extra year of schooling to help them catch up, and some have no doubt overtaken the average of their class. It is not easy to know which effect is more important (it is a pity this inquiry did not investigate this important issue). I would guess the former effect is slightly more important, and would thus hazard that the difference in calculated average scores that would arise according to whether we sample pupils by age (on the PISA-basis), or according to classes (on the IEA-basis), might be something like 20 points.

The UK's average scores would not be affected by this issue (since, as said, classes are formed strictly according to calendar-year of birth); but Continental countries, such as those mentioned in the opening paragraph to this Note, might show a fall in their PISA scores (as compared with the preceding year's IEA scores) of, say, some 20 points simply because of PISA's definition of the age-group (i.e. inclusion of 15-year-olds in lower grades). As will be remembered, we are looking for explanations for the relative

60 points' improvement in the UK scores in PISA: it seems possible, therefore, that about a third could result from the definitional change in the age-group—a kind of 'optical illusion' without any underlying real change in pupils' educational attainments.

In coming to a view on the important pedagogical issue of the benefits of flexibility in calendar-age in class-organisation, it may be as well to recall here that one of the important findings of the 1995 IEA mathematics tests was that the weakest 14-year-olds in England achieved scores not only below the corresponding proportion of pupils in Swiss (and other Continental) *eighth* grade classes, consisting mostly of 14-year-olds, but also below the scores of the corresponding proportion of Continental *seventh* grade classes consisting mostly of 13-year-olds [20]. That finding seemed to support those who hold that pedagogically it is better to attack the problem of slower maturation and low attainment by putting such pupils into a class of younger pupils, closer to those of similar *educational* attainments, and not to be rigidly tied to pupils' calendar age. Unfortunately little recognition can be found in the PISA report of the central pedagogical significance of such an issue in schooling organisation, and the light that a survey of this kind might cast on it, notwithstanding that the organisers had made a deliberate change in the basis of the age-group to be sampled.

Older Age-Group Chosen for PISA

We must next notice that the age of pupils covered in the PISA survey was about a year older than in previous surveys, being age 15—more precisely, in most countries 15:3 (read: 15 years and 3 months) to 16:2, in England 15:2 to 16:1—rather than ages 13/14 as, for example, in the previous IEA survey. That higher age in the PISA survey leads to fundamental problems for the sampling-coverage of the whole attainment-spectrum: we must expect poorer participation rates from those of lower academic attainment, and consequently greater upward bias in the measured average score for each country.

Previous surveys decided on younger ages of up to 14 because pupils in virtually all countries were still in obligatory schooling, and could therefore be reached satisfactorily through a sample of *schools* (rather than, say, a sample of *households*, which would be much more expensive); the reason PISA chose an older age was (presumably) that it would bring the sample nearer to the age of facing the problems of everyday life—a stated focus of their survey. That older age is, however, a 'transitional year' for many pupils: depending on the schooling requirements of the country in question, some pupils leave school before, or in the course of, that older year; others attend school only intermittently, some are in employment at the date of the survey, others are unemployed (in Brazil and Mexico, for example, 47% and 48%, respectively, were no longer enrolled in school at age 15). It is an age that, in my view, is difficult to sample representatively on the basis of schools—and amounted, in my view, to an error of judgement for an international survey of this sort, where full coverage of academically weaker pupils is important if any reliance is to be placed on calculations of average attainments and of the proportion of under achieving pupils.

The issue also affects developed economies. Let us contrast England, with its obligatory full-time schooling till 16, with Germany or Switzerland where obligatory full-time schooling is often—depending on the region (*Land, Kanton*)—only up to age 15 (it was 14 in Italy until very recently). Most pupils in England in the relevant age-group are in Year 11, the final year of obligatory full-time schooling when GCSE examinations are taken, and when schools and pupils are reluctant to take time off for

non-essential activities. More precisely, the age-range just quoted was specified to relate to the beginning of the 'assessment period' for this survey, say, close to April 2000 in England. Pupils in Year 11 were then aged 15:7 to 16:6, of whom the oldest third in that school-year (those aged 16:2 to 16:6) lay outside the defined PISA age-range; on the other hand, the oldest third of pupils in the immediately preceding Year 10 (namely, those who at the survey date were aged 15:2 to 15:6 inclusive) lay within the defined PISA range.

The simply-stated intention of PISA was to cover pupils born in the calendar year 1984; but the school-year does not usually correspond to the calendar year. The consequence was that *two* school-years had to be sampled, with greater disturbance and lower co-operation. Further worries ensue from considerable absenteeism in English schools in the final years of obligatory schooling, particularly after pupils reach the age of 16; the oldest sixth of those in the defined PISA age-range (those aged 16:0, 16:1) are particularly liable to be absent. Current legal provisions in England require that such pupils attend school till the end of that school year; in practice, the authorities mostly see little point in putting that right beyond formal letter-writing. The upshot is that we must expect a substantial fraction of pupils from that oldest part of the defined PISA age-group to be missing from the English sample (say, 10–15% of the whole sample not reached, mostly from the weakest academically). It must also be said, on the other hand, that since Year 11 is the year in which GCSE ('school leaving') tests are taken, some high-aspiring pupils may absent themselves from a voluntary test in order to concentrate on revision for GCSE (or are sitting GCSE oral tests in modern foreign languages, which are often held at that time of year). On the whole we must suspect that the former factor is more important, and that the attained PISA sample in England was biased upwards in its reporting of educational attainments [21].

The issue of response bias due to pupils having already left school, by the date of the survey, may be expected to be even more serious in countries where 15 is the age when obligatory full-time schooling ends as, for example, in many parts of Switzerland and Germany; but much depends on the survey's coverage of part-time vocational colleges [22]. Vocational colleges in these countries were sampled in respect of their 15 year-olds; pupil-response rates have been published for Germany, but reached only a disappointing 44% (compared to 86% for the inquiry as a whole in that country) [23].

In other countries many pupils leave school at the earliest possible moment and, at the time of the survey, may be in unskilled employment or unemployed. Youngsters not enrolled in some form of schooling were said by PISA to be *definitionally* excluded from the coverage of the survey. Such exclusions clearly affect the survey's recorded variations between countries of pupils' average attainments; but simply excluding such youngsters 'by definition' does not, of course, eliminate the underlying lack of comparability in the samples chosen to represent each country.

We see that the year-older age-group chosen as the focus for the PISA inquiry (as compared with previous IEA inquiries) has brought with it a penumbra of uncertainty due to non-respondents only nominally still at school, and due to the 'definitional' exclusion of those who have left schooling. That penumbra is unlikely to have been of equal size in different countries; and it is not clear whether, considered as a factor on its own, it artificially raised the recorded attainments of British as compared with Continental pupils. The more general issue of differences in countries' response (participation) rates of the randomly selected samples of schools and pupils will be seen below to be more important; the age-definition issue can be regarded as a specific factor contributing to those overall differences in response.

Were Special Schools Included in the PISA Survey, or Excluded from it?

Test-questions prepared mainly for pupils in mainstream schools are not, on the whole, appropriate for the kind of pupil educated in Special Schools. The latter are thus usually omitted from inquiries of this sort and, as such pupils usually account for only a small percentage of all pupils, not much is lost by that exclusion. The English PISA report put it this way: 'As the PISA assessment was not intended for students with special educational needs, for whom the assessment is likely to have been too challenging, schools which catered solely for such students were *not included* in the survey' [24]. The German report on their international sample put it, paradoxically, as follows: 'In accordance with criteria specified for the international comparisons of attainments of 15-year-olds, pupils in Germany at Special Schools for slow-learners and for those with behavioural problems were *included* in the inquiry' [25]. However, for the larger national inquiry within Germany (PISA-E), such schools were *excluded*; the effect of the exclusions of such low-attaining pupils was said to raise the calculated average attainment within Germany by some eight points [26]. How it came about that two OECD countries could interpret instructions from OECD HQ in such opposite ways may be left as an exercise for the student of Kafkaism [27]. For our purposes here in trying to understand why British pupils apparently did so much better in PISA than in previous tests, we may deduce (from this reported calculation for Germany) that the exclusion of Special Schools in Britain also gave something like an eight points advantage to Britain—equivalent, say, to 13% of the 60 points gap that (as said in the opening paragraphs of this paper) we need to explain.

(c) GREATER BIAS IN THE RESPONDING SAMPLE OF ENGLISH SCHOOLS

Variability among schools in their educational success is a familiar fact of life: parents in all countries consequently attach enormous importance to avoiding a weak or indifferent school for their children (insofar as they are permitted to exercise such choice!). It is clearly of equally enormous importance that schools chosen for a sample survey, such as this, are adequately representative. With the method of identifying PISA's sample list of 180 representative UK schools, everything went well: the difficulty in the UK was in persuading those sampled schools to participate. 'A minimum response rate of 85% was required for the schools initially selected', the PISA report asserted (p. 235); this was achieved in almost all countries: for example, for France, Germany, Hungary, Switzerland, the average school-response rate was 95%; but for the UK it was only 61%! The missing schools on the whole were probably low-attaining schools; and there must be grave suspicions of upward bias in the average score of responding schools as a result of such a low response rate.

To 'compensate' for the missing 39% of non-responding schools in the original UK sample, a second sample of 'replacement schools' (from a previously prepared list, matched as far as possible to non-responding schools by school-size, geographical region, attainment level) was approached; of this second sample, 55% responded, adding 21% to the number of schools included in the final total sample. The response rate for that 'matched replacement sample' was somewhat lower than for the original sample (55% against 61%), presumably reflecting the greater problems which schools of this type have to face, and which impede their participation [28].

By how much did these replacement schools raise the 'true' sampling response rate—the rate relevant to any judgement on potential bias due to possibly different

characteristics of non-respondents? The issue caused astonishing conceptual difficulties to the organisers of PISA (as also to their statistical advisers), so let us pause a moment and put it in the form of a PISA test-question:

Example E A sample survey is to be conducted of 100 representative schools; only 60 co-operate—a proportion judged to be unsatisfactorily representative. The organisers had prepared a parallel sample of 100 matched schools to allow for such an eventuality, and they next approached a replacement sample of 40 schools, of which 20 co-operated. What was the true total response rate for judging representativeness and possible bias? Give your reasons.

(i) 80% (since we wanted 100 respondents, and got 80).

(ii) A total of 140 schools were approached and 80 responded, so the response rate was 80/140 = 57%.

(iii) The replacement sample was not a simple probability sample, and its inclusion in forming a view on likely bias due to non-response is subject to theoretical complications—better rely solely on the first sample: so the only relevant response rate is 60%.

Astonishingly, thinking among PISA's experts did not seem to go beyond answer (i) above, and so for the UK a 'school participation rate after replacement' of 82% was reported. They did not even reach consideration of answer (ii) which would have led to a more relevant rate of 57%. To their credit it deserves notice that they were not entirely comfortable, saying: 'This procedure (bringing replacement schools into the sample) brought with it a risk of *increased* response bias' [29]. On the latter issue of increased bias, they must unhappily once again be judged wrong in principle. Bias has to be suspected in the original responding sample; insofar as the replacement sample came preferentially from the difficult kinds of school that were under-represented in the original achieved sample—because the two samples had been matched on the basis of region, size, academic attainment, etc.—then the inclusion of the replacement sample might be expected to *reduce* (not increase) the response bias—though it would have been no worse to proceed by stratification and re-weighting. Serious worries therefore arise as to the statistical logic employed by the organisers of PISA [30].

There is, therefore, a worry that, because of its exceptionally low school-response rate, the UK's reported average educational achievements in the PISA survey are biased upwards in comparison with most other countries (only the US and The Netherlands were lower, with 56% and 27% response respectively). It has to be added that a low school-response rate for the UK should have been expected by the organisers of the survey since, for example, in the IEA's 1995 mathematics tests of 14-year-old pupils (TIMSS) only 56% of England's representative sample of schools agreed to participate. With such a clear precedent, more should have been done to investigate the characteristics of non-participating schools (for example, an interviewed sub-sample of non-responding school-heads to elucidate reasons for non-participation, correlation of response-rates with proportions of *low*-attaining pupils, etc.) [31].

(d) BIAS IN THE RESPONDING SAMPLE OF PUPILS WITHIN RESPONDING SCHOOLS

What has just been said about the importance of ensuring representativeness when

selecting a sample of *schools*, applies equally when selecting a sample of *pupils* within each participating school: there is an obvious worry that low-attaining pupils will avoid the tests. By the age of 15, pupils in English schools are mostly divided for mathematics lessons into several 'attainment-sets' (between three and ten 'sets', usually) directed broadly to the 'tiers' of the GCSE mathematics examination for which they are expected to enter. School attendance by this age is often very irregular in the lower attainment-sets, with some pupils attending only rarely (if at all) [32]. For the PISA test, 35 pupils were selected for each of the 155 participating schools; of the sampled pupils, an average of 81% participated in England. This was the *lowest pupil-participation rate for any of the countries* in the survey; it may be compared, for example, with an average pupil-participation rate of 92% for France, Germany, Hungary and Switzerland.

In certain large cities in Germany (Berlin, Hamburg), pupil participation rates dropped to 70% (as no doubt they did in some English cities—though no such figures have been published); to avoid an upward bias in the overall results, the returns for these German cities were ignored in calculating an overall average for that country [33].

In the IEA-TIMSS tests of 1995, England's pupil-response rate was 10 percentage points higher, at 91%. A non-response rate by pupils of, say, 3–5% may be attributed to illness and similar random factors; but at 19% in PISA there must be more than a suspicion of lower representation of weaker English pupils, and a greater upward bias in the reported results, than in previous similar surveys in England and in comparison with the Continent.

No information seems to have been gathered on the background characteristics of non-responding pupils in England: for example, to which mathematics 'set' they were attached, or how they had been assessed in their previous public mathematics tests (SATs at age 14), nor teachers' advance estimates of their likely scores in the PISA test. Advance estimates of marks in public examinations are often sent by teachers to examination boards in England: this would thus not have been an unfamiliar request for teachers. Such information would have permitted re-weighting to reduce bias. Checking the characteristics of non-respondents is sound basic routine in statistical sampling, especially when non-response is as relevant and as serious as here. If there were worries about legal restrictions on personal privacy (the Data Protection Act), it should have been possible to arrange safeguards for privacy to avoid casting serious doubts on the accuracy of such an important survey [34].

This is far from the end of the reader's worries on pupils' participation. In deciding whether a *school* should be included as an acceptable respondent, PISA decided (quite properly) to have regard to the proportion of sampled *pupils* who participated in that school: a nominal minimum 'standard' of 50% of participating pupils was proclaimed—but, in practice, a 'cut-off point' of 25% was adopted (what is the point of stipulating a 'standard', if it is at once replaced by something else?)! The latter cut-off seems unacceptably low since, as PISA recognised, if only a quarter of sampled pupils in a school participated, they are likely to be higher-attaining pupils. That is not quite the end of this curious saga: in presenting a summary figure for each country of the proportion of participating pupils, those schools with between 25% and 50% pupil-participation rate were *excluded* from that calculated average participation rate (though their pupils' responses to the test papers were *included* in the country's average test score!) [35]. The *true* proportion of pupils participating, including all schools contributing to each country's published average score (even those with 25–50% pupil participation), has so far not been revealed. The shenanigans of the Official Mind at work here raise wider worries [36].

Let us finally take together non-response by schools and non-response by pupils: in relation to our originally drawn representative sample of English schools, we found that only about 60% of schools can be counted as responding; within those participating schools, only about 80% of pupils responded. Combining these two proportions leads to the conclusion that hardly half (only some 48%) of the original representative sample for England were included in the PISA finding. On the other hand, in Switzerland for example, the corresponding combined proportion was close to 90%; and similarly for France, Germany, Hungary …. The implications for England's estimated average score in PISA depend on what is reasonable to assume for the approximately 50% missing from the English sample. A rough calculation: if we assume that half of those missing were distributed fairly evenly over the attainment range, while the other half came from the lowest end (or, at least, that the missing 50% were on the whole roughly equivalent to such a mix), then an approximate calculation suggests that the English PISA sample average score would be raised artificially (simply due to poor response) by some 38 points [37]. As mentioned, in the previous IEA inquiry the pupil-response rate was higher by some 10 percentage points than in PISA; similar calculations lead to the conclusion that PISA's average score was artificially raised; in comparison with the previous IEA inquiry, by about 5 points.

SUMMARY AND CONCLUSIONS

We may all (including the present writer) wish to believe that the immense schooling reforms in England of the past decade or so, have yielded improved educational attainments for the majority of our school-leavers. We may also suspect that in some Continental countries a series of educational reforms in the past two decades or so has misfired (particularly clearly in Germany), and standards there have slipped: the gaps shown by previous international educational surveys between Britain and the Continent (to Britain's very substantial disadvantage, at least in mathematics) have thus probably narrowed: but—as all too clear from recent visits to Continental classrooms by the Institute's teams of teachers, inspectors and researchers—those gaps are very unlikely to have been reversed.

Our central concern in this Note has been whether the new PISA international tests of 15-year-olds cast any measurable light on this: but, as detailed above, serious doubts attach both to the designed *objectives* of the PISA survey and to the *methods* by which the survey was carried out. In brief, doubts attach to:

—the kind of questions asked in this survey: they were deliberately *not* related to the school curriculum (in contrast to previous surveys), and so unlikely to be of specific direct help to schools, or to educational policy-makers;
—the higher ages of the young people questioned in this, in contrast to previous surveys: 15 +, by which age in some countries—such as Germany and Switzerland—some pupils have left school and are in employment or unemployment, and others are in part-time vocational colleges and difficult to reach in a sample survey;
—the consequences of selecting a specific *year of birth* as the basis for sampling pupils, rather than a specific *school-grade*, as in previous international inquiries. For Continental schools PISA thus included slower-maturing pupils—entering schooling a year late, or having repeated a class—who would be in a lower school-grade and not performing as well in tests, even though such pupils would often be expected to remain in school on the Continent for an additional year and reach

higher attainments before facing the outside world. As compared with previous international surveys of educational attainments, Continental pupils would appear to be doing less well in PISA, but English pupils would not have been affected by this change from previous surveys;

—the response rates achieved both in respect of the schools that were sampled and in respect of the pupils within them. Response rates were particularly low in England where it seems that barely 50% of the original representative sample of pupils eventually participated (the United States and The Netherlands also had very low participation rates). Those who did not participate must be suspected of being, on the whole, in the lower-attaining groups of the population, thus biasing upwards the recorded average scores for England. To compare such results with France, Germany, Hungary, Switzerland, for example, where something like 90% of the originally drawn representative samples participated, runs the danger of being seriously misleading.

These reservations, taken together, are sufficiently weighty for it to be unlikely that anything of value for educational policy in the UK can be learnt from the PISA survey. We have tried to estimate an order of magnitude for the bias introduced by each of the above reservations, though this has not always been possible. We can perhaps summarise in two general points. First, while no single reservation by itself seems of overwhelming importance, on the whole they seem likely to have led to a substantial upward bias in the average English recorded score. Secondly, the probably over-riding non-quantifiable difficulty is that the overall objective of the PISA survey (in distinction from previous international surveys) was *deliberately not* connected to the school curriculum: the questions often seen as much a test of common sense or IQ. Whatever adjustments or allowances we may make for failings in the execution of this survey will thus not help us in drawing lessons for improvement in the school curricula or schooling policy.

The conclusions drawn from previous international tests (IEA-TIMSS and IAEP)—which were focused on pupils' mastery of the school curriculum—thus remain relevant for schooling policy; no grounds are provided by this latest PISA survey for Britain to relax its policy measures to raise pupils' schooling attainments.

Unless there are substantial changes in PISA's objectives and methods, consequential questions arise as to whether Britain should in future participate in—and whether the taxpayer should continue to finance—further rounds of these very expensive surveys (the next PISA round is planned for later in 2003; and a further IEA mathematics survey is also planned for this year!). The fact that an international organisation is the motivator of such a survey has not proved sufficient to ensure that it is sensibly carried out in many vital respects. Especially in the UK, but probably also in the US and The Netherlands where response was also very low, it seems that much further exploratory research is necessary (guided by a wider advisory committee, including additional experts on sampling methodology) on the nature and treatment of non-respondents; until that has been completed satisfactorily, a policy of diplomatic inactivity in relation to further rounds in the UK of this kind of survey seems the better option.

ACKNOWLEDGEMENTS

In preparing this cautionary Note I have benefited from, and been saved from many errors by, comments on earlier drafts by Geoffrey Howson (Emeritus Professor of

Mathematics, Southampton), Dr John Marks (previously, National Curriculum Council)—both of whom were associated with previous IEA (TIMSS) surveys—Professor John Micklewright (Southampton) and by officials of the Office of National Statistics (London) and of OECD (Paris); responsibility for remaining errors and misjudgements remains with me. For providing assistance in the preparation of this Note, my thanks go to the National Institute of Economic and Social Research, London.

NOTES

[1] OECD, *Knowledge and Skills for Life: First Results for PISA 2000* (OECD, Paris, 2001). Results are published there for the UK; first results for England—as distinct from the UK—were made available in a National Statistics release (4 December 2000) and showed negligible differences from those for the UK. We refer below to results for the UK or for England almost interchangeably, depending on available sources.

[2] Not all the countries listed took part in all three surveys and some approximation ('triangulation') is necessary to reach the above broad comparison. For the IEA sources, see A.E. Beaton *et al.*, *Mathematical Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study* (Boston College, 1996); I.V.S. Mullins *et al.*, *TIMSS 1999 International Report* (Boston College, 2000).

[3] R. Adams and M. Wu (eds), *PISA 2000 Technical Report* (OECD, Paris, 2002); B. Gill, M. Dunn, E. Goddard, *Student Achievement in England* (SO, 2002). We shall also refer below to two very full reports on Germany: J. Baumert, E. Klieme *et al.*, *PISA 2000: Basiskompetenzen …* (Leske + Budrich, 2001), J. Baumert, C Artelt *et al., PISA 2000: Die Länder der BDR im Vergleich* (Leske + Budrich, 2000); and to the full report on Switzerland: C. Zahner, A.H. Meyer, U. Moser *et al., Für das Leben Gerüstet? Die Grundkompentenzen der Jugendlichen—PISA 2000* (BFS/EDK, Neuchatel, 2002).

[4] The OECD approach categorised four areas of the mathematics curriculum, only two of which were tested in 2000: 'space and shape' and 'change and relationship'. The other two areas of their categorisation—'quantity' and 'uncertainty'—are to feature in later rounds (OECD, 2001, p. 237). Only half the sample took the mathematics tests (the other half took the science test; all took the reading-literacy test). This four-area categorisation of mathematics may well seem idiosyncratic to English readers; wider issues of the approach are discussed further below (n. 10, and associated text). While we focus here on the *mathematical* questions in PISA, looking particularly at Britain, Germany and Switzerland, the reader may wish to know that, based on the *reading-literacy* part of the tests, Professor E. von Collani (University of Würzburg) arrived at broadly similar critical conclusions, having looked particularly at the United States and German reports (OECD PISA—an example of stochastic illiteracy? *Economic Quality Control*, 2001, p. 227).

[5] 'Mathematical literacy is assessed by giving students 'arithmetic' tasks—based on situations which … represent the kinds of problems encountered in everyday life' (OECD, 2001, p. 23).

[6] In relation to the great limitations of that OECD International Adult Literacy Survey, it is worth recalling its anomalous finding that 'three-quarters of the

French population have an ability in terms of 'literacy' which prevents them handling the normal matters of everyday life: reading a newspaper, understanding a short text, payslip etc' ('Weaknesses and Defects of IALSs', by A. Blum and F. Guérin-Pace, ch. 4 in S. Carey (Ed.) *Measuring Adult Literacy* (London, Office for National Statistics, 2000), p. 34. That the French withdrew co-operation is less than surprising.

[7] German Report (2002), pp. 17, 98.

[8] For vocational training in occupations requiring much accurate numerical work, so I was told by a number of German *Meister* a decade ago, they often preferred to recruit pupils from *Hauptschulen*—rather than from the next higher tier of *Realschulen*—because of the former's greater emphasis on numerical work. The schooling balance was already changing then, and no doubt has changed further since then (with the growth of *Gesamtschulen*, etc.), but the underlying point still seems valid. German senior educationists, despite having rejected the relevance of the *international* PISA questionnaire, remarkably enough were prepared to use the international PISA results, when published, as a basis for voicing their despair in the mass-media and proposing great reform programmes for their schooling system.

[9] OECD, *Measuring Student Knowledge and Skills: A New Framework for Assessment* (OECD, 1999); the illustrative mathematics examples quoted here are from pp. 56–69, with slight condensation in examples B–D (below) of the original wording (it is not clear whether these examples were actually set, or were merely illustrative—but this probably does not matter too much here). A subsequent OECD publication (*Sample Tasks from the PISA 2000 Assessments*, OECD, May 2002) quotes five mathematical 'units', with a total of 11 questions actually set. Stimulated by Example A above, two detailed limitations will occur to the reader. (a) In an international comparative survey it may well be difficult to use local currency. The arithmetic involved, and apparent difficulty of the question, for English students dealing with pizzas at, say £3 or £4, would be vastly different from that facing their Italian or Turkish counterparts (in the pre-Euro era). It is important then to know whether or not—in questions involving money—all countries used the imaginary 'zed' currency, or whether some countries substituted a local currency with which students would be familiar. (b) Although this question tests whether a pupil realises that areas do not increase proportionally with lengths, it does not test knowledge of the general relation between areas and linear dimensions of *similar* figures. At no point in their discussion of this question did the OECD report mention the mathematical term *similar*: OECD, indeed, suggested that students might answer this question using squared paper and compasses—not a 'real-life' option for most buyers of pizzas. The question also lacks the numerical complexities that occur in real-life problems. A more discriminating question would have been to price the larger pizza at 50 zeds. It is, no doubt, good that students should be able to answer the question posed by OECD, but ability to do so does not mean that they are actually equipped to deal with the numerically more complex, *real* 'everyday' problems of the pizza parlour! [Thanks are due to Professor Howson for these points.]

[10] Example B is particularly 'unreal' (even 'imaginary', in the technical mathematical sense): what is here supposed to happen in the first 12 years of the life of a lichen?

[11] After the above text was drafted (and I have left it virtually as drafted), results

became available for the 31 mathematics questions, and could be matched to the 11 published questions (texts of other questions have not been published in order that they could be re-used in subsequent rounds). The final question mentioned above—matching a speed-distance graph to five possible track-circuits—showed as follows. Only 8% and 10% of Greek and Portuguese girls ticked the correct alternative (a policy of random ticking should have given them 20% correct!); Australian and German boys ticked correctly as to 43% and 38%; Swiss boys did better at 46%. Overall for OECD (total) only 28% of boys and girls ticked correctly (35% of boys, a mere 21% of girls)—as against 20% who would have done so on a policy of random ticking. Was this a case of inadequate piloting; or was it a deliberate choice to highlight differences between boys' and girls' attainments in mathematics?

[12] Hans Freudenthal was the author of a valuable critical appraisal of the first IEA survey of 1964; and subsequently of many learned works on mathematical pedagogy, for example, *Didactical Phenomenology of Mathematical Structures* Dordrecht, Reidel, 1983). He refers in the latter (p. x) to the 'assassination' of his mathematical institute—clearly a not uncontroversial academic! He was, however, honoured in the subsequently-named Freudenthal Institute of Mathematics at Utrecht. The current head of that Institute, Professor J. de Lange, was Chair of the PISA Mathematics Functional Expert Group. It seems regrettable that there was no representative of the UK, nor of Germany or Switzerland, on that Group.

Some recent historical developments in mathematics teaching may be sketched here (though some of what follows is, inevitably, debatable). The 'realistic mathematics' approach can be understood as an initiative to counter the abstractions of the New (or Modern) Maths approach—associated with Set Theory, the French Bourbaki movement, mathematics as a closed deductive system and a branch of logic (related to the *Principia Mathematica* of Russell and Whitehead; an insight into Bourbaki is provided by the autobiography of one of its members, A. Weil, *The Apprenticeships of a Mathematician*, Birkhäuser, Basel, English version, 1992). The Modern Maths approach was given a considerable boost internationally by the OEEC (predecessor of the OECD—the sponsor of the current PISA survey) seminar at Royaumont in 1959. Modern Mathematics was described by Professor de Lange in 1988 as the 'main misconception in mathematics education in the last decade' (J. de Lange and M. Doorman (Eds) *Senior Secondary Mathematics Education*, Ultrecht, 1988, p. 14). That vigorous debate was conducted at a high level of principle, rather than on the empirical basis of what actually has become accepted as didactically effective in the classroom (for an example of the latter, see Helvia Bierhoff's paper, Laying the foundations of numeracy: a comparison of primary school textbooks in Britain, Germany and Switzerland, *Teaching Mathematics and is Applications*, 1996, pp. 141–160).

[13] For example, the Swiss use *actual* maps and timetables (in all their complexity) while in English school-examples it is more usual to find simplified ones (my thanks, again, to Professor Howson for his studies of school textbooks).

[14] I have adjusted the published percentages-correct to allow for guessing in multiple-choice questions using the conventional formula. For further discussion, see my paper in *National Institute Economic Review*, July 1997, pp. 56–57; and, in a fuller version, ch. 4 in *Comparing Standards Internationally* (B. Jaworski and D. Phillips (Eds) *Oxford Studies in Comparative Education*, Symposium Books, 1999, pp. 86–87).

[15] I.V.S. Mullis *et al.*, *TIMSS Mathematics 1999* (IEA, Boston, 2000), p. 88. Belgian pupils showed 80% correct; Dutch pupils, 72%.

[16] Perhaps a final reservation on the nature of the questions is worth footnoting: previous IEA mathematical surveys took care to repeat several questions from preceding surveys ('anchor questions') to provide guidance on changes and continuities. There was no attempt to do so in PISA though, for most readers, it would have been useful if PISA had included some explicit linkages taken from (or based on) previous international studies.

[17] An exception was the first IEA international mathematics survey of 1964 which was conducted both on the basis of school-grade and on the basis of age (but certain problems arose in practice with the British sample; see my paper with K. Wagner, 'Schooling Standards in England and Germany', published in *Compare: A Journal of Comparative Education*, 1986, and also in *National Institute Economic Review*, May 1985, especially the section in Appendix B: 'How old are British 13 year old pupils?'; as it turned out, they were 14!).

[18] German report (2001), p. 473; of the 36% not in their 'normal' school year, 12% had entered their schooling late, and 24% had repeated a class (the latter includes some who had also entered late). In Switzerland 6100 pupils who were 15 years old were sampled, of whom 75% were in their 'normal' ninth grade (Swiss report, 2002, p. 17). The German 'extended' sample was based entirely on their ninth grade.

[19] OECD (2002), p. 183.

[20] The TIMSS survey of 1995 was unusual, and unusually helpful for the analyst, in that it covered two school-grades: the seventh and eighth international grades, corresponding to English Years 8 and 9. An example of the findings: the weakest 5th percentile score in mathematics by 14 year-old English pupils was 361 (again, based on a standardised international average of 500 and standard deviation 100), while pupils in correspondingly-aged classes in Switzerland (international eighth grade) had a 5th percentile score of 401, and those in a class one year younger (in Switzerland seventh grade) had a 5th percentile score of 387. Similar but less extreme contrasts emerged at higher percentiles (10th and 25th); and from comparisons between England and Austria, Belgium (Flemish and French), Czech Republic, France, Hungary and the Netherlands (Beaton *et al.*, pp. E2 and E3); issues of poor response, and differences in the surveys' coverage of pupils with special educational needs, complicate any simple interpretation of these statistics—but, based on the Institute's teams' direct observation of classes in Switzerland and England, the general implication of these statistics is as it appears.

[21] Some English schools agreed to participate only in respect of their Year 10 pupils (because of pressures in Year 11 due to GCSE); that is to say, only about the youngest third of such schools' pupils were included. The missing section of Year 11 pupils was 'replaced' from other schools; but a doubt is left whether they were equally representative.

[22] In the two Continental countries just mentioned, there is (what we might, for our purposes here, term) a vocationally-oriented stream of secondary schooling (*Hauptschulen* in Germany, *Realschulen* and *Oberschulen* in Switzerland's German-speaking Cantons) attended by about a third of all pupils; this stream continues only to their '9th Class' (corresponding in age, very approximately, to our Year 10); for example, the top year of that secondary vocational stream in Zürich caters for pupils normally aged 14:4 to 15:3 at the beginning of their school year. By

the following March, when the PISA survey took place, the normal age-range would be 14:10 to 15:9; most of that year's pupils are then beyond the age of compulsory full-time schooling. Some pupils in that stream continue, nevertheless, in their secondary school to the end of the school-year and receive their certificate of school-completion based on the standards of their '9th Class'. Others move into part-time vocational courses at *Berufsschulen* associated with their apprenticeship. The PISA *Technical Report* (p. 1, n.) gives the impression that compulsory schooling in Germany ends at 18; in fact, *full-time* compulsory schooling ends generally at 15, followed by three years part-time schooling for those undertaking an apprenticeship. Those not undertaking an apprenticeship are usually obliged to stay only for a further full-time pre-vocational course of one year's length, till age 16.

[23] German report (2001), p. 514. For the larger German national inquiry (PISA-E), vocational schools had an even lower pupil response-rate of 36% (German report (2002), p. 25).

[24] English report (2002), p. 19.

[25] German report (2002), p. 19 (my translation, and my emphases, throughout these quotations).

[26] German report (2002), pp. 19 and 115.

[27] Clue: there was proper concern at OECD that the policy of 'mainstreaming' children with learning difficulties could lead to too many exclusions in sampled mainstream schools of children with Special Needs. A ceiling of 2% was set for such exclusions. The proportion of children in Special Schools in Britain fell below that ceiling, and such schools could therefore be excluded; in Germany, the proportion of children in Special Schools was above that limit, and such *schools* were therefore included in the sample, though low-ability *pupils* within them could be excluded up to that ceiling.

[28] The second sample related only to replacements 'matched' for schools that did not participate originally (not chosen at random from the *whole* reserve sample, as might be understood from a quick reading of the procedure). Textbook random sampling theory cannot be applied to such a hybrid sampling process (even if a more sophisticated stratification model is brought into theoretical play—since schools that *did not participate* from the first sample may be expected to have characteristics differing from those which *did participate*). The PISA sampling process is closer (in part) to 'quota sampling' as used in commercial work, but this is not usually acceptable in scientific or governmental inquiries.

[29] Cf. OECD, p. 235, our emphasis.

[30] Curiously, a Research Brief on PISA prepared by the Department for Education and Skills (for England and Wales) includes the comment that the English sample of schools 'met the minimum international sampling requirements' (Brief no. RBX 25–02, p. 3)—when this was patently not so: even with replacement schools, only an 82% participation rate was achieved, while, as noted above, the explanatory OECD commentary said: 'A minimum response rate of 85% was required for the schools *initially* selected' (OECD, 2000, p. 235, emphasis added). No explanation of these inconsistencies has so far been obtained.

[31] No significant correlation was found by ONS between response-rates and (a) the proportion of *high*-attaining pupils in a school (proportion of pupils with 5 GCSEs at level of C and above) and (b) the proportion of pupils with free school meals; the correlation was calculated for only six groups of schools, and the lack

of statistical significance may reflect only the low number of degrees of freedom (i.e. based on only *six* grouped observations). It does not rule out that a comparison with the proportions of very low-attaining pupils might be more revealing.

[32] Often especially prone to absence when there is a voluntary test; cf. the discussion of German response, PISA, German report (2002), pp. 28–29.

[33] The lower response rate in these cities was attributed (in part) to the holding of local final school examinations close to the date of the PISA inquiry (loc. cit.).

[34] As, for example, in census of industrial production statistics, where *grouped* statistical returns used to be published only where there were a minimum of three respondents.

[35] The peculiarity of this procedure needs no emphasis here, and must have been an embarrassment to those who wrote the accompanying convoluted text on pp. 25–26 of the OECD *Technical Report*.

[36] Officials of the OECD responsible for PISA were questioned specifically on sample-representativeness by the House of Commons Select Committee on a visit to Paris, the HQ of OECD, on 20 March 2002 (quotations that follow are from the published minutes). On school-representativeness, the OECD official said in reply (p. Ev2) the UK 'got back just under 85%', when, as explained above, the correctly calculated rate (including the replacement sample of schools that was approached) was about 60%. On the pupil-response rate the official spoke (*loc.cit.*) of a target of 95%, and that for Britain 'there was not a problem with the response rate within schools'—when, in fact, the OECD report showed 81% as the UK's 'weighted student participation rate'—lower than any other participating country! A reader can be forgiven for worrying both about the correctness of OECD officials and about the adequacy of technical advice available to our Select Committees.

[37] Assuming that the lowest 25% has an *average* score close to the lowest 12th percentile of a Normal distribution, roughly 1.15 standard deviations from the mean, leading to a score on the PISA scale of 500–115 = 385. Removing that quarter from the full distribution would raise Britain's average to (500–385/4)/0.75 = 538. A more accurate calculation is possible based on the textbook theory of the truncated normal distribution; but it hardly seems warranted here.

Correspondence: Professor Sig Prais, National Institute of Economic and Social Research, 2 Dean Trench Street, Smith Square, London SW1P 3HE, UK. E-mail: mockenden@niesr.ac.uk

ANNEX: ON THE COMPUTATIONAL MYSTERIES OF PISA'S COUNTRY-RANKINGS

Doubts have been raised in the main paper on the precise purposes and on the methods of execution of the PISA tests; further doubts of an apparently elementary computational sort—but serious in their implications—came to light when scores for individual PISA questions became available after the writing of the main paper was completed; these scores give the percentage of pupils answering *each* question correctly in *each* country [1].

I began by comparing Switzerland and Britain, countries in which the National Institute of Economic and Social Research has had a research interest for many years;

on the basis of direct classroom observation we were satisfied that previous international test-surveys of mathematical attainments were right in reporting that pupils in Switzerland were well ahead of those in Britain. As mentioned in the opening paragraphs of the main paper, PISA now indicates that Britain has caught up with Switzerland in mathematics, with both countries showing the same standardised score of 529 (about a third of a standard deviation above the OECD international standardised average of 500). For the 31 mathematical questions that have more recently become available, I found (to my astonishment!) that Switzerland was *ahead* of Britain in answering correctly roughly two-thirds of the number of questions (21/31). Further, a simple average for each country of the percentages correct yielded: for Switzerland, 54.1% correct; and for Britain, 52.3% correct.

To put these figures in wider perspective, it is helpful to learn that Japanese pupils, for example, were still ahead at 61% correct—but no longer appeared phenomenally ahead (557 on the PISA scale); Canada, which had been put ahead of Switzerland by the PISA calculations, was now below it (at 53.5%); and Belgium, which had been put by PISA below Britain, was now ahead of Britain (at 53.2% correct). These differences between countries in the percentages correct are in reality small; for example, the lead by Japan of 7 percentage points over Switzerland, in effect amounted to answering correctly only 2 more questions out of 31 than Switzerland (19 correct as against 17): a gap that hardly seems insurmountable. One is tempted to wonder whether the questions in this survey—being based on ostensibly 'everyday life' situations—were adequately discriminating in relation to the wider objectives of the school curriculum.

It might be queried whether all 31 questions deserve equal weight in calculating an average for the test as a whole; perhaps it is thus worth notice that the greater success of the Swiss was slightly more evident for harder than for easier questions (judging a question's difficulty by the percentage of pupils for the OECD-total answering incorrectly). Perhaps internationally difficult questions deserve greater weight in calculating an average. A *weighted average* was therefore also calculated (with weights for each question given by the proportion of the OECD-total answering incorrectly); the result was that the average Swiss weighted percentage correct was even slightly further ahead of the Britain (a lead of 2.2%, instead of 1.8% on an unweighted basis, as above).

I also explored the consequences of certain questions being put by PISA into the category known, technically (in the IRM process to be described below), as 'dodgy items' ('poor psychometric fit'); three mathematical questions were put by PISA into that category for German-speaking Switzerland (questions 155/01, /03 and /04) and were omitted by them in calculating their national average ranking [2]. Accordingly, those three questions were left out of our next comparison and, on an unweighted basis, obtained averages of 54.8 and 52.7% correct for Switzerland and Britain respectively (a difference of 2.1 percentage points, very similar to, and slightly greater than, the difference of 1.8 points between the unweighted averages quoted above) [3]

The different rankings published by PISA—putting Britain on a par with Switzerland—might of course be due simply to an arithmetical or copying error; but that seems unlikely since, when *very great* computers are involved, they usually make only *very great* mistakes, which are usually quickly observed and quickly put right.

A more likely cause of the discrepancy is the use by PISA of a highly complex multi-step averaging process, being a variant of what is known as the Rasch transformation. One might have thought that an exposition of what has been done would start, empirically, from a table of the actual average percentages answering each question correctly in each country, and then show how each step of that complex transformation

process modified those percentages to reach, eventually, the table of country-rankings that are the pride-and-joy of the PISA exercise. The reader is however offered only a highly mathematical (algebraic) ten-page chapter, without illustrative intermediate statistical tables, on *Scaling PISA Cognitive Data* (by Professor R. Adams of the Australian Council for Educational Research, and Project Director of the PISA Consortium); the chapter is not easy reading even for professional mathematicians, and makes no concession to those who are not fully adept research-psychometricians [4]. 'Opaque' was the term used by three experts commenting on the use of such techniques in relation to the previous International Adult Literary Survey [5]. The use of Rasch modelling, or Item Response Modelling (or Item Response Theory; for short, IRM or IRT) has been the subject of intense academic controversy for many decades; it seems to be centrally posited on the 'supposed probabilistic nature of responses conditioned on an assumed trait called ability'—in this case, mathematical ability [6]. As noted in the main paper in relation to Germany, there was considerable controversy on the proper weight of arithmetical questions in the mathematical total; in other words, it has become clear in the present context that different countries' school curricula in mathematics lead to different response-probabilities for different types of mathematical questions. Consequently, one cannot always summarise observed inter-country differences into a simple sum (or product) of a *country-effect* and a *question-effect*, which is the essence of IRM analysis. If one tries to do so, inconsistent results ensue; and some questions have to be removed because of so-called 'poor psychometric fit'—that is, the simplified model adopted for the analysis is not really appropriate [7].

Faulty Weighting of Responses

In the course of correspondence with OECD, two specific possible sources of the above discrepancies came to light. First, in calculating the overall percentage of pupils in each country answering all questions correctly, it seems that the PISA computer was instructed—not to give the same weight to each question in all countries, but varied the weight in each country according to the number of respondents to *each* question in that country. The origin of this anomalous procedure was that different booklets with different mixes of questions had been prepared, mainly so that pupils sitting next to each other would not find it easy to copy. While, in very broad terms, all questions were tested in all countries, yet the precise relative numbers differed. Sometimes the difference was considerable, since a particular question may have been omitted entirely in one part of a country, or even in the whole country (the former happened, as mentioned, to three questions in German-speaking Switzerland; the latter to one question in Japan). The returns for a great many other questions (about half the total number of questions) are labelled, in the compendium of responses: 'Item statistics (i.e. percentages shown as correct) not based on the whole population'. The countries labelled in this way were Belgium, Germany, Hungary, together with the European country that was near top in everything—Finland! An enquiry to Finland was not able to cast any light on which parts of the population had been omitted—though it is obviously relevant to the confidence we can place in its PISA ranking.

The PISA computer program, in calculating the average percentage of questions answered correctly in each country, weighted each question according to the actual *number of respondents* to that question in *each* country—whereas each question should have been equally weighted in all countries. The actual number of pupils responding in each county is indeed relevant, but only to the calculation of *sampling errors*; it should

not have entered into calculations of the average number of questions answered correctly [8].

Questions 'not reached'

A second obscurity, minor in quantitative significance (clarified in course of correspondence), relates to the treatment of questions 'not reached' by candidates in the course of their test; the proportion in this category varied, according to question, from 0.7% to 8.9% for all countries. It is not obvious how markers decided whether a candidate had simply ignored a question because it was difficult, or because he ran out of time; whatever was done by the markers, in tabulating the results for the 'compendium' of scores for individual questions, questions 'not reached' were ignored; however, all this was reversed in the IRM ('black box') computation. The net effect of this element in our comparisons between Switzerland and the UK (in the opening paragraphs of this Annex) is to reduce by about 0.5% the gap between the percentage correct for Switzerland and Britain.

Until yet more light is cast by PISA on their view of the sources of the above discrepancies, it would seem necessary for the ordinary reader to proceed with great caution on the central PISA country-rankings. It is a matter of serious regret that PISA did not publish the simple percentage of questions answered correctly by pupils in each *country*, together with details and rationale of their subsequent transformation [9].

NOTES

[1] Available on the following FTP website (expert help may be necessary for downloading): ftp://ftp.acer.edu.au/pisapublic/Testitemcompendium3.doc.

[2] See PISA *Technical Report*, pp. 101, 153. No details of the subject matter of these omitted questions have been published; nor any explanation of why French- and Italian-speaking Switzerland tackled these questions satisfactorily. Separate averages for the three Swiss linguistic groups were published by OECD (2001, p. 317) but without indicating whether those three questions had been omitted.

[3] I am indebted to my colleague at the National Institute, Professor Ray Barrell, for help in these computations.

[4] PISA *Technical Report*, pp. 99–108.

[5] A. Blum, H. Goldstein, F. Guérin-Pace (IALS: an analysis of international comparisons of adult literacy) *Assessment in Education* 8 (no. 2), 2001, p. 231. They also argue for the publication of the 'proportion of correct responses' (p. 230)—something that seems to be doctrinally opposed by the proponents of IRM.

[6] Cf. H. Goldstein and R. Wood, Five decades of response modelling, *British Journal of Mathematical and Statistical Psychology* (1989), 42, pp. 139–167; the quotation above is from p. 163. A helpful exposition of IRM is given in H. Goldstein's chapter on the methodology of the International Adult Literacy Survey, in S. Carey (Ed.), *Measuring Adult Literacy* (London, Office for National Statistics, 2000), pp. 34–42. There are also worries as to how satisfactorily multiple-choice questions are treated by these models; A.E. Beaton, one of the American supporters of these methods, wrote: 'The Rasch model … does not fit well for multiple-choice tests in which there is substantial guessing' (Beaton, ch.

3 in Carey, *op. cit*, p. 28). About a third of the PISA mathematical questions were multiple-choice.

[7] Questions may be omitted by the 'poor fit' criterion even though they may form part of the curriculum particularly emphasised in that country.

[8] It is as if the average weight of fruits grown in different countries is to be compared by weighing baskets of apples, pears and oranges, each basket consisting of nine fruits (the statistician having in mind three of each variety). The officials directly involved, however, thought it would be in order to take unequal numbers of each variety, according to local convenience, so long as there were nine in total. Differences in average weights of fruits are thus confounded with differences in the *mix* of fruits (apologies for labouring what will be obvious to most readers).

[9] I have left the above considerations as an Annex, rather than incorporating them in the main paper, since I do not feel I have yet sufficiently narrowed the issues down to the 'mysteries of the black box' rather than just copying-type errors at some stage.