Draft revd 9 June 05
Not for publication
This preliminary version is circulated for readers to suggest comments and improvements prior to publications.

The National Institute of Economic and Social Research
Discussion Paper no. 258
June 2005

TWO RECENT (2003) INTERNATIONAL SURVEYS OF SCHOOLING ATTAINMENTS: ENGLAND'S PROBLEMS

by S J Prais

Abstract: *The two recent (2003) international surveys of pupils' attainments were uncoordinated, overlapped considerably, were costly and wasteful, especially from the point of view of England where inadequate response-rates meant that no reliable comparisons at all could be made with other countries. Sources of the problem are investigated in this paper and suggestions made for improvements in possible future surveys.*

Some astonishment was aroused by the recently published results of *two*, apparently independently organised, large-scale international questionnaire surveys of pupils' mathematical attainments towards the middle of their secondary schooling (age 14-15); nearly 50 countries participated in each survey, with some 200 schools in each country. Both surveys were carried out in the same year, 2003; previous surveys had generally been carried out at about ten-year intervals, and each one of these surveys only 3-4 years previously. Some questions on science and literacy were included in 2003, but the focus was on mathematics (and that is our focus here). A test towards the end of primary schooling, at age 10, was also carried out in association with one of these surveys. The total cost was probably over £1m for England, and probably well over $100m for all countries together, plus the time of pupils and teachers directly involved.[1] Results were published by the beginning of 2005 in several thick volumes, totalling some 2000 large (A4) pages; the two organisations behind the surveys are known as TIMSS and PISA (details of the organisations and publications at Annex A at the end of this paper). There does not appear, from these publications, to have been any coordination between the two organisations. Much wasteful overlap and duplication is evident; the interval between recent repetitions of these surveys was so tight as not to permit adequate consultation for lessons to be learnt.[2]

[1] Only limited information on costs of these surveys has been released. For England, a total of £0.5m was paid to the international coordinating bodies, but information on locally incurred costs were withheld (in reply to a Parliamentary Question on 7 March 2005) as publication could 'prejudice commercial interests' in the government's negotiating of repeat surveys in 2006-7. It is astonishing that expenditure on further surveys should have been put in hand before there has been adequate opportunity for scientific assessment of the value of the 2003 surveys and of the appropriate frequency of their repetition.

[2] The PISA (Programme of International Student Assessment) inquiry of 2003 was organised by OECD and followed their first attempt in this activity in 2000; the report on their first survey was critically reviewed in my article in the *Oxford Review of Education*, **29**, 2 (2003). The acronym TIMSS was originally short for *Third* International Mathematics and Science Study; subsequently it became short

Representativeness of samples

We shall try and assess here some of the main findings for England, ask whether further surveys of this kind are justified, and whether anything is to be learnt from these recent surveys which might improve future surveys. What can be said with any confidence about English pupils' attainments towards the end of their secondary schooling is much limited by poor sample response. From the TIMSS report on 14 year-olds we learn: 'England's participation fell below the minimum requirements of 50 per cent, and so their results were annotated and placed below a line in exhibits (= *statistical tables*) showing achievement'.[3] For the parallel PISA report, in all tables mentioning findings for the United Kingdom a footnote was attached to the line for the UK (and only for the UK!): 'Response rate too low to ensure comparability'.[4]

In other words, any differences that may appear between published results for England and other countries are not to be relied on. This reservation was not however attached to the tests of English 10 year-olds towards the end of their primary schooling (carried out by TIMSS, following a similar survey at that age in 1995); and those results, to first appearances, appear to be the most scientifically interesting and important for educational policy. We will need to examine below whether those results are indeed robust enough – that is to say, adequately representative – to be relied upon.

But before that, a short word on the recent historical background of Britain's schooling attainments may be helpful. Britain's economic capabilities – its motor industry, machine tool manufacturing industry, as well as other industries relying on a technically skilled workforce – led to much public concern by the 1960s: expressed subsequently, for example, in the official Cockcroft Committee's report on *Mathematics Counts* (HMSO, 1978), eventually leading to the National Curriculum, the National Numeracy Project, and then to nationwide annual testing of all pupils in basic school subjects at all primary and secondary schools (SATs at ages 7, 11 and 14 to supplement the longer-standing GCSE tests at 16). Detailed empirical comparisons in the 1980s and 1990s by teams centred at the National Institute of Economic and Social Research (London) were made of productivity and workforce qualifications. Site visits to comparable samples of plants in England and Germany clarified the nature of the great gaps in workforce qualifications; these gaps were not so much at the university graduate level, but at the intermediate craft-levels (City and Guilds, etc.) – the central half of the workforce. The difficulty in England in expanding that central category of trainees was traced to the secondary school-leaving stage when the standard of mathematical attainments required for craft and technician training, especially in numeracy, were much below Germany's. The IEA's First International Mathematics Survey of 1964 (FIMS – the original predecessor of TIMSS) was one of the important sources that confirmed this gap in secondary school mathematics; it was made evident to our teams of secondary mathematics teachers and inspectors on visits to secondary schools in France, Germany, the Netherlands and Switzerland, and in

for *Trends* in International… The previous occasion on which it had been carried out was 1999. More of the 2003 co-ordinating costs (76 per cent) were incurred by PISA, making TIMSS – which covered two age-groups – the better buy for the British taxpayer.

[3] TIMSS, *Mathematics Report*, p. 351.

[4] See, for example, PISA, Annex B, *Data Tables*, pp. 340 *et seq.*

discussions with heads of industrial training departments (*Meister).*[5] An important conclusion from visits to schools was that it was quite unrealistic to expect English *secondary* schools to be able to produce the numbers of students with levels of mathematical competence that had been seen abroad if they had to start with the standards delivered by our *primary* schools.

Shifts in research interests and in educational policy ensued for mathematics teaching, especially at primary level. Textbooks here and in Europe were carefully compared; teaching methods abroad were observed by practising teachers; new teaching schemes were prepared; and annual nationwide tests of pupils' attainments were administered nationally to all pupils at ages 2-3 years apart (SATs). Much more could be said on the details of what has amounted to a 'didactic revolution'; but perhaps the foregoing is sufficient here to indicate the interest attached to the 2003 TIMSS mathematics results at age 10 which can be compared with the similar sample inquiry eight years previously at that age (the 1995 TIMSS – Third International Mathematics and Science Survey). Had we now caught up with our competitors at least at the end of primary schooling?

The comparison was set out, clearly and apparently convincingly, in the *national* report for England for 2003 produced by the (English) National Foundation for Educational Research (which carried out the survey in England in coordination with the international body). It noted that England's mathematics scores showed the largest rise of any of the 15 countries that participated at the primary level in both 1995 and 2003 (the English rise was of 47 standardised points, from 484 to 531, where 500 is the notional average standardised score of all countries in these international tests, and the standard deviation is standardised at 100). Most test questions asked were different in the two years, but 37 questions were the same in both years; the proportion who answered those common questions correctly in England rose very satisfactorily from 63 to 72 per cent. The rise was even a little greater in questions relating to numeracy (arithmetic); this may all be taken as reassuring, since previous deficiencies in English students' attainments were, as said, particularly marked in that area – the foundation stone of mathematics.[6] The top countries at the primary school level were, once again, those bordering the Pacific: Singapore, Hong Kong, Japan – with scores averaging about 570; England's rise in performance in the nine intervening years, by 47 points to 531, can thus be seen as approximately *halving* the gap with these top countries – and in hardly more than a decade.

To first appearances, this seems a remarkably encouraging achievement; and, one must equally say, in a remarkably short time-span given the complexity of what amounted to changing almost the whole mathematics didactics system. But are these sample results to be relied upon? We have noted that at the secondary school level

[5] See my paper with K Wagner, Schooling Standards in England and Germany: Some summary comparisons bearing on economic performance, in *National Institute Economic Review,* May 1985 and in *Compare: A Journal of Comparative Education*, 1986, no 1. More generally, see the series of reprints re-issued by NIESR in two compendia entitled *Productivity, Education and Training* (1990 and 1995). Teachers and school inspectors, particularly from the London Borough of Barking and Dagenham, were invaluable in assessing school-visits here and abroad.

[6] See G Ruddock *et al., Where England Stands in the Trends in International Mathematics and Science Study (TIMSS) 2003,* (NFER), 2004, pp. 8-10.

(age 14) serious reservations were attached by the surveys' sponsors to response rates to the samples for England; at the primary level (average age 10.3, Year 5 in England) a cautionary footnote is always attached to the TIMSS results reported for England (not as serious as for secondary school results – but not to be ignored): 'Met guidelines for sample participation rates only after replacement schools were included'.[7] With that modestly expressed caution in mind, let us next patiently re-examine the actual response rates for England, bearing especially in mind that if response rates were lower in 2003 than in 1995 we might expect better average scores to be recorded simply as a result of 'creaming higher up the bottle'.

We first compare the response for schools; then the response for students within responding schools; and finally, the product of these two rates. In 2003 there were 150 primary schools in the original English representative sample, of which 79 schools participated, or 53 per cent.[8] For the previous primary school inquiry of 1995, 92 out of 145 sampled schools participated at the fourth grade – 63 per cent.[9]

The student participation rate (within participating schools) was 93 per cent in 2003, just a little below the 95 per cent recorded for 1995. Combining the two participation rates (schools *x* students) we have a participation rate of something like 50 per cent in 2003 compared with 60 per cent in 1995: there are thus grounds for worrying whether there has been a *genuine* improvement in scores in the population.[10]

But are either of these overall response rates adequate for anyone to place reliance on the representativeness of the results? Even TIMSS put the 'minimum acceptable participation rate' at 'a combined rate (the product of school and student participation) of 75 per cent'; but at Year 5 in England (as also in five other countries)[11] that criterion was said to be satisfied '*only after including replacement schools*'. This brings us to a long-standing thorny dispute on acceptable sampling practices. The sampling procedure adopted in these international educational inquiries is not at all orthodox. It starts with several parallel lists of schools, each list being equally representative.[12] If an inadequate response is received from the initial list, then 'corresponding' schools from the second list are approached, and from a third list if necessary. For England in 2003, as said, a sample of 150 schools was drawn from the initial list; in the outcome, 79 schools from that list participated (a mere 53 per cent) and 71 schools refused. A further 71 (replacement) schools were then chosen from the second list, an estimated 27 schools of which participated (38

[7] IVS Mullins *et al., TIMSS 2003 International Mathematics Report* (IEA, Boston), 2004, for example, p. 35.

[8] *Ibid.* p. 355.

[9] IVS Mullins *et al., Mathematics Achievement in the Primary School Years* (TIMSS), 1997, p. A 13.

[10] The reader will understand that the gradient of the response-rate with respect to attainment-level will be different according to whether it is amongst schools, at the school-level, or amongst students within schools; but the point is not worth elaboration in view of what is said in the next paragraph.

[11] Australia, Hong Kong, Netherlands, Scotland, United States (*ibid.*, p. 359). For the US a response rate (before replacement) of only 66 percent was recorded for the primary survey and the same for the TIMSS secondary survey. For England's secondary survey, the corresponding proportion was a mere 34 per cent!

[12] For example, starting from an initial list of schools organised by geographical area, size, etc., a random start is made; subsequent schools are chosen after counting down a given total number of *pupils* (so, in effect, sampling schools with probability proportional to their size). A second reserve list is yielded by taking schools, each one place above the schools in that initial list; and a third reserve, by going one place down the initial list.

per cent) and 44 refused; an estimated 44 were then approached from the third list, of which 17 participated. The total number of schools now participating totalled 79+27+17=123; the total number *approached* was (*nota bene*, since the organisers of these surveys do not agree!) 150+71+44=265; the overall response rate for schools was therefore 123/265=46 per cent (a little below the 53 per cent from the first list!).[13] Taken together with a response of 93 per cent of students in participating schools, the total combined response (schools and students) was thus only 43 per cent – all much below the proportion (75 per cent) originally laid down by TIMSS as acceptable!

Incredible as it may seem, the statisticians at TIMSS calculated a participation rate, not in relation to the total number of schools *approached* on first and subsequent lists (221), but in relation to the smaller number originally aimed at (150); they consequently published a misleading response rate of 123/150 = 82 per cent for schools, and of 75 per cent for schools and students combined – just falling into their originally stipulated requirements (whereas the correctly calculated combined response rate, as just said, was only 43 per cent). Was this merely a momentary slip? And forgivable? Or was it more in the nature of a scientistic *trompe l'oeil* (one hesitates to say sleight of hand) encouraging readers that all was fundamentally well, and had been placed in sound hands – including the hands of a so-called *Sampling Referee,* an expert to whom such technical statistical details had been safely relegated? Having discussed this issue with a number of British statisticians, I have regretfully come to the conclusion – putting it as kindly as I can – that these surveys' statisticians had misled themselves as a result of their commercial experience with quota sampling; and that any future such inquiry needs to be advised by a broader panel of social statisticians.[14] For the sake of clarity, I repeat that such an enlarged body will need to address two issues: first (a simple arithmetical issue), what is the correct method of *calculating* a total response rate if further, 'replacement', samples

[13] The official account gives only the total responding from the second and third lists combined, at 44 schools. For our purposes here, we may estimate the numbers at each stage, at least adequately correct for these purposes, as follows.

Assume that response rates were equal for both reserve lists at p, so that 71p schools agreed to participate at the second stage, and 71(1–p) refused. For the third stage, the number participating (assuming the same probability of response) would then be 71(1–p)p. We are told only that 44 schools participated from *both* reserve lists together, so that

$$44 = 71p + 71(1-p)p.$$

A little manipulation leads to the quadratic equation

$$p^2 - 2p + 0.62=0,$$

which solves, in the usual way, to yield p=0.38. The total of 44 replacement schools thus divides (at least approximately) into 27 schools from the second list and 17 schools from the third list.

The calculated response rate of 38 per cent for each of the second and third stages is below the 53 per cent recorded for the first stage, indicating that more difficult schools were being approached. (I have relegated this argument to a footnote so as not to distract from the main issue of the next paragraph on the misleading nature of the response rates published by the sponsoring organisations).

[14] Quota sampling is used in commercial work, and places greater emphasis on achieving the agreed total of respondents, rather than on their representativeness; it is avoided in scientific work. On the 'Sampling Referee', see TIMSS 2003, p. 441. The issue of replacement sampling was questioned in my previous paper on PISA 2000 (*Oxf. Rev. Education*, **29**, 2); see also the response by RJ Adams (*ibid.*, **29**, 3), and my rejoinder to that response (*ibid.*, **30**, 4). The need for representative sampling is so basic to scientific survey procedures that it is astonishing that those responsible for educational surveys, together with the government departments providing taxpayers' money for such exercises, could accept such an easy-going (slack) approach to non-response. But, as it now turns out, this was not the last word – as discussed below in relation to re-weighting with population weights.

are included; secondly, is there any substantial *scientific justification* for approaching a 'replacement sample' (rather, say, than an initially larger sample)?

Returning to the real issue on which we would all like to draw happy conclusions, namely, the tremendous rise in our pupils' attainments at age 10, we see from the previous paragraph that the sample of responding schools (at 43 per cent, not 82 per cent as reported by TIMSS) has to be judged as altogether too low to support any such conclusion.

A pity! But we cannot leave the topic of response rates without noticing a considerable improvement in the way that England's *secondary* school scores were calculated for TIMSS. As said at the outset, the whole of the English results were rejected for international comparability in the *international* reports because they did not satisfy their originally specified sampling requirements (the rejection applied equally to TIMSS and PISA). There was however an additional *national* report on England's TIMSS survey which outlines an alternative calculation based on re-weighting the sample results by population weights. It tells us that the TIMSS sample *over*-represented schools that were 'average and above average in terms of national examination (or test) results (*i.e. weaker schools were under-represented: SJP*). This sample was therefore re-weighted using this measure of performance to remove this effect'.[15] Presumably, the obligatory nationwide SAT test results were used to provide better weights, but details have not been released as to whether, for example, the re-weighting was for the country taken as a whole, or for the sampled schools or, indeed, the sampled students. The consequence of the re-weighting was that England was moved down in the TIMSS mathematics ranking below Australia, the United States, Lithuania, and Sweden (a reduction of England's international score from 505 to 498). Nothing of very great substance, it might be thought; but the new method of estimation is of great importance for future surveys.

Such an adjustment raises the reliability of English estimated average scores because, to put it simply, it employs *population* – rather than *sample* – weights for the various ability-strata. When educational surveys of this kind were first attempted in 1964 no routine nationwide tests of mathematical attainments were available for England; now that they have become available, and even on an annual basis, they can be used to provide population weights for a TIMSS-type of survey using internationally specified questions.[16]

[15] G Ruddock *et al., Where England Stands* (… in TIMSS 2003), *National Report for England* (NFER, 2004), p. 25. The (previous) view expressed by PISA was very different. 'A subsequent bias analysis provided no evidence for any significant bias of *school-level* performance results but did suggest there was potential non-response bias at *student levels*' (PISA, p. 328, my ital.). To emphasise, this is different from the TIMSS conclusion that it was weaker *schools* that needed up-weighting to improve representation (pp. 9, 25). [We still have to check the PISA *Technical Report*, promised for 2004, but now not expected till end-2005.]

[16] It is difficult to find more than a trace of a reference to this re-weighting in the *international* TIMSS report, though it is quite explicit in the English *national* report; the same average scores for England are published in both reports. The TIMSS *Technical Report* (ch. 7, by M Joncas, p.202, n. 7) offers the following light: 'The sampling plan for England included implicit stratification of schools by a measure of school academic performance. Because the school participation rate even after including replacement schools was relatively low (54%), it was decided to apply the school non-participation adjustment separately for each implicit stratum. Since the measure of academic performance used for stratification was strongly related to average school mathematics and science achievement on TIMSS, this served to reduce the potential for bias introduced by low school participation'. The PISA report

The upshot is that, first, while the TIMSS primary survey results for England are *less* reliable than would appear from the way they were reported, those for secondary schools are *more* reliable. Secondly, sampling errors ought properly to be calculated for the TIMSS secondary school survey as for a stratified sample. Thirdly, the poor response-rates achieved in both these secondary school surveys might yet encourage a refusal by England – at a political level – to support any such future surveys; but we see here that what is first really required is more research into *sampling design*, that is, better use of population information collected in any event for general educational objectives, so enabling more accurate results to be attained at lower cost.[17]

Objectives of international tests

When these international educational tests were introduced nearly two generations ago, it was widely understood that their main objective was not – as it seems to have become today – to produce an international 'league table' of countries' schooling attainments, but to provide broader insight into the diverse factors leading to success in learning. Despite current popular emphasis on 'league table' aspects (but usually without corresponding emphasis on sampling errors!), much space is devoted in the present reports to students' 'perceptions', attitudes towards learning and their relation to success. But the reader often finds himself questioning the direction of causation; for example, we are told such things as that students who are happy with mathematics tend to do better in that subject: but perhaps causation is more the other way round – those who do well in that subject are happier, or more willing to declare their happiness. Similarly, much space is given to watching TV, and its association with test scores; with reading books, and so on. But little space is given in these reports to what topics are taught at each age, to what level, and to what fraction of the age-group (see Annex B on the implications for the confirmation of longer basic schooling life in the US); nor to such a basic 'mechanism' of school learning as to how students, who inevitably differ in their precise levels of attainment, are *grouped into classes* – despite the obvious concern of this feature of schooling to teachers, parents, policy makers and, not least, to students.

The relation between the *size* of a class and its average achievement is tabulated in one of the studies and well illustrates the issue of *direction of causation*. For classes of up to 24 students, an average score of 479 was recorded for England at Year 9; for larger classes of 25-32 students, the average score was higher at 511; and

does not discuss any such possible improved estimation procedure (but the *Technical Report* for PISA is not yet available).

[17] The above discussion of response rates has been restricted, for the sake of brevity, to the primary school survey. More or less the same applied to both secondary school surveys, as follows. For the TIMSS secondary survey, the participation rate of the 160 sampled schools (before replacements were included) was a pathetic 34 per cent (TIMSS, p. 358); for the PISA inquiry, directed to 450 schools, it was 64 per cent (PISA, p. 327, col. 1). For the US, which deserves special attention because of its greater financial sponsorship, the corresponding secondary school response rates were 66 and 65 per cent (but would their financial contribution have been as great if the *true* response rates had been published, i.e. after correctly allowing for replacement sampling as explained above?).

The English Department of Education issued *Notes* of guidance for media-editors explaining that their 'failure to persuade enough schools in England to participate occurred despite…various measures including an offer to reimburse schools for their time…' (*National Statistics First Release* 47/2004, p. 4, 7 December 2004). Note the term 'reimburse'; there is no suggestion of motivating a sub-sample of schools by a substantial net financial incentive.

for yet larger classes of 33 or more students, the average score was higher still at 552 (much the same applied in the other countries).[18] Higher attainments in *larger* classes have previously been frequently observed – contrary to the presumption that *smaller* classes would do better; this 'statistical relation' has generally been attributed to the widespread recognition by schools that slower/weaker/younger pupils should be taught in smaller 'parallel' classes where possible. Whether schools allocated higher attaining pupils to larger classes as efficiently as possible can be debated; but it is clear that no one (least of all, the present writer) would draw the policy implication that if children were only to be taught in larger classes then they would attain better results at lower costs! Much care is similarly necessary in drawing conclusions from other statistical associations noted in these studies.

For example, very strong conclusions were drawn by PISA on how the schooling system should deal with *variability* of students' attainments and capabilities. But let us first spell out realistically the issue of variability of pupils' attainments in a class from the teacher's point of view. *Some* variability of students' attainments within a class is unavoidable but, once a certain level of variability is exceeded, the pace at which the teacher can teach slows, as does the pace at which learning takes place, not least amongst those students who are weaker (weaker for whatever reason – born at the later end of the school-year, illness last year, slow learning in a previous school, difficulties at home that weigh on the student's mind…), often with consequent 'playing up' in class; eventually the teacher finds it better to divide his 'class' into explicit sub-groups, or 'sets', which follow a more or less different syllabus of tasks, with consequences for the pace of learning, and the costly need for teaching assistants. All this is of course familiar; and it might have been thought that an elementary calculation of variability of attainments within a class would have been a natural, obvious, useful – indeed essential – part of such inquiries.

But No! The PISA sample was deliberately based *not* on whole classes, but on all those aged 15 in a school – whichever Year or attainment-set they were in. In England, as in other countries where promotion from one class to the next is based strictly on age, it might seem that nothing much is at issue; but to rely on that would ignore the widespread practice of 'setting' students into groups by attainment levels – a practice that becomes more widespread at higher ages. In most other countries some reference to attainment level usually influences promotion from one class to the next. But nothing of this can be investigated with the help of PISA since its sampling was based not on whole classes – but purely on age, irrespective of class or teaching-group.

The TIMSS sample, on the other hand, was different since it was based on whole classes, and thus may be expected to be better for our concerns; but that does not take us out of the woods! For the reality of a 'class' becomes tenuous in the upper reaches of secondary schooling, as 'setting' by attainment becomes more prevalent. In large English comprehensive secondary schools, a dozen 'parallel' mathematics classes for each age or 'Year', varying according to attainment, is not unusual; for TIMSS, just *one* of those classes was selected by some 'equal probability' procedure, except that when some classes were very small they would be combined with another

[18] TIMSS, *Mathematics Report*, p. 266; the same applied also to the primary inquiry at Year 5, p. 267.

to form a 'pseudo-classroom' for sampling purposes.[19] A small class for very weak pupils might be combined with another class next higher in its attainments; or, for all we are told, could be combined with a small top set. In any event, no statistical analysis of the extent of student variability *within teaching groups,* nor of the whole year-group within a school, seems to have been attempted as part of either of these sample inquiries, despite the central importance of that issue to difficulty and success in teaching and learning, and its interest to teachers and educational planners.

Despite the sampling design of both inquiries being so perverse that variability of students *within* teaching groups cannot be computed (to repeat: PISA did not sample whole classes, TIMSS generally sampled only *one* 'ability-set' out of each year-group), very strong policy conclusions were voiced in the PISA report against any form of differentiation: they were against dividing secondary school pupils into different schools according to attainment levels (in England: Grammar schools and Comprehensives); they were against dividing pupils within schools into streams or attainment sets; and they were against grade repetition which they 'considered as a form of differentiation'.[20] Throughout there is the assumption that differentiation is the *cause* of lower average attainments, rather than seeing it the other way round – where teachers are faced with a student body that is unusually diverse, they use any organisational mechanism at their disposal to reduce diversity, and so make the group more teachable. In other words, greater variability within the class needs to be understood as the cause, rather than the effect, of lower attainments. All their conclusions were announced by PISA with great conviction – indeed, with great presumption – despite, as said, no calculations having been possible from their data on the variability of attainments within *teaching groups, classes* or *year-groups*..

The future

How was it possible, the reader will ask himself – but not too loudly, for fear of offending all concerned – for such large inquiries, with their endless sub-committees of expert specialists, to arrange their sampling procedures to *exclude* the possibility of calculating the variability of attainments for each class/teaching group? Any student of Kafka will readily invent his detailed scenario; but their essence is probably that the specialists were too specialised – in particular, the statisticians did not understand, or give sufficient weight to, the pedagogics of class-based learning; and the educationists did not give sufficient attention to the implications of the sampling procedures proposed. Perhaps most important, those in overall command were not sufficiently alive to such deficiencies in their varied specialists. Better 'generalists', rather than more specialists, seem to be required.

From the point of view of more representative sampling, future international inquiries of this kind, it can now be seen more clearly, need to be re-designed to incorporate sampling features of *both* these recent inquiries. We need to focus (a)

[19] TIMSS, [*International*] *Technical Report,* p. 121 (see also *Mathematics Report,* p. 349, which is also not very helpful); the English *National Report* has an Appendix on Sampling (p. 287) but regrettably says nothing on this vital aspect of sampling.

[20] Parents in countries with low between-school variances, we are told, 'can be confident of high and consistent performance standards across schools in the entire education system' (PISA, p. 163). 'Avoiding ability grouping in mathematics classes has an overall positive effect on student performance' (though it is conceded 'the effect tends not to be statistically significant at the country level'!), (p. 258). 'Grade repetition can also be considered as a form of differentiation' [and therefore to be avoided] (p. 264).

initially on the original variability of attainments of a complete age-group of students (variability due to genetic and socio-historical elements), perhaps estimated by the PISA-approach or by sampling two (? three) adjacent school-grades as in previous TIMSS inquiries; (b) then we need to estimate the extent to which variability is reduced within teaching groups as they have been organised in practice; (c) finally, we need to estimate the separate contributions of various institutional factors to that reduction in variability – secondary school selection, ability-setting within Year-groups, class-repetition. Differences among countries in these elements may yield valuable and empirically-based policy conclusions.

From the point of view of the substance of the inquiries, more focus and debate would be valuable on syllabus issues within mathematics. For example, what is the proper share of arithmetic in the overall mathematics curriculum at younger ages, and how should it vary for different attainment-groups? In some countries (Switzerland, Germany), at least until recently, the less academic group of students often become more expert in mental arithmetic skills as a result of their different curricular emphases; has the wholesale use of calculators really made this otiose? At what ages, and to what fractions of pupils, should specific topics be introduced such as simultaneous linear equations, quadratic equations, basic trigonometry or even basic calculus? No more than these few hints can be thrown out within the ambit of the present Note to indicate what a proper Next Step should include (see also Annex B on the anomalously low average attainments in mathematics at age 15 by the world's economically leading country).

A final question: how much public breast-beating by the organisations that have carried out the two recent inquiries will be needed before they should be considered eligible for participation in such an improved Next Step?

Acknowledgements and apologies

This Note has benefited from comments on earlier drafts by Professor G Howson (Southampton), Professor PE Hart (Reading), Professor J Micklewright (Southampton), Dr Julia Whitburn and many others at the National Institute of Economic and Social Research, London; I am also indebted to the National Institute for the provision of research facilities. Needless to say, I remain solely responsible for all errors and misjudgements.

I take this opportunity also of offering apologies to the individuals who have innocently participated in carrying out the underlying inquiries here reviewed; but those who planned those inquiries must fully accept their share of blame for the inadequacies complained of here, and for too often uncritically following what was done in previous inquiries – instead of *improving* on those practices.

ANNEX A
Some background on the two international educational inquiries of 2003

The **I**nternational Association for the **E**valuation of Educational **A**chievement (IEA) has been active since the 1960s in sponsoring internationally comparative studies of secondary schooling – subsequently also primary schooling – involving tests set to representative samples of students. The school subjects covered were mathematics and science, plus some separate inquiries into reading/literacy. The year-groups focussed on were eighth and fourth grades on the international grading (Europe and the United States), corresponding to Years 9 and 5 in the UK, that is, to ages of about 14 and 10. Sampling was based on school classes. Before 2003 the IEA had carried out similar inquiries in 1995 (in some countries also in 1999). The number of countries expanded over time to reach 49 in 2003. The studies are now managed from Boston College, Mass., with substantial financial support from the US government mainly for the central organisation; and financial support for the surveys in each country is provided locally.

Three reports were published by TIMSS on their 2003 inquiries:-

IVS Mullins *et al., TIMSS 2003 International Mathematics Report* (Boston College, 2004), pp. 455.

IVS Mullins *et al., TIMSS 2003 International Science Report* (Boston College, 2004), pp. 467.

MO Martin *et al.* (eds), *TIMSS 2003 Technical Report* (Boston College, 2004), pp. 503.

The second inquiry considered here was sponsored by OECD (Organisation of Economic Cooperation and Development), an international organisation set up in Paris to assist European post-war economic reconstruction and development, with heavy support from the United States. It conducted its first assessment of educational attainments in 2000 under the name **P**rogramme of **I**nternational **S**tudent **A**ssessment, PISA for short; and a repeat was carried out in 2003. I have not been able to find any written justification for setting up an inquiry so close in its objectives to the IEA's; but two differences – not necessarily justifications – should be noted. First, PISA focuses on a certain *age*, 15 – rather than school *Year* (or grade) as for TIMSS – for those included in its survey (though for some countries, Brazil, Mexico, that age is beyond compulsory schooling and only about half that age-group can be contacted!). On average, the PISA age is about a year above TIMSS, and closer to the age of entering the workforce. Secondly, the focus of students' questioning in PISA is said to be on the 'ability to use their knowledge and skills to meet real-life challenges, rather than merely on the extent to which they have mastered a specific school curriculum'; whereas the focus of TIMSS is closer to the school curriculum.[21] It still remains to be shown whether the practicalities of written examinations held in a school room makes any substantial difference to the outcome whether one kind of question is asked or the other.

The PISA inquiry covered mathematics and science, just as TIMSS; and also had questions on literacy (reading). PISA's emphasis in 2003 was on mathematics.

[21] PISA (2004), p. 20.

Results were published in:-

[No attributed authorship] *Learning for Tomorrow's World: First Results from PISA 2003* (OECD, Paris, 2004), pp. 476. A *Technical Report* is advertised (p. 302) as forthcoming.

Of the 48 countries included in PISA (49 in TIMSS, as said), 19 also participated in TIMSS. A full investigation, with access to individual questions and results in both inquires would be needed for a proper comparison; here we may note only that Hong Kong and Korea were near the top scorers in both inquiries (scores of 586, 589 in TIMSS; 550, 542 in PISA); in Europe, Netherlands and Belgium were about equally high (536, 537 in TIMSS – Flemish Belgium only; 538, 529 in PISA); and the United States was very slightly *above* average in TIMSS (a score of 504) and more than slightly *below* in PISA (483). The different mix of countries in the two samples also affects the standardised marks published: such comparisons between the inquiries are not therefore more than suggestive.

ANNEX B
The proper objectives of internationally comparative educational research

That the US, the world's top economic performing country, is found to have schooling attainments that are only middling casts fundamental doubts on the value, and approach, of these surveys. It could be that the hyper-involved statistical methods of analysis used (known as Item Response Modelling) is, as many have suggested, wholly inappropriate (see also my comment of 2003 on the PISA 2000 survey, p. 161). Or it could be, as two US academics have suggested, that the level of schooling does not matter all that much for economic progress; rather, it is 'Adam Smithian' factors such as economies of scale, and minimally regulated labour markets that allow US 'employers enormous agility in hiring, paying and allocating workers…'.[22] Or – my own view – that the typical age of school-leaving in the US, at some three years above that in most European countries (say, 19 rather than 16), has the consequence that schooling attainments at 14-15 hardly provides a clear indication of the contribution of final schooling attainments to subsequent working capabilities. An older typical school-leaving age means that teachers can sequence their courses of instruction in a more graduated way; and that the kind of question set in the PISA inquiries – designed to be close to everyday life – is indeed something for which US students aged 15 are less ready than their European counterparts. But that does not mean that at later ages their schooling has not served US students as a whole at least as well as their European counterparts. More time may be spent by US students in consolidating fundamentals. No investigation, or even discussion, of such issues is to be found in the official reports on these inquiries; and the absence of a sufficient number of published individual questions makes it impossible for the reader to take the issue further.

So far we have treated both surveys (TIMSS, PISA) as showing much the same schooling performance for US pupils – namely, as indifferent, or even weak, when judged in relation to the tremendous economic performance of that country. But we should also notice, and express surprise, that it is precisely in that survey with questions emphasising practical and 'real life' aspects, namely, the PISA survey, that average US 15 year-olds are shown at being *below* world average (whereas, in the more school-task oriented TIMSS survey, the US students were – even if only modestly – *above* the world average). Indeed, it is not too fanciful to suppose that the poor performance of US students in school-curriculum oriented questions in the earlier TIMSS surveys provided part (much?) of the impetus for carrying out a further survey with a more practical emphasis in its questioning. But, anyone who expected better results for the US via that line of questioning must have been sorely disappointed by the outcome. That outcome, it may also be concluded, casts further doubt on the value of repeating a PISA-type survey. Until wider-ranging pilot inquiries, on alternative lines, have been carried out and analysed, it is difficult to see that further inquiries of the present sort and scale are justified.

[22] See A P Carnevale and D M Desrochers, The democratization of mathematics, in *Quantitative Literacy* (eds. B L Maddison and L A Steen, National Council on Education and the Disciplines, Princeton NJ, 2003), esp. p. 24: 'if the United States is so bad at mathematics and science, how can we be so successful in the new high-tec global economy? If we are so dumb, why are we so rich?'