

Response to ‘Cautions on OECD’s Recent Educational Survey (PISA)’

RAYMOND J. ADAMS

ABSTRACT In his recent paper, ‘Cautions on OECD’s recent educational survey (PISA)’ (Oxford Review of Education, 29, 2), S. J. Prais questioned the outcomes of the Organisation for Economic Cooperation and Development’s PISA survey of the reading, mathematics and science attainments of 15-year-olds. Prais suggested that methodological flaws in PISA had resulted in an apparent improvement in the attainment of British students—particularly when compared to their Swiss and German counterparts. This paper responds to Prais’s criticisms, noting that when Prais’s conjectures are tested with empirical data they are not supported. Further it is noted that many of Prais’s criticisms are due to an incomplete understanding and knowledge of the methodology of international studies, and of PISA in particular.

INTRODUCTION

The purpose of studies such as PISA and TIMSS is to stimulate debate about the relative merits of policy choices that are made in education systems. International studies provide a natural complement to in-depth system level analyses by systematically examining educational outcomes, practices and relationships across educational settings.

Professor Prais’s major concern seems to be with understanding why PISA reports mathematics literacy scores for the United Kingdom and Switzerland that are essentially equal, while in the TIMSS 1995 study the average scores of Swiss students was some 40 points higher than the average scores for students from England. This indeed is an important question for the English and Swiss education authorities and deserves considered attention. At the same time, it appears that many of the claims made by Professor Prais are based on misunderstandings related to the methodology underlying these international studies and a lack of research of the relevant technical documentation. The objective of this article is to provide some methodological background that will allow proper examination of the claims made by Professor Prais.

Before turning to the specific issues raised by Professor Prais, the article will first review the key premises that underlie Professor Prais’s argument. First, the PISA and the TIMSS assessments are not statistically linked and include different sets of countries. The meaning of the TIMSS scale scores was defined in 1995 so that 500 was the mean of the countries that participated in that study, and 100 was the standard

deviation of students' scores. The TIMSS 1999 results were linked to this scale and are therefore directly comparable with it. The PISA scale was set so that the mean of the participating OECD countries was also 500 and the standard deviation of student scores in those countries 100. However, because the set of countries that participated in each of these studies was quite different, and because the target populations in each country were different, there are no reasonable grounds for suggesting that either the mean of 500 on each scale, or the standard deviation (the unit size) are in any way comparable.

Second, Professor Prais notes that the TIMSS 1999 study did not show that the students in the UK had improved in their performance relative to those in Switzerland, France, Belgium (Flemish Community), the Czech Republic and Hungary. Since neither France nor Switzerland (nor more than half of OECD countries [1]) participated in TIMSS 1999, there is no scientific basis for comparing trends in TIMSS performance between England and these countries. It would also have been helpful to the reader if it had been made clear that the performance of England [2] relative to that of the Flemish Community of Belgium and the Czech Republic was similar in PISA 2000 and TIMSS 1999.

Third, it is argued that international surveys have consistently shown that UK students show low average performance levels and a wide dispersion of scores. An examination of the various international assessments reveals, however, a much more complex and varied picture. In the 1964 IEA mathematics assessment, England showed average performance among 13-year-olds and above average performance in the final year of secondary school. Performance in the 1974 IEA science study and the 1974 IEA reading study was also at the average level. Finally, England scored seventh among the 26 OECD countries that participated in the 1995 IEA TIMSS science assessment and showed similar performance levels in the 1999 IEA TIMSS assessment. In the most recent of the IEA studies, PIRLS 2001, a study of reading literacy at fourth grade, England ranks third out of 15 participating OECD countries and third out of 35 participating education systems.

Fourth, and finally, it is suggested, without any supporting argument or citation, that Switzerland is a model education system within the European context. It seems that such a judgement would always depend on the criteria that are used to identify and define a model education system. In the OECD context, such criteria are usually defined with regard to the level and distribution of educational outcomes. Since the performance of Switzerland in international comparisons has not been particularly high, despite extraordinary investments in education, Switzerland has not generally been among the countries that were considered as model school systems among the policy and scientific expert groups working within the OECD context at least. Obviously, Professor Prais may have different criteria on which he bases his judgement that Switzerland is a model education system but, unless he makes these criteria transparent, it will not be possible to discuss these in a scientific manner.

Let us now turn to each of the specific concerns that are expressed by Professor Prais. He lists five major concerns:

- (a) the nature of the mathematics questions;
- (b) differences between TIMSS and PISA in target population definitions;
- (c) the representativeness of the UK's sample of schools;
- (d) the representativeness of the UK's sample of students in schools;
- (e) the scaling and data processing errors.

NATURE OF THE MATHEMATICS QUESTIONS

Prais is incorrect when he asserts that the PISA mathematics literacy domain definition is one that is primarily concerned with everyday functional mathematical literacy [3]. One of the strengths of PISA is that, as opposed to previous international assessments of this kind, it began the study with the establishment of a theory-based assessment framework that defined and operationalised the knowledge and skills to be assessed. Had Professor Prais presented a more detailed discussion of the frameworks and the assessment material, he would have reported that each assessment area in PISA was defined to reflect: first, knowledge of a set of fundamental skills and understandings that are specific to the respective assessment area, and second, a capacity to use those skills to address real-life issues and problems. It is therefore correct that PISA sought to assess students' abilities to apply mathematical concepts, skills and understanding to authentic problems that arise in real world settings. However, to deduce from this that the main or even a major objective of PISA was the assessment of 'everyday functional mathematical literacy' is entirely inappropriate.

It is also quite explicitly stated that authentic settings are not primarily focused on day-to-day (everyday) applications of mathematics. Instead, the primary focus of PISA is on the ability to apply mathematical knowledge and thinking to a whole variety of situations, including what PISA calls inner mathematical settings [4].

The PISA approach, which is indeed different from that of TIMSS, was deliberately adopted for various reasons. The most important distinction between PISA and TIMSS lies in the assessment development process. The starting point of PISA was a review of the intended outcomes of schooling rather than a review of the specifics of the particular curricula of participating countries. PISA adopted this approach because of concern that focusing only on the common elements of school curricula across a range of countries would limit the assessment to a set of intersecting features that are perhaps not objectionable to any participant but are not likely to be of high value or interest to educational policy-makers or practitioners. This is because the focus on the common international denominator of curricular materials limits the potential of international studies in enabling countries to learn from curricular differences and instructional innovations pursued by other countries. Furthermore these intersecting areas are simply a coincidence of the participating group of countries—something that varies widely from study to study and over time.

The debate of the relative merits of the two approaches is of course open to others. Our own view, however, is that it is a worthwhile objective to seek to assess the extent to which pupils are able to apply to a rich array of problems, what they have learned in the compulsory phase of their education. This is particularly relevant when pupils come to an age when leaving school is an option for most. It is not that this information is more or less important than knowledge about the extent to which a curriculum has been absorbed, but that it is different, relevant information that sheds light on the relative preparedness of pupils for the real world at the chosen age.

Prais makes a variety of observations concerning the German education system and its additional PISA-related studies to support his argument that a TIMSS-like approach would be superior to the PISA. Indeed it is true that the German authorities added an additional 86 mathematics questions to the international PISA survey, and that these questions were more directly curriculum-focused. This approach was adopted so that assertions like those made by Prais could be explicitly tested. Of key interest to them was whether the international scale and a national (more curriculum-focused) scale

would be distinct and provide different results. In the German national report it was shown that while there was a difference in the flavour of the items, there was clear evidence that both assessments were fundamentally assessing a common underlying outcome [5].

In view of the comments about links to TIMSS and the PISA test being simply a test of intelligence or common sense (*Oxford Review of Education*, 29, 2, pp. 144–145), perhaps it is worth extending the discussion of the findings of the German study.

The German study noted that there were indeed some differences between the PISA test and its national, more curriculum-focused, test. The national items have a somewhat lower correlation with PISA reading scores, yet they had a *higher* correlation with an intelligence test that was also administered as part of the German PISA assessment. Furthermore, in the German study, some 2174 students worked on 15 TIMSS-items in addition to the PISA items (31 international and 14 national). The correlation between TIMSS and PISA (in this combination, the international item parameters were fixed) was 0.91. The conclusion was that, at least in Germany, PISA and TIMSS lead to similar estimates of student outcomes [6].

It seems, therefore, that in each case where an assertion about the nature of the PISA construct made by Prais was empirically tested, his argument was found wanting [7].

In his discussion of the nature of questions asked, Prais included a discussion of three sample items. We shall not extend this response by discussing each item in detail, but we shall make some remarks on Example A (p. 142) since such remarks will again highlight Prais's misunderstanding of the PISA Mathematical Literacy construct.

In Prais's comments on the pizzeria problem he did not seem to realise that the item is not simply meant to assess what students would do in a pizza shop. That would be much too particular—and would perhaps confound the central purpose of this item by introducing a computational dimension. Rather we were testing how students could use their mathematical knowledge (including their mastery of various hands-on calculation techniques) to answer questions set in a real-world context. Some students might try to make conscious use of the mathematical topic of *similarity* to solve this problem. Others may apply the skills arising from the similarity concept, which may have been integrated into their *automatic* armoury of skills, while others might resort to a *counting squares* strategy. Using a term such as *similar figures* in the question may well have hampered students in their approach—the absence of any such reference in the OECD commentary is surely not relevant. Indeed the main mathematical idea needed here relates to the functional relationships, and the difference between linear and quadratic growth [8].

At the risk of being repetitive, Mathematical literacy, and the emphasis on authentic use of mathematical concepts in situations that arise in the real world, are not the same as everyday or ordinary problems of living.

THE TARGET POPULATION

The objective of PISA is to provide an assessment of the cumulative yield of education and learning at a point at which most young adults are still enrolled in initial education. A major challenge for an international survey is to operationalise such a concept in ways that guarantee the international comparability of national target populations.

Differences between countries in the nature and extent of pre-primary education and care, the age of entry to formal schooling, and the institutional structure of education systems do not allow the definition of internationally comparable grade levels of schooling and thus render the kind of grade-based comparisons that Professor Prais

favours inadequate for international comparisons. Some previous international assessments, including TIMSS, have defined their target population on the basis of the grade level that provides maximum coverage of a particular age cohort. The problem of this approach is, however, that slight variations in the age distribution of students across grade levels often lead to the selection of different target grades in different countries, or between education systems within countries, raising serious questions about the comparability of results across, and at times within, countries. In addition, because not all students of the desired age are usually represented in grade-based samples such as those used in TIMSS, there may be a more serious potential bias in the results if the unrepresented students are typically enrolled in the next higher grade in some countries and the next lower grade in others. This would exclude students with potentially higher levels of performance in the former countries and students with potentially lower levels of performance in the latter.

In order to address this problem, PISA uses an age-based definition for its target population, i.e. a definition that is not tied to the institutional structures of national education systems: PISA assessed students who were aged between 15 years and 3 (complete) months and 16 years and 2 (complete) months at the beginning of the assessment period and who were enrolled in an educational institution, regardless of the grade levels or type of institution in which they were enrolled, and regardless of whether they were in full-time or part-time education.

As a result of this population definition, PISA 2000 makes statements about the knowledge and skills of a group of individuals who were born within a comparable reference period, but who may have undergone different educational experiences both within and outside school. In PISA, these knowledge and skills are referred to as the *yield* of education at an age that is common across countries. Depending on countries' policies on school entry and promotion, these students may be distributed over a narrower or a wider range of grades. Furthermore, in some countries, students in PISA's target population are split between different education systems, tracks or streams.

If a country's scale scores in reading, scientific or mathematical literacy are significantly higher than those in another country, it cannot automatically be inferred that the schools or particular parts of the education system in the first country are more effective than those in the second. However, one can legitimately conclude that the cumulative impact of learning experiences in the first country, starting in early childhood and up to the age of 15 and embracing experiences both in school and at home, have resulted in higher outcomes in the literacy domains that PISA measures.

As Prais also raises the issue of exclusions let us consider the comparison of the TIMSS and PISA coverage and exclusions. The values reported in Table I are extracted from the technical reports of the three respective studies.

The UK exclusions figures for PISA are somewhat higher than they were for Germany and Switzerland. The 5.00% figure for the UK is made up of 2.4% of school exclusions (special schools), 2.0% of student exclusions due to special needs and 0.5% of student exclusions for other reasons. In Switzerland the values were lower, largely because these countries used the PISA special booklet and therefore reduced their rate of school level exclusions. However, it should be noted that the PISA exclusions for countries all satisfied the PISA standards of a ceiling of 5% of exclusions. In contrast, all of the TIMSS figures, with the exception of England in TIMSS 1999, had much higher exclusions. This would suggest a bias upward in the TIMSS 1995 results for Germany, Switzerland and England relative to the PISA figures.

TABLE I. Coverage and exclusion rates for TIMSS 1995, TIMSS 1999 and PISA 2000

Country	Study	Exclusions (%)	Coverage of National Target Population (%)	Coverage of ALL 15-year-olds (%)
England	TIMSS 1995	11.00	100	NA
England	TIMSS 1999	5.00	100	NA
England	PISA 2000	5.36	95	93
UK	PISA 2000	5.00	95	88
Germany	TIMSS 1995	9.70	100	NA
Germany	PISA 2000	1.68	98	89
Switzerland	TIMSS 1995	5.30	100	NA
Switzerland	PISA 2000	2.74	98	89

1. TIMSS 1995 figures are the upper grade of population B
2. England—not the UK—participated in the IEA studies
3. Neither Germany nor Switzerland participated in TIMSS 1999
4. PISA presents a number of coverage indices; the two figures we report are index 1 and index 3. Index 1 describes the coverage of 15-year-olds in schools and index 3 describes the coverage of all 15-year-olds, not just those in schooling. TIMSS has no index comparable to index 3.

On what grounds then is it an *injustice* to pick an arbitrary age (15-year-olds) and make comparisons across countries, and on what grounds does the TIMSS definition lead to better grounds for comparison? Let us begin discussing this by considering the TIMSS population definitions and their cross-national comparability. The population definition for TIMSS 1999 was the modal grade for 13-year-old students. For TIMSS 1995 it was the two grades that covered the largest percentage of 13-year-olds [9]. In what sense does such a population definition provide comparability across countries? Does it result in students who have been engaged in schooling for a matched length of time [10]? Does it result in students who are of comparable ages? Does it result in students who are comparable in terms of the number of remaining schooling years? A closer analysis reveals that the answers to all of these questions are negative.

In TIMSS 1995 the Swiss students had been in school for one to two years (depending upon the Kanton) fewer than their counterparts from England, while the German students sampled in TIMSS 1995 had been in school one year less than English students. The average age of the German sample was 14.8, whereas for England the average was 14.0. In fact in TIMSS 1995 the range of average ages for students was 13.6 to 15.7! The range of the number of years of schooling was from seven to nine years. In TIMSS 1999 the range of average ages was 13.8 to 15.5 while the range of the number of years of schooling was seven to nine and a half years. On what grounds then does TIMSS provide stronger cross-country comparability—the key element of international studies?

When comparing PISA and TIMSS results for England, Switzerland and Germany, it should also be noted that 60% of the English students were already in Year 11, the second year of upper secondary education. In Switzerland and Germany over 60% were still in grade 9, the last year of lower secondary education. This fact, not explicitly mentioned in Prais's article, provides a possible explanation for the differences in the relative performance of each country in PISA 2000 when compared to TIMSS 1995.

While we would suggest, therefore, that it is clear that a grade based definition does

not provide a population definition that has any common-sense advantages, in terms of comparability, over an age-based definition, it might still be reasonable to assume that there are pedagogical, educational policy or methodological advantages to such a definition.

Prais makes the claim that the age-based approach is *unfair* to the continental countries because it penalises them for recognising that some students are slower in maturing and may, as a consequence require more years of instruction. One can, however, equally argue that countries with high grade-retention rates will be favoured by a grade-based sampling design because their students will, on average and other things being equal, have had more years of instruction. In grade-based studies countries with grade repetition have *modal-age* repeaters who attend lower grades that are not tested, while repeaters who attend the grade tested are at least one year older than their *modal-age* classmates. By using an age-based sample, PISA confirms the literature about grade repetition, which indicates that retention has, in general, a negative relationship with achievement [11].

From a methodological point of view, Prais argues that because a grade-based population definition permits grade-based sampling that the sampling will be less disruptive and that it permits an examination of within-class variance in attainment. Prais is correct in stating that this is typically less disruptive to the school timetable, which might result in the higher participation of schools. However, he fails to acknowledge that in sampling only one class within a school, there is the risk of obtaining a less efficient sample of students than if the students are selected at random from across the whole year-group. This design would typically lead to larger design effects due to the increased clustering of pupils and hence would lead to less precise survey estimates. The current PISA sample therefore would likely give more precise estimates than the design Professor Prais is advocating.

Finally, it is indeed true that research on school effects indicates the importance of breaking down the variation in student performance into three components: between-school variance, between-class within school variance and variance between students within schools [12]. It is important to recognise however, that neither the grade-based sampling (as in TIMSS) nor the student sampling (as in PISA) can provide such a breakdown. By sampling a single intact class in each school, TIMSS-like designs confound the between school and between class variance components. By sampling students at random within schools a PISA-like design confounds the between-student and between-class variance components. Therefore neither design can produce all three variance components [13].

REPRESENTATIVENESS OF THE UK'S SAMPLE OF SCHOOLS AND STUDENTS

The PISA report is clear in its declaration that the UK school response rate of 59% (prior to replacement) and 82% after replacement was a matter for concern, as was the student level response rate of 81%. Such low response rates bring a threat of bias that cannot be formally addressed through any follow-up survey or studies.

Table II shows the response rates for England (and for PISA the UK) and for Germany and Switzerland in the TIMSS 1995, TIMSS 1999 and PISA 2000 studies. In each case the figures for England were markedly lower than the figures for Switzerland and Germany and the England school response rate prior to replacement was consistently low—in fact for the UK, PISA had the highest school response

TABLE II. Response rates for TIMSS 1995, TIMSS 1999 and PISA 2000

Country	Study	School	School after	Student (%)	Overall (%)
		Before Replacement (%)	Replacement (%)		
England	TIMSS 1995	56	85	92	78
England	TIMSS 1999	49	85	90	77
England	PISA 2000	59	82	81	66
UK	PISA 2000	61	82	81	66
Germany	TIMSS 1995	72	93	87	81
Germany	PISA 2000	94	94	86	81
Switzerland	TIMSS 1995	93	95	99	94
Switzerland	PISA 2000	92	96	95	91

rate prior to replacement. Two other observations that might be worth noting are the increase in the German school response rate prior to replacement in PISA and the decline in the student response rate in the UK for PISA [14]. Following Prais's arguments, does this suggest that the apparent decline in German performance between TIMSS 1995 and PISA 2000 was the result of improved response rates in PISA or perhaps, as discussed above, was it the lower number of exclusions in PISA when compared with TIMSS 1995? The fact is that one cannot know the direction or exact magnitude of bias caused by non-response—although bounds can be placed on its magnitude.

The Characteristics of Non-responding Schools

In PISA, unlike TIMSS, follow-up studies were undertaken to bring additional evidence to bear on the possible impact of non-response bias. Without any supporting evidence, Prais asserts that the missing schools were probably low attaining schools. The national report for England and the PISA Technical Report show that this was not likely to be the case. The report examined the response rate among schools, before replacement sampling occurred, according to the proportions of their students attaining at least five GCSEs at grade C or above [15]. Among LEA schools, response was highest (68%) in those schools with moderate attainment (i.e. with between 34 and 42% of their students getting five or more good GCSE passes) and was lowest in the highest attaining schools (47% where at least 60% of students had five or more such passes). Lower attaining schools had participation rates of 60% (schools with GCSE pass rates of 25–33%) and 56% (schools with GCSE pass rates of 25% or under). A more detailed analysis showed there was no significant relationship between schools' average GCSE point scores and whether or not they took part in the survey. Furthermore, an additional study that is documented in the PISA 2000 National Report reveals that there is no substantial evidence of bias in terms of the percentage of students eligible for free school meals which can be considered a reasonable proxy for the socio-economic background of the school's student population.

Further in note 29 Prais suggested examining school participation according to its proportion of students who were low attainers. The national report showed that no significant correlation was found between school response rates and their average point scores at GCSE level [16].

Again we see that when the assertions made by Prais were tested with available data they were not supported.

The Characteristics of Non-responding Students

Professor Prais made the same assumption about student participation as he did about schools, namely that it was the low attainers who refused to take part. While this is certainly a valid concern, there is documentary evidence from telephone calls from parents and students, and feedback from school staff, that students who refused to take part were not only the low attainers. Many cited the interference of PISA with preparations for their GCSEs in which they were keen to do well.

Prais's suggestion of gathering information about responding and non-responding students is a good one, and was considered in detail during the preparations for PISA 2000. These have been described in the UK national report [17]. A pre-requisite was that the data collected for comparing these students should be objective and standardised across schools. This ruled out the collection of school-based data describing in which mathematics set students were. The availability of Unique Pupil Numbers for the PISA 2003 cohort will allow us to link the PISA sample with national assessment data, and will be important in describing the coverage of the PISA and allow comparisons to be made for the current cycle of PISA.

Replacement Sampling

Prais raises a number of points on the subject of replacement sampling. The methodology behind selecting replacement schools so that their characteristics are matched to those of the selected schools, and ensuring these represent a random sample is well established [18]. This is achieved by listing schools according to important characteristics at the time of drawing the random sample of schools to be approached. PISA sampled two replacement schools for each selected school. These were the ones on either side of the selected school on the list. As the replacement sample was drawn at the same time as the main sample, it has also been drawn at random [19]. This procedure has been described in the UK national report in the appendix on sampling.

Prais's argument that this replacement procedure brought the PISA sampling process closer in part to quota sampling used in commercial work is groundless. Quota sampling has no element of randomness and as such one cannot know the probability with which a respondent has been selected and therefore how many others in the population they represent. In PISA, the probability of a school and student being selected is known, and this is why it is possible to make statistical inferences from the survey findings.

There is no standard method of calculating response rates when replacement sampling has been used. However, survey and non-response experts agree that approach (iii) as outlined by Prais is the most logical for representing response *at the initial sampling stage*. This is why the OECD and national reports always present the before-replacement response rate. The after-replacement response rate has been calculated using approach (i), as this gives an indication of what the potential non-response bias might be, assuming replacement schools are perfect matches for non-responding schools. Approach (i) is also generally used to calculate response at the final stage. Approach (ii), which is favoured by Prais, is not a method we have found to be widely used, as it does not reflect the principles underlying the use of replacement samples, which attempt to reduce non-response bias. Rather, approach (ii) treats the replacement schools as if they are an add-on to the sample, trawling through the remaining un-sampled schools in an attempt to increase the numbers participating [20].

SCALING AND DATA PROCESSING ERRORS

We will be somewhat briefer, yet more specific in response to the various points made by Prais about the scaling methods used in PISA. It is indeed a pity that this section has been included in the paper at all. As Prais states regarding his own understanding in annex note 9: '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'.

Again let us remind the reader that Prais is raising concerns because he does not see the apparent narrowing of the gap between Switzerland and Britain as a credible observation and he is looking to the scaling as a reason as to why there might be a change from previous international studies to PISA 2000 [21]. What Prais fails to note in his annex discussion is that the scaling procedures used in PISA 2000 were identical to those that were used in TIMSS 1995 and were fundamentally the same as those used in TIMSS 1999 [22]. Let us turn now to some specific observations.

Prais seems to work from an expectation that the national averages were some direct transformation of item percentage correct values (item facilities). This is not the case as modern scaling techniques provide an indirect transformation of student percentage correct scores to a scale score. Students are assigned scale scores, not items, because the purpose of such studies is to make inferences about what students can and cannot do, as well as looking at distributions of students' scores and examining relationships between student scores and other characteristics of those students, or characteristics of their learning contexts. This can only be done if students, rather than items, are assigned scores.

On pp. 160–162 Prais makes a number of incorrect assertions. First, he states that items were deleted from the final scaling for certain countries because the scaling model was not appropriate. This assertion is incorrect. Item deletion occurred only where an item was found to be translated erroneously, or where there was an error in the item format or if there was a misprint in a country's booklet [23]. Second he erroneously reported that multiple booklets were used to prevent cheating. This is not correct. Multiple booklets were used to maximise the coverage of the assessment domain while placing limited demands on individual students. Third, he complained that the documentation of the sampling was incomplete because the compendia of item statistics included the statement 'item statistics not based on whole population'. This annotation occurred for those items not included in the PISA special booklet in countries that used that booklet [24]. Because the standard booklets were not administered in special schools it follows that the reference population for the items not included in the special booklet (that is items included in the standard booklets only) was a subset of the full population. Fourth, Prais erroneously asserted that the items were weighted by the number of responses to them. This is erroneous and a misrepresentation of the scaling methods that were fully described in the technical report and that have been fully described in a large body of psychometric literature. Fifth Prais suggested that markers were provided with some rule to decide whether 'a candidate simply ignored a question because it was difficult, or because he ran out of time'. This, too, is erroneous. The procedure employed has been fully described on page 130 of the technical report. Finally, Prais asked why PISA did not publish the simple percentage of questions answered correctly by pupils in each country. These results have indeed been published and can be downloaded from http://pisaweb.acer.edu.au/oced/oced_pisa_data_s1.html.

CONCLUSION

There are legitimate questions to be asked about the apparent differences among TIMSS 1995, TIMSS 1999, PISA 2000 and more recently PIRLS. But the article by Prais has not taken a systematic view of this issue. It simply argued that PISA results for the UK must be biased upwards, because the outcomes did not meet with the preconceived ideas of the author, ideas which when tested with empirical data were consistently refuted.

ACKNOWLEDGEMENTS

PISA is a very large project and it cannot succeed without significant contributions being made by many individuals. Similarly this response has drawn heavily on the input of a number of individuals who have been central to the success of PISA. Helpful assistance with this reply was provided by: Alla Berezner, Aletta Grisay, Eckhard Klieme, Christian Monseur, Keith Rust, Andreas Schleicher, Wolfram Schulz and Ross Turner.

NOTES

- [1] The reader should be aware that many of Europe's leading economies chose not to participate in TIMSS 1999. For example, none of the following countries participated: Germany, France, Switzerland, Austria, Spain, Denmark, Norway, Sweden, Finland.
- [2] The participant in PISA was the United Kingdom, whereas England participated separately in the TIMSS studies.
- [3] On p. 141 the phrase 'Knowledge and Skills for Everyday Life' is used with capitalisation that implies this is the title of the PISA initial report. Perhaps it is worth alerting the reader to the fact that this is *not* the title of the initial report. The report title is *Knowledge and Skills for Life: First Results from PISA 2000*.
- [4] There appears to be an erroneous quote in Prais's note 5. The reference is to page 23 of the initial report. Such a quote cannot be found there.
- [5] The national and the international scale had correlations of 0.91 and the item parameters from the two-dimensional and the one-dimensional solution are almost perfectly linearly related ($r = 0.98$). For details see J. Baumert *et al.* (2001) *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (Deutsches Pisa-Konsortium, Eds).
- [6] Note also that in detailed comparisons between eastern and western Germany it is shown that eastern Germany has a relative strength in the national tasks, while western Germany has relative strengths in the international part. For details see J. Baumert *et al.* (2003) *PISA 2000: Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (Opladen, Leske + Budrich). The authors argue that these differences can be traced to different pedagogical traditions. These differences are very small, and they do not have any substantial impact, say, on the relative ranking of the 16 German Landers.
- [7] Perhaps this is not surprising as our advice on the German education system is that Prais is exaggerating the amount of mathematical instruction for apprenticeship students. Firstly, students from lower tracks (*Hauptschule, Realschule*) do not automatically enter an apprenticeship school. They need to find an

- apprenticeship—which has become problematic in the last few decades. Secondly, apprenticeship students have one day of instruction per week, so the amount of mathematical instruction is necessarily limited to this segment. To call these ‘final years of obligatory full-time schooling’ is hardly a correct statement.
- [8] As for note 9 (p. 155) regarding the use of currency in this item, the reader is referred to the translation chapter by Grisay in the PISA Technical Report (R. Adams and M. Wu, *PISA 2000 Technical Report*, OECD, Paris, 2002). There the reader will find that potential problems with different currencies raised in note 9 are well recognised, and led to universal use of the fictitious zed for such items. Only when the numbers do not matter would countries be allowed to consider substituting their national currency.
- [9] In fact many of the leading TIMSS 1995 results are reported for a sub-population which is in fact the upper of the two grades that contains the most 13-year-olds. It is this sub-population that is most commonly discussed.
- [10] In fact experience with international studies has shown that it is not possible to well define the notion of years in schooling. Induction process and the way in which something akin to formal instruction is introduced vary widely across countries. This results in the near impossibility of defining a commencement date for (formal) schooling.
- [11] See for example HOLMES, C.T. (1990) Grade level retention effects: a meta-analysis of research studies, in: L.A. SHEPARD & M.L. SMITH (Eds) *Flunking Grades. Research and Policies on Retention* (Bristol, Palmer Press).
- [12] In fact Prais (on page 145–146) suggests only the importance of looking at variation with classes.
- [13] Variations on the basic one-class per school design have been used from time to time in international studies in an attempt to deal with this issue, although they have met with limited success.
- [14] Note that the improved German response could be taken as evidence against Prais’s assertion that grade-based sampling would lead to a higher response rate.
- [15] Table A.2, Office for National Statistics (2001) *Student achievement in England: Results in reading, mathematical and scientific literacy among 15-year-olds from the OECD PISA 2000 study*.
- [16] Appendix A: sample design, response and weighting, Office for National Statistics (2001) *Student Achievement in England: results in reading, mathematical and scientific literacy among 15-year-olds from the OECD PISA 2000 study*.
- [17] Office for National Statistics (2001) *Student Achievement in England: results in reading, mathematical and scientific literacy among 15-year-olds from the OECD PISA 2000 study*.
- [18] D. ELLIOT (1933, July) The use of substitution in sampling, *Survey Methodology Bulletin*, no. 33, OPCS. Further it is a methodology that is consistently applied in IEA studies.
- [19] This assumes that all schools in the sampling stratum have the same likelihood of responding.
- [20] The PISA 2000 Technical Report, pp. 135 and 136, gives a whole table of (iii), followed on pp. 138–140 by a whole table of (i). In response to Prais’s question however, (ii) would be the correct answer if the replacement sample were a random sample from the entire population, of size equal to the number of original refusals, and all approached to participate. The correct answer by the way is (iv), somewhere between (i) and (iii).

[21] Prais, p. 160.

[22] In fact the original TIMSS 1995 scaling and the PISA 2000 scaling were performed with the same black box, and by the same data analysts!

[23] It is indeed unfortunate that the study's technical report does not report why the items were deleted.

[24] The PISA special booklet was used in special schools; it is used to minimise exclusions as discussed earlier.

Correspondence: Ray Adams, Australian Council for Educational Research, Private Bag 55, Camberwell, VIC, 3124, Australia. E-mail: Adams@acer.edu.au

