Denis MEURET

Some considerations on how far we can trust PISA 2000

Contribution to the National Agency for Tackling Illiteracy's international colloquium on the evaluation of low levels of competence, Lyon, 5 November 2003

I have been invited to speak here on the basis of a study I did on the possible reasons why the performance of French pupils was below that of eight countries, Anglo-Saxon, Asian and Scandinavian in the PISA 2000 study (Meuret, 2003), a study which obviously rests on the assumption that "cultural biases", the issue we are discussing here, if they do exist, are not such as to invalidate this type of comparison.

I would therefore like to present this study (3) here and to press the case for the legitimacy (4) of these international comparisons starting with PISA 2000. First though I would like to set this discussion in its political context — to say why I see PISA as good news in political terms (1) — and technically — in the sense that it is legitimate and desirable to criticise the way PISA was conducted, provided however that we take account of the progress it has made (2).

(1) A political advance

I see it as good news that 43 countries[1] measure the competence of their pupils on a common scale at the end of compulsory schooling and hence that they are in agreement on the nature of the competences that matter as well as the means for measuring them. It would be wrong to see as a simple technocratic process the fact that 43 governments collectively agree as legitimate a particular idea of reading, maths and science literacy, still more so if there is a compromise between different educational traditions. In a world that is subject to the crypto-religious conflicts that we are going through, that seems to me to be important. Furthermore, in a world run by economic forces and the inadequately controlled exercise of might, the existence of this rational process of reaching agreement, whatever might be its imperfections, seems to me intrinsically positive.

From the political point of view, for me, PISA has two important features.

The first is the fact that the construction of the items was regulated at the same time by experts, by groups drawn from the countries and a group of representatives from education ministries. The country representatives validated the theoretical canvass which underpins the nature of the tests; on this basis the countries sent suggested test items, from which a choice was made by specialists in evaluation, then these items were sent to the countries for scrutiny and submitted to the expert group for each subject (reading, maths, science); The resulting test was piloted in each country, this

---

This translation by Jason Tarsh, August 2005. The French original is at: www.anlci.gouv.fr/documents/actes112003/meuret.pdf

[1] Including the 12 countries in "PISA+".

pre-test leading to a set of items revised in the light of each country's comments, then re-submitted to country representatives[2]. And, added to all this, is the transparency of the procedures (they are described in the technical report, downloadable at no charge from the PISA web-site) and the speed with which the data were made available to researchers and to the general public, two factors which are politically significant and not just technical matters.

The second is the nature of the skills measured by PISA. The difference between the approach of the IEA studies and that of PISA is not, just, a simple technical matter. The IEA approach is that of searching for the greatest common denominator across the curricula of the different participating countries while that of PISA is not, as Professor Prais accuses it of, "testing the maths skills needed for everyday life" — with the implication "low-level" — but, as Adams, one of those in charge of PISA counters, "understanding of fundamental ideas and the ability to handle real-life problems" which is not the same thing at all. In brief, in the case of PISA, countries are agreed that they are aiming, each in their own way, a certain basic level of competence at the end of compulsory schooling and they define this basic level in terms of providing the intellectual capacity for everyday life, a very secular definition (Meuret, 1998) of the aims of education and particularly welcome in a country such as France where it is very much the elite view to think, for example, that "teachers have no other duty than to the internal logic of their subject" (Régis Debray, 1998).

(2) Technical advances

General opinion is that PISA has raised the technical standard of international assessments, in part, probably because of the political importance of its results. This progress has affected many aspects, for example, the definition of the population (an age group rather than a school grade), the rigour of the sampling procedures, the coverage (fewer very low performing young people were excluded than in previous assessments, in particular by the provision of tests which do not include the most difficult items, strict procedures for the exclusion of very low-attaining pupils), marking (markers were asked to rate one question after another and not one pupil after another to avoid halo effects which can lead a marker to be influenced by the quality of the pupils' preceding answers, numerous checks were made on the inter-marker reliability, which showed a very high degree of consistency between marks on the open-ended questions, of 91% on average and 80% in the countries where this was least good).

---

[2] That 60% of items originate from Anglo-Saxon countries or that 88 items out of 141 are in English needs to be seen in the context where countries had at least three opportunities to assess whether or not these items were educationally appropriate for them. These precautions explain why, in spite of this initial imbalance, France's ranking hardly changes when countries are ranked on all items, on only the set of items which France itself judged were most suited to evaluating its pupils or even only with the items that other countries, including the Anglo-Saxons, judged were most suitable to assess their pupils. Thus, France does better in the ranking on the items that New Zealand identified than on those it itself highlighted (OECD 2001). This result does not completely rule out the existence of cultural biases but excludes that they are so significant that the experts of the different countries perceive them when examining item content.

I would here like to dwell a little on two of these areas of progress, which closely relate to our theme. The first is the definition used of reading literacy. Lafontaine (2000) has compared the concept of reading used in the first international studies (starting in 1971) and in PISA. She shows how PISA, then PIRLS[3] were the first to set the definition of reading on a genuine conceptual framework, drawing on significant advances in research in the 1980's. This framework defines reading as an "interactive process between a text, a reader and a context: the reader, in order to achieve their purpose uses strategies and draws on their culture to work out the meaning and to respond to the text". Hence the three scales (retrieving, interpreting and reflecting on a text) and that they are not in an hierarchy. Hence, also, the use of questions to which there had to be a constructed response. This was not a simple concern for authenticity but a desire to measure the way in which the reader relates the text to their prior knowledge. It follows that there are several possible correct answers to these questions. That results in a set of quality test questions. I do not know if they were submitted to French teachers. There were two groups of Swiss teachers, initially rather hostile to PISA, but who nevertheless were impressed, based on what they told me, by the quality of the items presented.

There has also been progress in the translation of the tests into the different languages (Grisay, 20012). Whereas preceding assessments essentially used back-translation (translation to German from English is re-translated into English and compared to the original text), PISA used double translation from two source languages, English and French. For example, in Germany, one translator translates from English, another from French and an umpire compares the two translations and the final version is made through agreement of the three people. This is done because back-translation does not identify errors due to a too literal translation of the source version and to guarantee the equivalence of the final versions (between the translation in German and that in Japanese for example) by equivalence with two source versions and not one, which improves the semantic equivalence and "reduces the impact of cultural factors introduced by the reference language when there is only one.". In fact, only six countries proceeded in this, costly, way recommended by the PISA Consortium, which meant it was possible to observe that the recommended solution was indeed more reliable than double translation from a single language (Grisay, 2002, p 52). However, despite the care applied to this feature, it remains the case that in practically all countries some or other imperfection has made the scores a little higher or lower on some items. These errors, which generally cancel out across the totality of the items, are not such as to significantly change the average country rankings.

(3) French pupils' performance in PISA

It is possible to explore France's position on particular aspects of the test and on particular sub-scales, what the DEP call a "pedagogical" reading which is concerned with the direction and size of the differences between French pupils' results and the OECD average. We then can see, for example, that:

---

[3] Progress in International Reading literacy study, an evaluation of fourth grade pupils' reading literacy, undertaken by IEA in 35 countries. For a presentation on the results for France see MEN-DEP, Note d'Information 03.22

- "while young French people prove to be able readers when it is a matter of picking out information or interpretation, they seem to experience difficulties when they have to express their point of view or offer a judgment" (Robin, 2002).

- French pupils are less used to being presented in class with non-literary texts and do relatively worse on these than on exercises based on literary texts (MEN-DEP, 2002)

- French pupils have one of the highest rates of non-response amongst OECD countries for questions calling for a constructed response (MEN-DEP, 2002).

It is also possible, as I have done (Meuret, 2003), to take the risk of comparing individual countries, but using all the items and trying to find explanations for the difference in average performance between young French people and those in countries which do better than them in the three subjects tested (reading literacy, maths and science): two Asian countries (Japan and Korea), two Scandinavian countries (Finland and Sweden) and four Anglo-Saxon countries (Australia, Canada, New Zealand and the UK).

I first observed that this superiority remains after controlling for family characteristics related to academic success (wealth, parental education, whether two-parent family, speaking the language of instruction at home etc). I then compared France and these countries from the point of view of four major dimensions which empirical studies have shown were linked to effective teaching[4]. The *organisation of the education system* seems more favourable in New Zealand and the UK than in France but this is not the case for the other six countries[5]. For three of the four variables (discussion of cultural matters, homework, educational resources in the home) which represent *support and stimulation by the family* the position is significantly more favourable in France than in these eight countries and, for the fourth (*parents offer help*) it is more favourable than in three countries, less so in the other five. *Disciplinary climate* and *pupil engagement* are better in Korea and Japan than in France but less good in the remaining countries. These therefore are not the factors that can explain the poorer performance of French pupils, with the exception of help from parents with their children's work and, for the Asian countries, pupil discipline. There remains the fourth dimension, *effectiveness of class teaching* measured by three constructed variables in PISA (academic press, quality of pupil-teacher relations, support given to pupils) which are also linked, all other things equal, to pupil progress in French and maths in French lower-secondary schools (Grisay, 1997). France is, on these three variables, in a significantly less good position

[4] For those I drew on classic research studies on teaching effectiveness and not on the relationships that can be seen in PISA itself between process variables and attainment because, without a value added measure of attainment in PISA, these relationships are affected by omitted factors which result in lack of correlation or very low correlations (OECD, 2003) between variables where the literature often shows there to be stronger relationships.

[5] The description of the organisation of education systems is in OECD "Education at a Glance, 1998". The subsequent three dimensions were measured in the PISA "pupil" questionnaire.

than the other countries, apart from Japan and, for one variable only, Korea[6]. It therefore seems clear that the inferior performance of French pupils compared with the better-performing Scandinavian and Anglo-Saxon countries can partly be explained by the attitudes of French teachers compared to those of teachers in these countries.

This study does not show that French teachers are less "demanding and attentive to pupils in difficulty" than their colleagues in these countries, it shows that there are fewer of them (the size of the difference is: 75% on the one side, 55% on the other), which is not the same. Furthermore, it does not show that here is the sole explanation of the lower attainment measured by PISA: the existence of year repeating, teaching that is less oriented to the acquisition of some of the competences measured by PISA[7], other factors clearly, including poorer parental support, probably figure amongst the explanation. The study only allows us to identify one negative factor and hence an area where progress is possible.

I would like here to generalise from this study by emphasising that an education where teachers are less likely to be responsive to pupils having difficulties, seems to me to bring us back to various key characteristics in the DEP's analysis (lower motivation, higher rate of non-response to questions calling for extended answers[8], difficulties in presenting an argument, in applying a critical spirit) as well as to certain findings from PIRLS (the fact that French pupils under-estimate their abilities more than in other countries). In broad terms, compared with these countries, French pupils who have less frequent exposure to teachers who care about helping them in their learning, less frequent exposure to teachers who care about getting them to appreciate the intrinsic

---

[6] The differences measured here can be large: 90% of UK pupils as against 45% of French said that "the teacher expects pupils to work hard", 74% of UK pupils compared to 57% of French answered that "the teacher carries on explaining until the pupils have understood" (PISA database on the open-access PISA web-site). These results are additionally in line with the fact that UK teachers are, more often than their French equivalents, confident of their teaching effectiveness, according to a recent comparison of teaching in England, Luxembourg and France at the end of primary (RERPESE, 2002).

[7] The same study notes that English primary teachers give more weight than French to teaching the stating of an opinion and argument. It might be supposed that that leaves its mark at age 15 and helps explain why French youngsters do less well particularly when they have to show a critical spirit.

[8] Across all items, the rate of non-response (missing items) in France is close to the average rate of non-response in PISA. It is higher than the rate of non-response seen in the eight countries which outscore us. However, analysis of the phenomenon of omitted questions shows a very high (negative) correlation between the rate of omission or items not reached and the score the pupil would have got if only the items for which the pupil gave an answer are taken into account. That therefore is a feature rather than a cause of our poorer performance relative to those countries. However, there are a few more items missing in France than in countries whose performance are not significantly different from ours (Austria, Belgium, Iceland, Norway, USA, Denmark, Spain): 2.44 in France compared to 2.1 all country average for session 1, 2.99 compared to 2.79 for session 2 (OECD, 2002, p. 157). Though "fear of making a mistake" — which might well have been seen as a factor here — is perhaps a little stronger in France than elsewhere, but not much.

value of skills, less frequent exposure to exercises drawing on the outside world, and hence one might think, subject to lessons from a higher standpoint — since that is our reputation — but only mastering the easy parts; French pupils acquire rather less competence and more academic skills than in these countries, more often feel incapable of offering their opinion, and will be less inclined to "work for nothing". This picture makes sense, which leads me to consider, for example, the percentage of non-responses to complex items or the lower willingness to apply a critical appraisal, not as sources or outcomes of bias, tending as a result to invalidate the overall scores which I used but as notable results and which are, instead, consistent with the results of my own work.

(4) Why have I not taken into consideration that, if France got only average results in PISA, this was because of "cultural biases"?

First I would like to emphasise that "cultural biases" are one out of a whole group of biases (sampling, definition of the parent population[9],…) which can affect an international assessment and that you can probably examine what role they play in our country's reception of PISA. But, since this is why we are here, let us think about cultural biases.

If I decided to treat these as negligible in my work, it is first, but I won't go over this again**,** because of the care taken in PISA to eliminate items which might have brought this about. This is also because it was possible to show that these biases were minimal. The measures of differential functioning of items — i.e. the difference between the average score of a country on an item and that which would have been expected taking account of the average score in the domain — enables us to say that the differences and difficulty of a given item from one country to another are minor and in particular, that if certain items slightly favour one country, others have the opposite effect so that the overall score — which is what matters to me — is not affected.

It seems to me that we must add to this that, if an item is better done in one country than it should be, taking account of the overall score of that country, that can be because of a cultural bias in the item (a greater familiarity of pupils in the country with the content of the text for example) but also that it measures competencies that are particularly developed in this country, in which case it is not a matter of a bias, but an interesting result for analysing the actual curriculum of the country. The same applies for the test materials used: if, in science, academic texts are used more than press articles to present scientific issues (MEN-DEP, 2002, p 159), we can criticise the PISA items for not adopting French educational practices as much as we can question the consequences of this practice in terms of the capacity to handle scientific facts. Were we to judge, in the end, this to be an acceptable cost in relation to the advantages in French practice, the questioning would not have been without use. Similarly, the practice, in the maths questions, of asking problems of increasing difficulty, and put forward as an explanation, indeed a possible one, of certain non-responses (MEN-DEP, 2002, p 146): were we — though I doubt it — to conclude from trying this practice, … is probably also a good example of how international comparisons can be used. It is however difficult to mark a clear line between a psychometric bias and an interesting

[9] This bias shifted France's results upwards in the 1991 Reading Literacy Study.

result; where each of us put this depends on how far the French situation must serve as the norm[10].

There remains the question of the length of the text (Grisay, 2002). The chapter on translation in the PISA technical report does indeed show that texts and words are slightly longer in French than in English. It therefore seems that their language does slightly handicap young French speakers[11], which nevertheless does not prevent young Quebecers from achieving enviable scores[12] and which therefore might be a (small) part of the explanation of our score, especially in reading literacy. It is therefore legitimate to reckon that the very characteristics of their language are a slight handicap for the French relative to certain countries which do better than them in PISA 2000 (English-speaking countries, Finland). However, we must ask ourselves along with Grisay (2002) what response this situation requires: must we give slightly shortened texts to pupils in less concise languages or refuse to do this, judging that this would be just as biasing as comparing pupils in different languages on texts whose semantic weight was different?.

The ideal solution would probably be to include the conciseness of language as one of the factors influencing the effectiveness of teaching reading literacy. It must be stressed, moreover, that an international study which consisted of a pre-test and a post-test and which measured the effectiveness of an education system over a given year by "value added" would avoid this difficulty, as it would many others which are raised when the issue is of relating the characteristics of an education system and the performance of its pupils.

Denis Meuret
University of Burgundy (Université de Bourgogne)
IREDU

---

[10] The present author was placed in a similar position when he tried hard, with others, to promote the use of uniform funding in schools: each school explained that its specific characteristics required that it be treated differently from the rest, and criticised the crudeness and insensitivity of the indicator (H/E), and indeed it is only after having thought for a long time that indeed this indicator was unacceptably crude, that I thought that its merit rested in its very crudeness which brought out the cost of these famous "specificities". In the meantime, we had tried to "weight" in various ways, all the same feeling confusedly that by pushing the weightings, we put at risk the essence of what we were doing.

[11] The coefficient of elongation of the texts is 15% compared to English. Similarly, German, Italian, Spanish, various Scandinavian languages, have significant elongation coefficients. One of the least concise is Korean (30%). That the difficulty of a understanding a text does not only depend on its length explains why the effects of coefficients as high as this can be low and that Korea can have one of the best scores in reading … and that the correlation between performance in the three PISA subjects is high, whereas this language factor affects them differently.

[12] 536 in reading literacy, only 10 points less than Finland and 31 points above France, or 0.3 of a standard deviation: the median Quebec score is higher than 62% of the French. Statistics Canada, 2001 *At the top: the performance of young Canadians in reading, maths and science*, OECD PISA study, 93 p.

Bibliography

Adams, R.J., 2003, *Response to "Cautions on OECD's Recent Educational survey (PISA)",* Oxford Review of Education, 29(3), 377-389.

D'Haultefeuille,X., Murat,F. et Rocher, T., 2000, *La mesure des compétences, les logiques contradictoires des évaluations internationales*, Intervention aux VIIème journées de méthodologie statistique, Paris, 37p.

Grisay, 2002, *Translation and cultural appropriateness of the test and survey material,* PISA technical report, 42-54.

Grisay, A. 1997 *Evolution des acquis cognitifs et socio-affectifs des élèves au cours des années de collège*, MEN-Direction de l'Evaluation et de la Prospective , Dossiers Education et formations, n°88.

Lafontaine, *Quoi de neuf sur la littéracie ? Regard sur trente ans d'évaluation de la lecture*, Cahiers du SPE, Université de Liège, n°7-8, 71-95.

Meuret, D. 2003 *Pourquoi les jeunes français ont-ils à 15 ans des performances inférieures à celles des jeunes d'autres pays ?* Revue française de Pédagogie, n°142, 89-104.

Meuret, D., 1998, *Intérêt, justice, laïcité*, 8p, in (Le) Télémaque, Presses universitaires de Caen, N° 14.

Prais,S. J. , 2003*, Cautions on OECD'S Recent Educational Survey (PISA)* Oxford Review of Education, 29(2), 139-163

Robin, I., 2002, *L'enquête PISA sur les compétences en lecture des élèves de 15 ans_: trois biais culturels en question*, Ville, Ecole, Intégration, Enjeux, n° 129, CNDP.

Men-DEP, 2001, *Les élèves de 15 ans, premiers résultats d'une évaluation internationale des acquis des élèves,* NI 01-52, 6p.

MEN-DEP, 2002, *Les élèves deCM1, premiers résultats d'une évaluation internationale en lecture (PIRLS),* NI 01-52, 6p.

MEN-DEP, 2002, *Les compétences des élèves français à l'épreuve d'une évaluation internationale, Premiers résultats de l'enquête PISA 2000*, Dossier Education et formation 137, Paris, 182p.

OCDE, 2001, *Choice of assessment tasks and the relative standing of countries in PISA 2000, a first analysis*. Document pour la treizième réunion du conseil des pays participants, Paris, 10 p.

OCDE, 2002, *Pisa Technical Report*, Paris, 262p.

OCDE, 2003, *La lecture, moteur de changement, Performances et engagement d'un pays à l'autre, résultats de PISA 2000*, Paris, 279p.

Réseau européen des responsables des politiques d'évaluation des systèmes éducatifs, (RERPESE) 2002, *Enseignement et Compétences en lecture à la fin de la scolarité primaire dans trois pays européens*, MEN-DEP, 105p.