

# Ensuring marker reliability in international assessments

Jenny Bradshaw

National Foundation for Educational Research

IAEA Conference 2008 Cambridge, UK

**7 – 12 September 2008**

## ***Introduction***

In any marking which depends on judgment of responses written by students, careful training and moderation of marking is essential to ensure that markers apply judgments consistently and reliably. In the case of a large international survey of educational achievement, such as the OECD PISA survey or the PIRLS and TIMSS studies organised by IEA, there are particular challenges. This paper discusses the issue in the context of the OECD PISA survey, using one question from the 2006 survey as an example.

The OECD Programme in International Student Assessment (PISA) assesses the attainment of 15-year-olds in reading literacy, mathematical literacy and scientific literacy, and also gathers a large amount of data on student and school factors and on student attitudes. In each PISA survey one of the three subject areas forms the main focus, which in 2006 was on science. A total of 57 countries participated in PISA 2006.<sup>1</sup>

The test questions and mark schemes were developed by an international consortium led by the Australian Council for Educational Research (ACER). In the United Kingdom, the National Foundation for Educational Research (NFER) was responsible for administering the survey in England, Wales and Northern Ireland. Scotland participated separately.

The assessment items used in the PISA survey are a mixture of closed and open items. Some of the open items require only short responses which are relatively easy to mark, but others require longer constructed responses from students, and these are the items where ensuring reliable marking poses a particular challenge. Such items will always need care in training and moderation of markers, but when the additional requirements of ensuring reliability across a range of languages, situations, marker expectations and local conditions are added, the challenge for the PISA consortium is great.

This paper uses the example of one item in the assessment to illustrate the process and to draw some general conclusions.

### ***An example of the PISA marking process***

On the following page is the stimulus text for one of the items used in the PISA 2006 assessment.<sup>2</sup> This item was one which had been used in previous PISA surveys, and in common with many of the earlier items had a relatively large reading load. Science items which were newly-developed for PISA 2006 tended to have a reduced amount of text. The item which is included here has now been released and will not be used in future PISA surveys.

---

<sup>1</sup> For full details of PISA go to <http://oecd.pisa.org>

<sup>2</sup> The full unit and other released items are available at [https://mypisa.acer.edu.au/index.php?option=com\\_content&task=view&id=69&Itemid=445](https://mypisa.acer.edu.au/index.php?option=com_content&task=view&id=69&Itemid=445)

## Figure 1 GREENHOUSE

Read the texts and answer the questions that follow.

### THE GREENHOUSE EFFECT: FACT OR FICTION?

Living things need energy to survive. The energy that sustains life on the Earth comes from the Sun, which radiates energy into space because it is so hot. A tiny proportion of this energy reaches the Earth.

The Earth's atmosphere acts like a protective blanket over the surface of our planet, preventing the variations in temperature that would exist in an airless world.

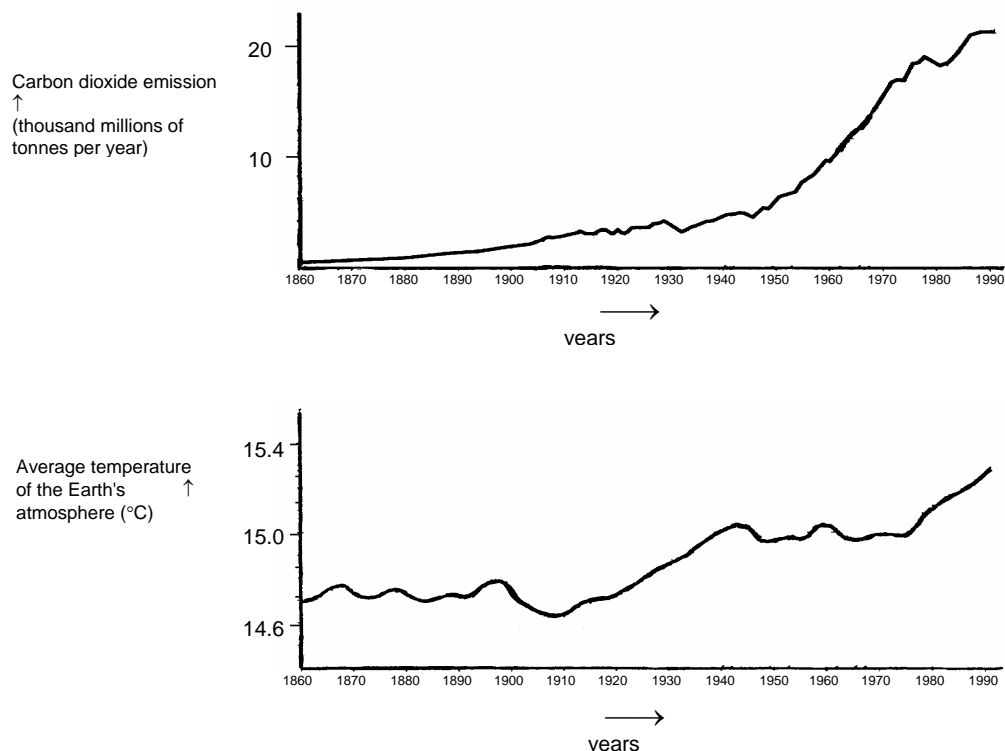
Most of the radiated energy coming from the Sun passes through the Earth's atmosphere. The Earth absorbs some of this energy, and some is reflected back from the Earth's surface. Part of this reflected energy is absorbed by the atmosphere.

As a result of this the average temperature above the Earth's surface is higher than it would be if there were no atmosphere. The Earth's atmosphere has the same effect as a greenhouse, hence the term *greenhouse effect*.

The greenhouse effect is said to have become more pronounced during the twentieth century.

It is a fact that the average temperature of the Earth's atmosphere has increased. In newspapers and periodicals the increased carbon dioxide emission is often stated as the main source of the temperature rise in the twentieth century.

A student named André becomes interested in the possible relationship between the average temperature of the Earth's atmosphere and the carbon dioxide emission on the Earth. In a library he comes across the following two graphs.



André concludes from these two graphs that it is certain that the increase in the average temperature of the Earth's atmosphere is due to the increase in the carbon dioxide emission.

There were three constructed response questions based on this text. The first of these was:

**Figure 2**

What is it about the graphs that supports André's conclusion?

.....

.....

In the coding guide (mark scheme), there were two possible types of full credit response:

**Figure 3**

**Full Credit**

Refers to the increase of both (average) temperature and carbon dioxide emission.

- As the emissions increased the temperature increased.
- Both graphs are increasing.
- Because in 1910 both the graphs began to increase.
- Temperature is rising as CO<sub>2</sub> is emitted.
- The information lines on the graphs rise together.
- Everything is increasing.
- The more CO<sub>2</sub> emission, the higher the temperature.

Refers (in general terms) to a positive relationship between temperature and carbon dioxide emission.

*[Note: This code is intended to capture students' use of terminology such as 'positive relationship', 'similar shape' or 'directly proportional'; although the following sample response is not strictly correct, it shows sufficient understanding to be given credit here.]*

- The amount of CO<sub>2</sub> and average temperature of the Earth is directly proportional.
- They have a similar shape indicating a relationship.

This is an example of a question in which responses need careful interpretation to distinguish between a full credit answer and one which is given no credit. This is particularly the case with the second type of ‘full credit’ response, which may be more general and which is acknowledged in the coding guide to be ‘not strictly correct’ (see figure 3).

The coding guide also gave detailed guidance on answers which, while they may have been partially correct, were not given full credit:

**Figure 4**

<p><b>No Credit</b></p> <p>Refers to the increase of either the (average) temperature or the carbon dioxide emission.</p> <ul style="list-style-type: none"> <li>• The temperature has gone up.</li> <li>• CO<sub>2</sub> is increasing.</li> <li>• It shows the dramatic change in the temperatures.</li> </ul> <p>Refers to temperature and carbon dioxide emission without being clear about the nature of the relationship.</p> <ul style="list-style-type: none"> <li>• The carbon dioxide emission (graph 1) has an effect on the earth’s rising temperature (graph 2).</li> <li>• The carbon dioxide is the main cause of the increase in the earth’s temperature.</li> </ul> <p>OR</p> <p>Other responses.</p> <ul style="list-style-type: none"> <li>• The carbon dioxide emission is greatly rising more than the average Earth’s temperature. <i>[Note: This answer is incorrect because the <u>extent</u> to which the CO<sub>2</sub> emission and the temperature are rising is seen as the answer, rather than that they are both increasing.]</i></li> <li>• The rise of CO<sub>2</sub> over the years is due to the rise of the temperature of the Earth’s atmosphere.</li> <li>• The way the graph goes up.</li> <li>• There is a rise.</li> </ul>
---

Questions such as this, in which some partially correct answers are given full credit while others are not, presented some challenges in the marking process. In some other questions, a partially correct response would be given partial credit (a score of 1) and a fully correct response was given full credit (a score of two). During the marker training process it was necessary for the markers to set aside any feelings they may have had that it was unfair to the students not to give them some credit for partial correctness – or, conversely, that only fully correct answers should be given credit. They also needed to pay careful attention to the way in which students phrased their answers in order to decide which side of the credit/no credit borderline a response should be placed.

## ***Marker training***

The training of markers was done first on an international level. Marking supervisors were taken through the process which they themselves would need to carry out with the national markers. The first stage of this training was familiarity with the mark scheme, followed by practice coding of some workshop examples taken from responses by students in the trialling done by the test developers.

The answers below are some of those used in the international training workshop:

### **Figure 5**

#### ***Full credit***

- The temp went up, with the amount of CO<sub>2</sub>.
- That they stay together and as the years go by the earth atmosphere goes up.
- While the carbon dioxide is constant the temperature is. But if the level of carbon dioxide increases the temperature increases too.

#### ***No credit***

- With more carbon dioxide, the temperature changes.
- The average temperature of the atmosphere is related to the increase of the carbon dioxide.
- The level of carbon dioxide is rising over the years.

This was followed by discussion of the codes given and the reasons for the codes on which the test developers had decided. In some cases, particularly in a question like this where careful interpretation of the students' meaning was required, this discussion became lively.

The international training gave marking supervisors an overview and helped with interpretation of the marking guide. The workshop examples which were used in the international training were used in the first stage of training for local markers. However, it was essential that examples of national student responses should also be used to train the local markers. In the case of countries which translated the questions into another language, especially if the mark scheme and examples were also translated, the ways in which students typically worded their answers were crucial. Even in countries which used the English versions, there may have been local differences in the way in which answers were expressed.

Below are some of the local examples used in the marker training in the UK (England, Wales and Northern Ireland).

**Figure 6**

***Full credit***

- as the temperatures go up the carbon dioxide does aswell.
- The both roughly increase and decrease at the same point
- They both steadily increase over time

***No credit***

- similarity
- They show that the increase in the average temp of the atmosphere is due to the increase in the carbon dioxide emission
- There's a big increase in the amount of carbon dioxide been released over the years

The local examples were used, firstly, as a method of checking marker reliability after the international workshop examples had been discussed. The discussion of coding of local examples also served as a check for the supervisor on the markers' interpretation of the mark scheme. Individual markers who had problems could then be re-trained or could be supervised closely while marking the item, or the whole group could be re-trained if necessary.

During the marking process, regular spot-checks were done. The marking was centre-based with markers working full-time under supervision, which enabled immediate discussion or re-training where necessary. It was also possible to send queries to the International Consortium on responses which were difficult to interpret, which did not quite fit the mark scheme or which caused debate and disagreement among markers. Despite the many examples in the coding guide, the international workshop and the local examples, there were nevertheless occasions when a student would give an original response which did not totally fit the available examples.

As well as local checks on marker reliability there were international checks. A proportion of scripts were marked by four separate markers and this data was analysed internationally. In addition, a number of these multiply-marked scripts were also marked by an international team of markers. National teams were sent the results of these checks both for individual markers and for individual items, as a report on the reliability of the national marking.

## *Discussion*

The marking of this type of question inevitably has an element of subjectivity. However, as the process described in the previous section demonstrates, subjectivity needs to be kept to a minimum in an international marking process. Markers need to be trained to justify their choices about whether to give credit in terms of comparison with an example in the mark scheme, the international workshop materials or the local examples. They need to be able to look beyond the student's possibly faulty expression to interpret the student's meaning, but to do this without over-interpreting what they think the student meant to write.

The amount of attention given to ensuring the quality of marking is much greater in an international survey than in a marking operation which involves a test and a mark scheme produced for just one country. In the latter case the quality of marking is of course equally important, but both test trialling and development of the mark scheme will have been done with students who are from the same background as those to whom the test is administered. In an international survey, the same assumptions cannot be made and the responses given by students in one's own country will not necessarily fit the examples which have been based on trials in another country.

One effect is that marking takes longer than with a locally-produced test as there tends to be more discussion and more queries. The training process needs to be given sufficient time for all to become thoroughly familiar with the mark scheme and to understand the reasons for giving credit to particular examples of student responses. The marking supervisors are limited in the extent to which they can arbitrate on problematic responses as they need to consult the international team, to ensure that local interpretations are acceptable. This sometimes leads to delays in marking while waiting for a response, because of international time differences.

In summary, great care is needed in all aspects of international studies to ensure comparability of results. The marking of tests is one aspect which could be particularly susceptible to local variation if care is not taken. The national centres which are responsible for carrying out the marking in their own country share this responsibility with the international organisers. While the example discussed in this paper is specific to the PISA study, any test with constructed response items which is used in more than one language or which is used in more than one context would need an equally rigorous approach to marking.