NATIONAL FOUNDATION FOR EDUCATIONAL RESEARCH

Ensuring marker reliability in international surveys

Claire Hodgson

The NFER is involved in a number of international comparative surveys. These focus on attainment in a variety of subjects and ages including reading in primary schools (PIRLS), maths and science in primary and secondary schools (TIMSS) and the skills in the core subjects of students nearing the end of compulsory schooling (PISA). With many different countries involved, it is crucial that rigorous methods are in place to ensure that marking is comparable both across countries and over time.

Background

The Progress in International Reading Literacy Study (PIRLS) survey is conducted by the International Association for the Evaluation of Educational Achievement (IEA). It is a comparative study of the reading achievement of 10-year-olds which is carried out every five years in order to measure trends. It was first carried out in 2001 and involved more than 140,000 pupils in 35 countries. The second survey was carried out between October 2005 and June 2006 and involved 40 countries, including England and Scotland. The survey took place between October and December 2005 in the southern hemisphere and between March and June 2006 in the northern hemisphere to take account of the school year in these different parts of the world. The results of the survey are due to be published in late 2007.

As well as the actual reading tests themselves, the survey includes a series of questionnaires for pupils, teachers, headteachers and parents. There are ten different reading tests, which assess comprehension of both literary and informational texts, of which pupils are required to complete two. The questions included a mixture of closed response formats (for example, multiple choice and ordering) and open response formats, requiring pupils to construct their own answers.

NFER involvement

NFER were contracted by the Department for Education and Skills (DfES) and the Scottish Executive Education Department (SEED) to carry out the surveys in England and Scotland. Samples of pupils in year 5 (England) and in primary 5 (Scotland) were selected to take part in the survey. Administrators, appointed and trained by the NFER project team, were responsible for visiting the sampled schools and giving the nominated pupils an allocated reading test. The tests were then returned to NFER for marking.

Marking training

Training for marking took place by means of a cascade system. The IEA devised the marking guide in consultation with reading experts drawn from a number of countries involved in the study. The marking guide was finalised and agreed during the first training session which was held in the southern hemisphere. No changes were allowed to be made following this finalisation. Closed response questions are scored automatically and training for marking open response questions was provided at intensive week-long meetings, one in the southern hemisphere and one in the northern hemisphere. They were attended by representatives from each country and thorough training on the agreed marking guide was provided. This involved completing Anchor papers (on which scores were already agreed) and Practice papers (in which trainees had to allocate and discuss marks before being informed of the agreed mark by IEA trainers).

Following this tutoring, the NFER researchers trained seven highly experienced reading test markers who had been appointed by NFER to mark the test papers for England and Scotland. The training followed a similar pattern to the international model, with the marking guide being discussed and Anchor papers and Practice papers being completed. Training took place over a period of two weeks, allowing time for consolidation marking and the development of pupil response exemplar banks. Following this period, markers were allocated batches of tests to mark – a process which involved marking a total of around 8,000 tests over a period of six weeks. All the marking was centre-based. Completed tests were checked either by a researcher or by a fully trained nominated marking checker.

Ensuring reliability

In order to ensure reliability of marking across countries, every country also had to mark a specific set of items in English, drawn from the field survey (a trial carried out before the main survey). To enable this to happen, the IEA provided a CD containing a total of 4,800 pupil responses to open response questions, drawn from four of the tests, to be marked. This included scripts from England, Scotland, Canadian province of Ontario, United States, New Zealand, South Africa and Singapore. At least two markers were required to mark each item on computer with these scores being logged on a database. This data was then sent to the IEA for analysis. For countries where the tests are not conducted in English, it was necessary to recruit markers who *can* speak English in order to complete the cross-country reliability study.

Similarly, in order to ensure reliability of marking over time, every country which had been involved in the 2001 survey also had to complete trend reliability items. These were items which appeared in both the 2001 and the 2006 surveys.

A total of 5,200 pupil responses (200 responses to 26 different questions) were provided – these were taken from scripts relating to that specific country from the 2001 survey - and at least two markers were required to mark these on-screen. Once half of these responses had been marked, the results were exported to a spreadsheet and the agreement between the two 2006 markers as well as the agreement between each of the 2006 markers with the mark assigned in 2001 were analysed. Where agreement for any comparison was below 85%, retraining was required. Once this had been provided, the marker re-marked the first half of the responses and a new agreement was calculated. If the agreement was above 85 per cent, then markers were able to continue marking the remaining half. Final results of all the marking on this CD was submitted to the IEA and will be reported in the international report.

Conclusion

Ensuring reliable marking with open response questions is always going to be challenging, especially in a subject such as reading. Many measures are in place for the international surveys to ensure that marking is done as consistently as possible and that it is as reliable as it can be. Marking in different countries happens in a variety of ways, for example, in terms of the previous experience and numbers of markers involved and it is not possible for the IEA to directly monitor each set of marking. However, by rigorously following the systems laid down, it is possible to have some confidence that the results for these international surveys are based on careful and accurate marking which is of comparable standards both between the different countries involved and over time.