Impact of independent judges in comparability studies conducted by awarding bodies

AUTHOR: Mike Forster
ADDRESS: Research and Evaluation Division,
University of Cambridge Local Examinations Syndicate,
1 Hills Road,
Cambridge.
CB1 2EU.
☎ 01223 553851
FAX: 01223 552700
💻 forster.m@ucles.org.uk

AUTHOR: Elizabeth Gray
ADDRESS: Assessment Quality Department,
OCR,
1 Hills Road,
Cambridge.
CB1 2EU
☎ 01223 552789
FAX: 01223 327904
💻 gray.e@ucles.org.uk

Abstract

The Unitary Awarding Bodies (formerly Examination Boards and Groups) in England, plus CCEA in Northern Ireland and WJEC in Wales, are responsible for providing public examinations for school students. The Awarding Bodies regularly conduct comparability studies as part of their procedures for maintaining standards. These usually involve experienced examiners from each of the boards making judgements on a range of scripts from each board. This paper will consider the impact of using independent judges in these studies. It will look at the nature of previous comparability studies, which more recently have tended to use board-affiliated judges, and the methodologies that were used. Recent studies, performed on the 1998 GCSE English, Mathematics and Science examinations, used the Thurstone paired comparisons technique, a methodology relatively new to comparability studies. This technique required the judges to compare pairs of scripts, and record which script they felt was of the higher standard. It did not require an explicit knowledge of grading standards, as some of the previous studies had. This enabled independent (i.e. not affiliated to any of the boards in the study) judges to be included in the study. Evidence from these three studies will be used to demonstrate that the independent judges produced very similar results to the board-affiliated judges, reinforcing the validity of these results.

Impact of independent judges in comparability studies conducted by awarding bodies

Introduction

The examination boards[1] of England, Wales and Northern Ireland are responsible for providing public examinations for school students. Each board offers examinations in many different subjects, and so different boards have offered examinations in the same subjects for many years. The debate about the equivalence, or comparability, of these examinations is also long-standing. Mathews (1985:139) noted that there was:

> *…a feeling that some boards are more generous or more severe than others either in general or in particular subjects.*

Davies (2000:7) noted the words of a retired secondary head teacher:

> *If you are clever you can improve your GCSE results by picking the right exam board for the right subjects.*

There is, quite clearly, a perception that the boards do not always offer comparable examinations, and that schools can shop around for the 'easiest' ones. However, for many years there have been procedures in place to monitor the comparability of examinations in the same subject offered by the different boards. Straight comparison of results is not an appropriate methodology, as each board has its own candidate entry, which differs from every other board's entry. As Bardell et al (1978:15) noted:

> *There is, for example, a common pattern which shows that candidates from independent schools tend to achieve higher grades than those from maintained schools. It is therefore to be expected that a board such as O&C, which draws most of its entry from the independent sector, will award a greater percentage of high grades than a board which draws its candidates principally from maintained schools.*

Studies have been undertaken using grade distributions from different boards, but with statistical adjustments made for the different candidate entries. Indeed, ten such studies were undertaken between 1966 and 1975, across the range of subjects (Bardell et al, 1978).

Comparing grade boundary marks is also inappropriate, since the boundary marks themselves cannot be separated from the difficulty of the paper. A paper with high boundary marks could be 'easier' than one with low boundary marks, depending on the relative difficulty of the papers. An alternative method that has been used is the monitor test, such as Aptitude Test 100 (Nuttall et al, 1974), also known as a reference test (Johnson and Cohen, 1983). A monitor test is a test which is separate from the two syllabuses being compared, which acts as a common test against which the results in the two syllabuses can be compared. It tends to be either a test in the subject under investigation, or a general aptitude test (Mathews, 1985). Such tests must not be biased towards the candidates taking one syllabus, and must be equally relevant to candidates taking both syllabuses (e.g. a mathematics test would not be a good monitor for comparing two French syllabuses). Ten studies between 1968 and 1976 used monitor tests to compare different boards' syllabuses in Physics; Economics; French; Biology and History (Bardell et al, 1978).

The monitor test, however, is based on statistical comparisons, and so a preferred methodology is that of the cross-moderation study. This involves judgement of script quality by the most experienced examiners, the same process that is undertaken when grades are awarded. Scripts from different boards, at the same grade boundary, are compared, and any differences in quality noted. This has proved the most popular technique

[1] For the sake of clarity, throughout this report all of the examining bodies will be referred to as examination boards. The individual boards will be referred to by their pre-Unitary Awarding Body names throughout. See Appendix 1 for more detail.

with the examination boards, with 55 such studies taking place between 1964 and 1997. Indeed, Bardell et al (1978) noted that:

> *…cross-moderation involving the boards' examiners (possibly with outsiders too) is the most fruitful and sensitive of the methods available.*

History of cross-moderation studies

Cross-moderation studies are based on the assumption that the most experienced examiners can scrutinise scripts from different boards, and decide whether the same grades have been awarded to candidates of comparable levels of attainment. There have been two main types of study: identification studies and ratification studies (Forrest and Shoesmith, 1985). Identification studies require examiners to identify where in a range of scripts the boundary mark should lie. Ratification studies require examiners to look at boundary mark scripts and either agree, or disagree, with the board's decision that the scripts are worthy of the grade. Between 1964 and 1983 there were 19 ratification studies, 9 studies classed by Forrest and Shoesmith (1985) as 'other' studies, and only 5 identification studies. Since 1983 there have been a further 22 studies: 18 ratification studies; 2 identification studies; 1 Thurstone Pairs analysis; and 1 'other' study.

These cross-moderation studies traditionally involved senior examiners from the examining boards. However, since the late 1970s there has been a tendency to include subject specialists, external to the examining process, in the studies. One reason for this is to provide a more objective element to the studies. Forrest and Shoesmith (1985) noted the outcome of the O level History cross-moderation study in 1978, whereby each of the examiners thought that the other boards' syllabuses were generous. This is an example of how bias can affect the results. The addition of independent specialists should provide an unbiased monitor against which the board examiners can be judged. If they are drawing conclusions in a biased way, their results should differ from those of the independent judges.

There are, however, problems associated with incorporating independent judges in the cross-moderation exercise. Firstly, since they are considerably less familiar with the syllabuses than the board examiners, they are, perhaps, less able to make judgements about scripts from different boards. They will also, inevitably, make their judgements at a slower pace than the more experienced board judges. This can be an important issue. The cross-moderation process is a somewhat tiring and tedious process, and hence after more than two days the judging process becomes unproductive. This means there is a limited amount of time available for the judges to make the decisions which provide the data for the study. Independent judges might provide less data in the time permitted, and this may affect the reliability of the results.

Case study – the 1998 cross-moderation studies in English, Mathematics and Science

Background

In 1998, the examination boards began a series of cross-moderation studies as part of an ongoing programme, under the guidance of the Joint Council for General Qualifications (JCGQ). The JCGQ is made up of representatives from the five awarding bodies in England, Wales and Northern Ireland that offer A level, GCSE and GNVQ. Through the JCGQ the awarding bodies seek to ensure that individual awards are of a comparable standard each year, and over time. Six examination boards took part: CCEA; Edexcel; MEG; NEAB; SEG and WJEC. Studies were undertaken in English, Mathematics and Science. In 1999, a study began in a fourth subject, French, but it will not form part of this report. The methodology used in the study was different from that which had been used in many previous studies. The first part of the study involved a syllabus review. Examiners compared syllabuses, question papers and mark schemes from each board, so that a number of similarities and differences emerged. The second part of the study was the cross-moderation exercise, reviewing candidates' work. Rather than using the traditional ratification method, a technique known as Thurstone Pairs was used. Gray (1999) reports that this technique had been used successfully in the 1996 A level Biology and Mathematics studies. It followed the methodology first proposed by Thurstone in 1927 (van der Linden, 1994). Examiners consider the whole work of boundary scripts from two different boards, and then have to decide which demonstrates the higher level of achievement. No ties are allowed, and scrutineers do not make judgements involving scripts from their own boards.

There are a number of advantages of using this method over some of the methods previously used. Since examiners are comparing only the quality of two pieces of work, they do not need to be able to judge whether a piece of work is on the boundary, only whether it is better or worse than another piece of work. This methodology can thus be used by any subject expert, without the need to refer to any particular standard. Another advantage is that, by forcing judges to choose one of the scripts (although this does have its own problems), there is no dilemma regarding how far equivalence has to stretch before two scripts are considered different. Finally this is a useful method because there is a well-defined technique for analysing such data based on the Rasch model. Using raw data from the comparisons, the model estimates parameters for each script, based on the number of times any script is the 'winner'. It also takes into consideration the calibre of the other scripts against which it is being judged. For example, a 'success' against a 'better' script is valued more highly than a 'success' against a 'less worthy' script (Fearnley, 2000). The model then produces an order of merit, with the 'best' scripts at the top, but it gives no indication of the calibration of the scale in terms of mark differences.

Three boundaries were considered in the study: grades A and C on the Higher tier, and grade C on the Foundation tier (except in Mathematics, which examined grade A on the Higher tier, and C on the Intermediate tier). Each board provided five scripts for each boundary. There were, potentially, 25 comparisons that could be made between each pair of boards, and there were 6 boards which could be compared in 15 different pairings. This meant there were 375 possible comparisons on each boundary, although board judges could only make 250, since they could not make judgements on their own board's scripts. It was accepted, however, that no judge would be able to make all the comparisons in the time available.

The judges were arranged into two groups of nine, each consisting of a judge from each board, and three independent judges. The two groups looked at different boundaries in turn, so that both groups had looked at all three boundaries by the end of the study. The judges chose two scripts randomly from the pile of scripts in the centre of the table. Once they had compared them, they retained one and picked up another to compare it with (watching out for duplicate comparisons), thereby reducing the amount of time spent choosing scripts. Initially, the judges had 7 minutes to make their comparisons, although this was soon reduced to 4.5 minutes. Their comparisons were based on first impressions of the two scripts being compared, as Thurstone had intended. Part of the output of the Rasch analysis shows how the judges performed relative to all the judges together. If a judge made comparisons which did not follow the pattern of the other judges, he or she would be described as *misfitting*. If his or her judgements followed the pattern of the other judges very closely, then he or she would be described as *overfitting*. In other words, a judge who made unexpected judgements would be a misfitting judge, whilst one who made very predictable judgements would be overfitting. Further information on Rasch analysis techniques can be found in Andrich (1978).

Results

The full results from the comparability studies have been discussed in the reports by Gray (1999), Pritchard et al (1999), and Fearnley (2000). This paper will concentrate on the results involving the independent judges. In particular, it will look at the number of misfits per judge at each grade; who made the biggest misfitting judgements at each grade; and which judges were overfitting/misfitting at each grade. The report will deal with each subject in turn, starting with English, examining the boundaries individually. To ensure confidentiality, the names of the judges have been replaced by a code letter and number: I1 is the first independent judge; I2 is the second; W1 is the first WJEC judge; W2 the second, and so on. To make the tables clearer, the independent judges (**I1-I6**) are in bold.

English

Table 1 shows the number of misfitting judgements made by each judge at each grade for English. For example, judge I1 made no misfitting judgements at Higher tier grade A, 2 at Higher tier grade C, and none at Foundation tier grade C. As can be seen, the independent judges tended to be quite well spread out, although there are two in the top three of the table. This suggests that, in English, the independent scrutineers were making similar numbers of misfitting judgements as the board affiliated judges were.

Table 1: English – number of misfits by judge

		Misfitting judgements			
Judge	*Affiliation*	*High A*	*High C*	*Foun C*	*Total*
I1	SQA nominated	0	2	0	2
W1	WJEC	1	2	0	3
I5	NEAB nominated	0	2	1	3
C1	CCEA	0	4	0	4
E1	Edexcel	0	3	1	4
N1	NEAB	2	1	1	4
N2	NEAB	1	1	2	4
I3	LEA nominated	1	3	0	4
M1	MEG	5	0	0	5
I4	LEA nominated	1	3	1	5
I2	IB Nominated	3	2	1	6
E2	Edexcel	2	3	2	7
W2	WJEC	0	1	6	7
I6	WJEC nominated	1	2	4	7
M2	MEG	5	1	2	8
S1	SEG	1	4	3	8
C2	CCEA	1	3	5	9
S2	SEG	4	3	3	10
Total		**28**	**40**	**32**	**100**

Table 2 shows the judges who made the biggest misfitting judgements at grade A on the Higher tier, i.e. the judgements that were most unpredictable. The table includes a figure for the standardised residual (Std Resid), and the calculated probability. The standardised residual is a measure of the difference between the judgement expected and the judgement observed: the bigger the standardised residual, the more unexpected the judgement. The calculated probability shows how 'likely' that judgement was. For example, the judgement made by judge I4 had a probability of only 0.07 (i.e. 7%). As can be seen, of the four most misfitting judgements, two were made by independent judges, suggesting that at grade A in English, independent judges are slightly more likely to make big misfitting judgements than the board judges.

Table 2: English Grade A Higher - biggest misfitting judgements

	Misfitting judgements	
Judge	*Std Resid*	*Calculated prob*
I4	3.64	0.070
I2	3.27	0.086
M1	3.23	0.087
M1	3.20	0.089

As well as particular judgements, the Rasch analysis can detail which judges were making predictable, or unpredictable, judgements. Note that the number and size of the (un)predictable judgements made by the judge will determine whether or not a judge is classed as mis/over-fitting. Table 3 shows which judges were deemed to be misfitting or overfitting at grade A in English. Also included in the table is the mean square, which shows the degree of overfit/misfit of the judge. The only overfitting judge was a board judge, whilst two of the three misfitting judges were independent judges. This suggests that independent judges are more likely to be 'unpredictable' than the board judges, at grade A on the Higher tier in English.

Table 3: English Grade A Higher - misfitting/overfitting judges

	Misfitting/overfitting judges	
Judge	*Mean sq*	*Mis/Over –fit*
C1	0.790	overfit
I2	1.251	misfit
I1	1.291	misfit
S2	1.407	misfit

The same tables were produced for grade C on the Higher tier in English. Of the most misfitting judgements, only 2 out of 7 were made by independent judges, suggesting the independent judges were no more likely to make such judgements than the board judges.

Table 4: English Grade C Higher - biggest misfitting judgements

	Misfitting judgements	
Judge	*Std Resid*	*Calculated prob*
E1	8.34	0.014
I2	6.81	0.021
W1	5.05	0.038
I4	4.18	0.054
S1	3.92	0.061
S1	3.25	0.087
S2	3.06	0.096

When the judges themselves were analysed, it was found that there were no overfitting independent judges, whilst there were three board judges who overfitted. The only misfitting judge was independent. This suggests that independent judges did not overfit as much as board judges, but where there was misfitting, it came from an independent judge (Table 5).

Table 5: English Grade C Higher - misfitting/overfitting judges

	Misfitting/overfitting judges	
Judge	*Mean sq*	*mis/over –fit*
M1	0.597	overfit
M2	0.622	overfit
N1	0.764	overfit
I2	1.365	misfit

Looking at grade C on the Foundation tier, it can be seen in Table 6 that only one of the eight misfitting judgements came from an independent judge, suggesting the independents are less likely to make such judgements than the board judges.

Table 6: English Grade C Foundation - biggest misfitting judgements

	Misfitting judgements	
Judge	*Std Resid*	*Calculated prob*
C2	5.97	0.027
I2	4.90	0.040
W2	4.43	0.049
S1	3.96	0.060
W2	3.92	0.061
C2	3.65	0.070
W2	3.22	0.088
C2	3.17	0.091

Of the judges themselves, all four of the overfitting judges were board affiliated, whilst two of the four misfitting judges were independents. Again this shows that board judges tend to overfit more than independents, who are slightly more likely to misfit.

Table 7: English Grade C Foundation - misfitting/overfitting judges

	Misfitting/overfitting judges	
Judge	*Mean sq*	*Mis/Over –fit*
M1	0.518	overfit
C1	0.682	overfit
E1	0.782	overfit
W1	0.787	overfit
I6	1.229	misfit
W2	1.255	misfit
I4	1.317	misfit
C2	1.445	misfit

These results suggest that for English, the independent judges did not stand out from the board judges in terms of the number of misfitting judgements they made. The biggest misfitting judgements varied according to the grade and tier. At grade A it was the independents who made the biggest misfitting judgements; at C on the Higher tier there was no difference between the two; whilst at C on the Foundation tier the board judges made the biggest misfitting judgements. A higher proportion of the board judges tended to be overfitting than was the case for the independents, who were more likely to be misfitting judges.

Mathematics

The same tables were produced for the judges used in the mathematics study. The independent judges tended to make either very few, or a large number, of misfitting judgements, although the two judges making the most misfitting judgements (by some distance) were board judges (Table 8).

Table 8: Mathematics – number of misfits by judge

		Misfitting judgements		
Judge	*Affiliation*	*High A*	*Inter C*	*Total*
M1	MEG	1	2	3
I1	IB nominated	2	1	3
I4	LEA nominated	4	0	4
I2	LEA nominated	2	2	4
S1	SEG	5	0	5
N2	NEAB	3	3	6
W2	WJEC	4	2	6
S2	SEG	4	4	8
C2	CCEA	2	10	12
W1	WJEC	7	5	12
N1	NEAB	4	9	13
C1	CCEA	7	7	14
E1	Edexcel	2	13	15
I3	LEA nominated	10	5	15
I5	SQA nominated	12	4	16
E2	Edexcel	9	12	21
M2	MEG	20	5	25
Total		**98**	**84**	**182**

Three of the eight biggest misfitting judgements at grade A were made by independent judges, suggesting that the independents were no more likely to make misfitting judgements than the board judges (Table 9).

Table 9: Mathematics Grade A Higher - biggest misfitting judgements

	Misfitting judgements	
Judge	*Std Resid*	*Calculated prob*
C1	3.76	0.066
I5	3.56	0.073
M2	3.17	0.091
I1	3.11	0.094
I5	3.11	0.094
M2	3.05	0.097
S1	3.05	0.097
W1	3.05	0.097

Of the four overfitting judges at grade A, only one was independent, and similarly only one of the three misfitting judges was independent (Table 10). These proportions are similar to what would be expected if there were no differences between the two groups of judges.

Table 10: Mathematics Grade A Higher - misfitting/overfitting judges

	Misfitting/overfitting judges	
Judge	*Mean sq*	*Mis/Over –fit*
C2	0.661	overfit
I2	0.725	overfit
M1	0.731	overfit
S2	0.784	overfit
I5	1.268	misfit
E2	1.426	misfit
M2	1.480	misfit

At grade C on the Intermediate tier, three of the ten most misfitting judgements were made by independent judges, a proportion in line with the number of independent judges in the study (Table 11).

Table 11: Mathematics Grade C Intermediate - biggest misfitting judgements

	Misfitting judgements	
Judge	*Std Resid*	*Calculated prob*
E1	4.81	0.041
W2	4.81	0.041
E2	4.81	0.041
I5	4.30	0.051
W1	4.30	0.051
I5	3.35	0.082
W1	3.35	0.082
N2	3.07	0.096
M1	3.02	0.099
I3	3.00	0.100

Table 12 shows the overfitting and misfitting judges at grade C on the Intermediate tier. Of the three overfitting judges, one is an independent, whilst all four of the misfitting judges are board judges.

Table 12: Mathematics Grade C Intermediate - misfitting/overfitting judges

	Misfitting/overfitting judges	
Judge	*Mean sq*	*mis/over –fit*
I4	0.736	overfit
W2	0.737	overfit
S1	0.772	overfit
C2	1.223	misfit
E1	1.251	misfit
N1	1.370	misfit
E2	1.422	misfit

The results for Mathematics suggest there is little difference in the judgements of the independent judges compared with the board judges.

Science

Looking at Table 13, it can be seen that the independent judges tended to be found in groups. Two of them made very few misfitting judgements, three made 5 or 6 misfits, whilst one made 21 misfits. Overall, the independents did not stand out from the board judges.

Table 13: Science – number of misfits by judge

		Misfitting judgements			
Judge	*Affiliation*	*High A*	*High C*	*Foun C*	*Total*
I2	LEA nominated	0	0	1	1
I3	LEA nominated	0	1	2	3
S1	SEG	2	2	0	4
C1	CCEA	0	3	2	5
W1	WJEC	2	3	0	5
M1	MEG	0	3	2	5
I6	WJEC nominated	1	4	0	5
I1	IB nominated	2	2	1	5
I4	SQA nominated	4	2	0	6
E1	EDEXCEL	2	4	3	9
E2	EDEXCEL	2	0	7	9
M2	MEG	2	7	1	10
C2	CCEA	2	4	4	10
W2	WJEC	3	9	0	12
N1	NEAB	4	2	6	12
N2	NEAB	12	1	0	13
I5	SQA nominated	11	9	1	21
S2	SEG	1	18	3	22
Total		**50**	**74**	**33**	**157**

At grade A, four of the six biggest misfitting judgements were made independent judges, although three of these were made by the same judge (Table 14).

Table 14: Science Grade A Higher - biggest misfitting judgements

	Misfitting judgements	
Judge	*Std Resid*	*Calculated prob*
N2	4.06	0.057
I5	3.77	0.066
I5	3.58	0.072
I5	3.48	0.076
N2	3.33	0.083
I4	3.10	0.094

One of the four overfitting judges was independent, as was one of the two misfitting judges (Table 15).

Table 15: Science Grade A Higher - misfitting/overfitting judges

	Misfitting/overfitting judges	
Judge	*Mean sq*	*Mis/Over –fit*
S2	0.701	overfit
W1	0.742	overfit
I2	0.744	overfit
S1	0.786	overfit
N2	1.341	misfit
I5	1.909	misfit

At grade C on the Higher tier, four of the twelve biggest misfitting judgements were made by independent judges (Table 16), a figure in line with the proportion of independent judges in the study.

Table 16: Science Grade C Higher - biggest misfitting judgements

	Misfitting judgements	
Judge	*Std Resid*	*Calculated prob*
W1	5.42	0.033
W2	4.53	0.046
I5	3.79	0.065
S2	3.65	0.070
W2	3.61	0.071
S2	3.46	0.077
I1	3.42	0.079
I4	3.33	0.083
S2	3.30	0.084
W2	3.22	0.088
S2	3.04	0.098
I5	3.01	0.099

Of the overfitting judges at grade C on the Higher tier, two out of five were independent, whilst one of the three misfitting judges was independent. This, too, reflects the proportion of independent judges in the study.

Table 17: Science Grade C Higher - misfitting/overfitting judges

	Misfitting/overfitting judges	
Judge	*Mean sq*	*mis/over –fit*
I2	0.703	overfit
N2	0.744	overfit
W1	0.768	overfit
I3	0.776	overfit
N1	0.790	overfit
I5	1.348	misfit
S2	1.432	misfit
M2	1.439	misfit

On the Foundation tier at grade C, all three of the biggest misfitting judgements were made by board judges (Table 18).

Table 18: Science Grade C Foundation - biggest misfitting judgements

	Misfitting judgements	
Judge	*Std Resid*	*Calculated prob*
C2	3.57	0.073
M1	3.36	0.082
E1	3.01	0.100

Two out of the three overfitting judges were independent, but all three of the misfitting judges were board affiliated (Table 19).

Table 19: Science Grade C Foundation - misfitting/overfitting judges

	Misfitting/overfitting judges	
Judge	*Mean sq*	*Mis/Over –fit*
I2	0.734	overfit
I4	0.752	overfit
S1	0.792	overfit
E2	1.216	misfit
N1	1.233	misfit
C2	1.253	misfit

Summary

English

- Looking at all three grade boundaries together, there was no evidence that the independent judges were making more misfitting judgements than the board judges. Indeed, two of the three most ‘fitting’ judges were independent.

- Some evidence of independent judges being more likely than board judges to make big misfitting judgements, and for independent judges to misfit also, at grade A.

- Evidence of one independent judge making a big misfitting judgement, and this judge misfitting also, at Higher grade C, but no overall pattern.

- Slight evidence of independent judges misfitting, probably through consistently small misfitting judgements, but no overall trend.

Mathematics

- Looking at all three grade boundaries together, the independents tended to either make few misfitting judgements, or make many misfitting judgements – there were none ‘in the middle’.

- One independent judge made some big misfitting judgements, and was classed overall as a misfitting judge at Higher tier grade A. There was no evidence, however, that the independents were more/less likely to misfit/overfit than the board judges.

- The same independent judge (as at Grade A) made more big misfitting judgements at Intermediate tier Grade C. One independent judge was overfitting, but there was no evidence of independent judges differing from the board judges.

Science

- ❑ In terms of the number of misfitting judgements, the independent judges varied, with two judges making less than 4, but one judge making 21.
- ❑ At grade A (Higher tier), one independent judge made 3 of the 4 biggest misfitting judgements, and was also the most misfitting judge. If he is excluded, there is no strong evidence to suggest the independent judges were different from the board judges.
- ❑ At grade C (Higher) the picture was similar, whilst at grade C (Foundation) there was no clear picture, although 2 of the 3 overfitting judges were independent.

Conclusion

Whilst there are small differences at some grades in the different subjects, there is no evidence to suggest that independent judges are any less reliable than board affiliated judges at making the type of judgements associated with this type of study. It can be concluded that the inclusion of independent judges in the Thurstone pairs comparisons adds an important dimension to the study, and credibility to the results.

Acknowledgements

We would like to thank the judges and the members of the examining boards who contributed to the comparability studies on which this paper is based. In particular we would like to thank Austin Fearnley, Ben Jones, and Jeremy Pritchard for allowing us to use the data they had collected.

References

Andrich, D. (1978) Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, 2, 2, 449-460.

Bardell, G. S., Forrest, G. M. and Shoesmith, D. J. (1978) *Comparability in GCE: a review of the boards' studies, 1964-1977.* [Joint Matriculation Board, Manchester]

Davies, N. (2000) Fiddling the facts and figures in the struggle for success. *The Guardian* 11 July

Fearnley, A. (2000) *A Comparability Study in GCSE Mathematics: a review of the examination requirements, and a report on the cross-moderation exercise.* [Joint Forum for GCSE and GCE]

Forrest, G. M. and Shoesmith, D. J. (1985) *A Second Review of GCE Comparability Studies* [Joint Matriculation Board, Manchester]

Gray, E. (1999) *GCSE Science (Double Award) 1998 Comparability Study Report: syllabus review and cross-moderation report.* [Joint Forum for GCSE and GCE]

Johnson, S. and Cohen, L. (1983) *Investigating Grade Comparability through Cross-moderation.* [Schools Council Study, London]

Mathews, J. C. (1985) *Examinations: a commentary.* [George Allen and Unwin, London]

Nuttall, D. L., Backhouse, J. K. and Willmott, A. S. (1974) *Comparability of Standards between Subjects.* [Evans/Methuen Educational, London]

Pritchard, J., Jani, A. and Monani, S. (1999) *A Comparability Study in GCSE English: syllabus review and cross-moderation exercise – a study based on the summer 1998 examination.* [Joint Forum for GCSE and GCE]

van der Linden, W. J. (1994) Fundamental Measurement and the Fundamentals of Rasch Measurement. In *Objective Measurement: theory into practice Volume 2.* [Ablex Publishing Corporation, New Jersey]

Appendix 1 – The move from examination groups to unitary awarding bodies

examining groups/boards	change	unitary awarding body
CCEA	has remained as	CCEA
WJEC	has remained as	WJEC
MEG	now forms part of	OCR
Edexcel	has remained as	Edexcel
NEAB	now forms part of	AQA
SEG	now forms part of	AQA