Language Learning Journal, September 1996, No 14, 28–34

Reliability and validity in the GCSE Oral Examination

Brian Richards and Francine Chambers
University of Reading University of Southampton

INTRODUCTION

We have argued previously that too little attention has been paid to the reliability and validity of assessments by teachers. We have also claimed that moderation procedures are themselves flawed and cannot compensate for inconsistent marking (Chambers and Richards, 1992, 5). In the study below we consider a particularly problematic area – open-ended production tasks – taking as our focus the GCSE Higher Level conversation in French.

Research questions

Four questions were addressed, three of which were concerned with reliability, and one with validity:

1. To what extent do the linguistic background, training, and length of experience of teachers affect the reliability of their marking? How do native speakers compare with teachers who have learnt French as a foreign language?
2. To what extent does the ability range with which teachers are familiar affect the severity or leniency of their marking? Do teachers who are more experienced with the top end of the ability range mark more severely?
3. Are some types of marking criteria used by GCSE groups for open-ended tasks more reliable than others?
4. How valid is the assessment of features of performance such as 'range of vocabulary' and 'complexity of language'?

"The 30 conversations were re-recorded in two randomised sequences"

METHOD

Twenty-four teachers were asked to mark Higher Level conversations of 30 GCSE French candidates. They did so on two occasions, one month apart, using three types of assessment criteria on each occasion. Thirteen PGCE students completed an identical task on one occasion only.

The sample of candidates

Authentic recordings of 72 candidates doing the GCSE French oral were obtained from one school. From these, 30 Higher Level conversations were selected. Because we wanted teachers in the experiment to apply the assessment criteria rather than rely on their own intuitions, we made the task more difficult by excluding candidates who received grade 'A' in the overall examination. Final grades are shown in Table 1. The number of lower grades simply reflects the fact that not all children who are entered for Higher Level produce a Higher Level performance.

The 30 conversations were re-recorded in two randomised sequences, one for each occasion of marking. This prevented teachers hearing candidates in the same order on each occasion, thus minimising the influence of the first marking on the second.

The sample of teachers

The study involved 24 teachers of French from comprehensive and selective schools (12 from each). The latter group included teachers in both maintained grammar schools and independent schools. Each group of 12 teachers comprised six teachers whose first language was English and six whose first language was French (Table 2).

The teachers were all experienced and one third were heads of department. However, they varied extensively in the number of years they had taught; the average was 17.1 years (range: 2 to 37 years). An analysis of variance showed no differences between the groups of teachers in their years of experience. One additional group took part in the experiment: 13 PGCE students during the final fortnight of their course. All were graduates in French, but none were native speakers.

The assessment criteria

The three sets of assessment criteria were derived from existing GCSE practice at Higher Level.

A. *A criterion-related global impression scheme* (see Appendix A) assigned marks on a nine-point scale (0–8) supported by band descriptions of overall performance (cf. WJEC and SEG).

B. *A criterion-related categorical scheme* (see Appendix B) gave a separate mark on a four-point scale (0–3) for 'content', 'accuracy' and 'pronunciation'. Each point within each category received a description of performance (cf. NEAB and ULEAC).

C. *A norm-referenced categorical scheme* (see Appendix C) assigned an impression mark on an eight-point scale (0–7) for 'range of vocabulary', 'complexity of language' and 'fluency'. There were no descriptors.

To remove effects of the order in which teachers used the three schemes, a latin squares counterbalanced design was used, which systematically varied the order of use, both between occasions of marking and between the four groups of teachers.

Analysis

The statistical analysis was carried out by a researcher who was unaware of the identity of the teachers or their schools. This division of labour was important because it reduced the possibility of research bias, and individual participants and their schools could be sure that the quality of their performance would remain confidential, even from those carrying out the research.

The analyses made use of rank order correlation or product moment correlation (where appropriate), and Kendall's coefficient of concordance as measures of marking consistency or agreement between markers. Analyses of variance and *t* tests were carried out in looking for differences between groups of teachers or between marking schemes.

The three types of marking criteria produced seven sets of scores:

1. A global score (scale of 0–8) (Scheme A)
2. Content (0–3) (Scheme B)
3. Accuracy (0–3) (Scheme B)
4. Pronunciation (0–3) (Scheme B)
5. Fluency (0–7) (Scheme C)
6. Range of vocabulary (0–7) (Scheme C)
7. Complexity of structures (0–7) (Scheme C)

RESULTS

Self-consistency

For each teacher the rank order of the 30 candidates on the first set of marks was compared with the rank order of the same candidates on the second marking a month later. This was carried out for each of the seven sets of scores, giving a total of 168 (7 × 24) correlations.

Grade	Number
A	0
B	7
C	7
D	10
E	6

Table 1 **Grades obtained by the 30 candidates**

French native speakers		English native speakers	
Selective	Comprehensive	Selective	Comprehensive
6	6	6	6

Table 2 **Number of teachers, their first language, and type of school**

Correlations quantify the amount of agreement between two sets of values. They range from 1 (perfect agreement) through 0 (no relationship) to –l (a perfect *negative* relationship, i.e. if children who had been ranked highest on the first marking did worst on the second marking, and *vice versa*). The teachers with the highest correlation coefficients are therefore the most consistent markers – they placed the 30 candidates in a similar rank order on both occasions. Subsequent analyses which compared the self-consistency of the four groups of teachers or compared the reliability of all 24 teachers on the different marking criteria used logarithmic transformations of the correlation coefficients to calculate the average correlation.

General findings

1. Coefficients of self-consistency across the seven scores range from .93 (highly consistent) to .20 (no more consistent than if you had thrown dice on each occasion); the average value for the 24 teachers on the seven sets of scores ranged from .82 (Scheme A) to .50 (Pronunciation 0–3).
2. Despite huge differences between teachers in the consistency of their marking, there is no correlation between self-consistency and teachers' years of experience, except for a weak relationship for pronunciation, where experience is an advantage.
3. If you rank the 24 teachers for consistency on each of the seven sets of marks, there are close similarities between the seven sets of rank orders; consistent markers tend to be consistent (and inconsistent markers to be inconsistent) regardless of the marking scheme. Nevertheless, even the best markers can go astray on pronunciation and range of vocabulary.

Differences between groups of teachers

Overall, there is a slight tendency for the French native speakers in comprehensive schools to be most consistent, and for the non-native speakers in selective schools to be least consistent. However, only the following two differences are statistically significant:

1. In assessing Fluency (0–7) comprehensive school teachers were more consistent than those in selective schools.
2. In assessing Accuracy (0–3) native speakers were more consistent than the non-native speakers.

"there is a slight tendency for the French native speakers in comprehensive schools to be most consistent"

Differences in consistency on the seven sets of scores

An analysis of variance comparing the seven sets of scores showed that teachers' consistency varied across different criteria. Their ability to mark consistently was therefore significantly affected by the mark scheme.

There was greatest consistency on Scheme A (global) and least consistency on Scheme B (scales of 0–3 with descriptors). The following differences were statistically significant:

1. Scheme A is more reliable than all three components of Scheme B (0–3).
2. All three components of Scheme C are more reliable than Accuracy (0–3) and Pronunciation (0–3) in Scheme B.
3. Within Scheme B, Content (0–3) is more reliable than both Accuracy (0–3) and Pronunciation (0–3)

When self-consistency for the *aggregated* marks for the three schemes is compared (i.e. Scheme A compared with the totals on Schemes B and C), there are again significant differences between the three schemes: Scheme B is less reliable than Schemes A and C.

Relative severity or leniency of groups of teachers

The following results were statistically significant.

1. Native speakers marked more severely than non-native speakers on all sets of marks except Pronunciation (0–3), Fluency (0–7), Vocabulary (0–7) and the aggregate scores for Scheme C.
2. Teachers in selective schools were more severe than teachers in comprehensive schools on all sets of marks except Accuracy (0–3), Pronunciation (0–3) and the aggregate for Scheme B.
3. Results for Scheme A, Fluency (0–7), Vocabulary (0–7), Complexity (0–7) and the aggregate of Scheme B, show that while native speakers in selective schools are the most severe markers, native speakers in comprehensive schools are the most lenient. This interesting finding suggests two distinct responses by French native speakers depending on their experience of the UK educational system.
4. On Scheme A, teachers in selective schools assess able children more severely than other teachers, but they are more *lenient* with children below the mean. They appear to be more reluctant to award marks in the lower range or to give zero.

Comparison with PGCE students

Comparisons between PGCE students and the 24 teachers show students to be significantly more lenient in four areas: Complexity (0–7), Accuracy (0–3), Pronunciation (0–3), and the Aggregate scores for Scheme B.

Separate comparisons between the students and each of the four groups of teachers showed that:

1. The students were always significantly more lenient than native speakers in selective schools.
2. For Complexity (0–7), Accuracy (0–3), and the Aggregate of Scheme B (0–3) they were more lenient than non-native speakers in comprehensive schools.
3. On Accuracy (0–3), Pronunciation (0–3), and the Aggregate of Scheme B (0–3) they were more lenient than native speakers in comprehensive schools.
4. However, on some measures (Fluency 0–7, Vocabulary 0–7, Aggregate of Scheme C) they were more severe than native speakers in comprehensive schools.

Degree of agreement within each group of teachers

The reliability of assessment of candidates' performance depends not only on the ability of individual scorers to apply the marking criteria consistently; the criteria also have to be given a common interpretation across different markers. In other words there has to be both intra- and inter-marker reliability.

The inter-marker agreement for each of the four groups of teachers on the seven sets of scores was estimated using Kendall's Coefficient of Concordance (*W*). This statistic is calculated from the rank orders given to the 30 candidates by the six teachers in each group. Kendall's *W* ranges from1.0, which would result from perfect agreement between all scorers, to zero, which would indicate maximum disagreement. The resulting 28 coefficients were all statistically significant. In other words, despite differences between teachers' marks, overall agreement on the rank order given to the 30 children was always greater than chance, regardless of the scheme being used or the group the teachers belonged to. Nevertheless, strength of agreement varies substantially; coefficients range from .41 to .83.

Differences between the groups of teachers

The four groups of teachers were ranked from 1 to 4 according to their inter-marker reliability (as measured by *W*) on each of the seven marking schemes. The groups were then compared by calculating their average rank across the seven scores (Table 3). In fact groups varied little in their ranking on the different marking criteria – there was a clear tendency for the native speakers in comprehensive schools to have the highest level of agreement (on six schemes out of seven), and for the non-native speakers in selective schools to have the lowest level of agreement (on five schemes out of seven). It can be seen from Table 3 that we obtain the same rank order for agreement between markers as we did for self-consistency. In other words, groups containing teachers with the highest intra-marker reliability also achieved the highest inter-marker reliability.

Agreement among PGCE students

The PGCE students received a morning's formal induction in oral assessment. This began with an introduction to key concepts such as reliability, validity, discrimination, etc. There followed a comparison of marking criteria and speaking tests, and an evaluation of video material. The aim was to sensitise the students to key issues, rather than provide intensive training.

What is surprising is that, despite such meagre preparation, the students performed no worse than experienced teachers in respect of inter-marker agreement. If we look at the five groups of markers (four of teachers and one of students), coefficients of concordance placed the students in second or third place on each of the seven mark schemes.

Relative agreement between all 24 teachers on the different scoring criteria

Kendall's *W* was computed between all 24 teachers on each of the seven mark schemes. It can be seen from Table 4 that, while differences on most sets of scores are minimal, the coefficients for Accuracy (0–3) and Pronunciation (0–3) are lower than the others. These were the same aspects of performance which fared worst on self-consistency.

Validity

Chambers and Richards (1992) drew attention to the variation in terminology used by examining groups to describe performance, and the lack of definition of terms such as 'complexity of language', 'accuracy' and 'fluency'. Interviews with teachers (Chambers and Richards 1993) revealed some concern over the 'nebulous' nature of criteria, but, more importantly, teachers differed greatly in their interpretation of these concepts.

We chose two such terms from our own assessment criteria which we thought we could assess objectively if we took sufficient time and trouble. Teachers' assessments could then be correlated with our objective measures as a test of their validity. The two aspects selected were *range of vocabulary* and *complexity of structures*.

Range of vocabulary

An objective measure of vocabulary range was produced by simply counting the number of different words used by each candidate. To ensure that this was done accurately the conversations were transcribed in computer-readable format and the word counts carried out using the Computerised Language Analysis (CLAN) software (MacWhinney and Snow 1990).

One difficulty is that it is in the nature of conversations that the longer they are, the wider the range of vocabulary they contain. In selecting our 30 candidates we tried to standardise the length of conversations. Nevertheless, substantial differences remain in the quantity of speech produced in a conversation of approximately three minutes; in our 30 WJEC Higher Level conversations, the number of words spoken by the candidates ranges from 23 to 227 (mean = 114.7), while the number of different words ranges from 20 to 105 (mean = 62.7). It could be argued therefore that we are giving an unfair advantage to those who talk the most. Nevertheless, we decided to enter unweighted numbers of different words into our correlations, taking the view that it was valid to give credit to those whose lexical diversity was boosted by greater fluency. The rank order correlations between the teachers' assessment of vocabulary range and our objective measure varied between .87 and .48 with a median value of .77 and all were well above chance levels of agreement. Overall, therefore, teachers' assess-

Table 3 **Average rank order of the four groups of teachers on inter- and intra-marker reliability across seven sets of scoring criteria**

	Reliability	
Teacher Group	inter-	intra-
Native speaker/selective	2.9	2.7
Native speaker/comprehensive	1.1	1.1
Non-native/selective	3.6	3.9
Non-native/comprehensive	2.4	2.3

Table 4 **Coefficients of concordance between 24 teachers on each marking scheme**

Marking scheme		*W*
A. Criterion-related global impression	(0–8)	.69
B. Criterion-related categorical scheme	(0–3)	
Content		.68
Accuracy		.49
Pronunciation		.46
C. Norm-referenced categorical scheme	(0–7)	
Fluency		.69
Range of vocabulary		.70
Complexity of structures		.64
Aggregate of B		.64
Aggregate of C		.73

ments of vocabulary range had a high validity, in some cases remarkably so.

Complexity of structures

'Complexity of language' and use of 'complex structures' have been included in the assessment criteria of SEG and ULEAC respectively (see Chambers and Richards 1992). However, complexity has been a particularly elusive concept for linguists. Crystal concludes that

> 'it has not yet proved feasible to establish independent measures of complexity defined in purely linguistic terms .... largely because of controversy over the nature of the linguistic measures used, and the interference stemming from other psychological factors....' (Crystal 1991, 68).

"Overall, therefore, teachers' assessments of vocabulary range had a high validity"

We approached the task of developing an objective measure by asking our teachers to tell us what kind of linguistic features they would be looking for in assessing complexity (see Chambers and Richards 1993). We then allocated one point for each feature present in the candidates' conversations. In this way points were received for each different tense and mood, subordinate clauses, use of auxiliaries, and the correct use of certain prepositions and portmanteau words (e.g. *au, aux, du*). Some features expected by the teachers, such as complex uses of pronouns, relative clauses, constructions with *après avoir, avant de*, and participle constructions were not present in our sample (recall that Grade A candidates were not included). This gave us a maximum possible score of 9, out of which the candidates obtained between 1 and 7 with a mean of 4 points. We have no way of determining the external validity of our measure, but we found it to be powerfully correlated with the overall number of GCSE points obtained by the candidates. The measure appears therefore to be sensitive to an important dimension of linguistic performance in the examination.

Rank order correlations between our complexity measure and teachers' rating on the 0–7 scale varied from .60 to .23 with a median of .44. These coefficients are much lower than those reported above for vocabulary. In fact, for five of the 24 teachers the agreement between candidates' rank order on their assessments and their rank order on our own measure is no greater than chance. Since we based our measure on what teachers thought was important for assessing complexity, this low level of agreement could be the result of the diversity of their opinions, and the arbitrary fashion in which we converted them into a points system. On the other hand, it might also reflect the difficulty of assessing an undefined aspect of oral performance – there may be a discrepancy between what markers consciously believe to be complex linguistic behaviour and what they are sensitive to when making on-the-spot judgements.

DISCUSSION

It is important to remember that the teachers in our experiment were not assessing their own pupils, nor were they present for the examination – their role resembled that of moderator or external examiner rather than teacher-assessor. Nevertheless, the variation between the teachers in the consistency of their marking was surprisingly large. Depending on the marking scheme, performance on the dual marking extended from an astonishingly high consistency, to levels which could have been achieved by chance. Furthermore, even though some marking criteria were more reliable than others, consistent markers usually performed better than inconsistent markers regardless of the criteria.

Small differences between groups of teachers suggest slightly higher reliability for native speakers and for teachers in comprehensive schools. However two points need to be made in relation to this finding. Firstly, these effects were comparatively weak and were not always statistically significant. Secondly, one would expect highest reliability where teachers were marking the range of ability they were used to. Since some of the teachers from selective schools informed us that they expected most or all their pupils to obtain grade 'A' at GCSE, it is hardly surprising that they were less comfortable with the range of performance encountered in our experiment. On the other hand, the consistency of the native speaker group is a rebuttal to the occasional prejudice which we have encountered against native speakers. Again, this finding needs replicating with a larger sample, but an advantage for native speakers would be consistent with the idea that in making on-the-spot decisions, their greater efficiency in linguistic processing would leave more processing capacity spare to attend to the assessment criteria.

These results, and the fact that reliability was unrelated to years of teaching, suggests that quality, rather than quantity of experience is crucial. The additional finding that students with a minimum of preparation, and little experience in oral testing, could attain levels of agreement as high as experienced teachers, draws attention to the potential value of even a small amount of training. However, it must be remembered that we were unable to make comparisons with a control group of students who received no training at all.

Our prediction that teachers in selective schools would mark more severely proved to be correct. Native speakers also proved to be less lenient: we might have expected that native speakers would be less tolerant of deficiencies than someone who has learnt French as a foreign language. Nevertheless, closer inspection of our data revealed a more complex pattern – while native speakers in selective schools were the most severe markers, those in comprehensive schools were relatively lenient. We interpret this as indicating the influence of teaching experience on native speakers, with experience of the average and lower ranges of ability tending to reduce expectations.

With regard to the quality of the three types of assessment criteria, it is interesting to note the superiority of Scheme A (criterion-related global impression) and Scheme C (categorical without descriptors) over Scheme B (categorical with descriptors). Teachers who were interviewed after the experiment (Chambers and Richards 1992) expressed serious reservations about Scheme C because of its lack of descriptors, yet its reliability was equal to Scheme A and better than Scheme B. Further research is required to determine the reasons for the performance of the three types of criteria; the aspect of language proficiency to be rated, the presence and quality of descriptors, the use of a global *versus* categorical approach are all confounded in our experiment. Nevertheless, it is notable that within Scheme B, Accuracy and Pronunciation caused significantly more problems than Content, indicating that the linguistic domain and/or the quality of descriptors are important factors. On the other hand, even Content in Scheme B had lower self-consistency than Scheme A, suggesting benefits from a global scheme.

In investigating validity we chose two areas, Range of Vocabulary and Complexity of Structures, to determine the extent to which assessments were genuinely tapping the features of performance intended. Despite differences between the teachers, they generally demonstrated valid judgements of candidates' vocabulary. Complexity, by contrast, proved to be much more difficult, even though the consistency of assessments by individual teachers was relatively high. This seems to be a case of reliability without validity, and, given the differences of opinion among our teachers about what they understood by complexity, it demonstrates the need for the intended meaning of such terms to be made explicit by those devising marking schemes. In general, little attention is given to the precise definition of descriptors and the linguistic terminology they contain, as in: 'He (*sic*) shows a reasonable range of vocabulary and structures', 'He can use language flexibly' (Scottish Standard Grade examination). One suspects that many teachers eventually ignore the criteria, or parts of them, and rely on their own subjective impressions. This might explain the surprisingly high consistency on Scheme C.

One problem is that the choice of assessment criteria usually seems to be arbitrary. They are neither derived from a model of language proficiency, nor from an analysis of communicative development which would identify features of performance which cluster at different levels of proficiency. Our video recordings of French children being interviewed in their own language (Chambers and Richards 1995) indicate that notions such as 'complexity' need careful validation if they are to be included in assessment criteria. All the French children were skilled conversationalists in their own language on GCSE topics, but instances of the type of complexity thought important by teachers we interviewed were surprisingly elusive.

ACKNOWLEDGEMENTS

We should like to thank all the teachers, students, and schools who took part in the project, as well as David Satterly of the University of Bristol who helped to design the study, and Kim Dawes and Linda Williams of the Department of Applied Statistics, University of Reading who helped with the analysis.

"Native speakers also proved to be less lenient"

APPENDICES

ASSESSMENT SCHEDULES

A. CRITERION-RELATED GLOBAL IMPRESSION SCHEME

Choose one of the four bands most appropriate to the performance assessed and then allocate a high or low mark within this band.

Example:

0	1 2	3 4	5 6	7 8

Description of bands:

0 Not a Higher Level performance.

Band 1 The responses are brief and simple, with some inaccuracies and a good deal of hesitation. There is little attempt to achieve good pronunciation or intonation.
Responses could tax comprehension by a tolerant native speaker.

Band2 The responses are brief and simple but clear and accurate.
If more complex constructions are attempted inaccuracies occur and fluency is affected.
The interviewer may have to intervene or repeat questions.
Pronunciation is strongly influenced by the mother tongue.

Band 3 The responses are more sophisticated with a wider range of vocabulary and structures.
Factual information is conveyed accurately and without hesitation.
Expression of opinions and/or attitudes are less fluent. A genuine attempt at correct pronunciation and intonation is obvious.

Band 4 Both factual information and opinions/attitudes are expressed with ease and confidence, although there may be some hesitation.
Lexical and grammatical accuracy is high.
Pronunciation and intonation are increasingly good.

B. CRITERION-RELATED CATEGORICAL SCHEME

For Content, Accuracy and Pronunciation give 0, 1, 2 or 3 marks.

CONTENT

0 Not the appropriate quality for Higher Level.

1 Some information is conveyed, though this may be limited to simple facts.

2 A good deal of information on most aspects of the topics; some opinions are given.

3 Full descriptions or accounts are given; opinions are given freely.

ACCURACY

0 Not the appropriate quality for Higher Level.

1 Error incidence is quite high; comprehensibility may be impaired.

2 Occasional errors are of a minor nature and do not interfere with communication.
A sympathetic native listener would understand without difficulty.

3 Very few consistent or conspicuous grammatical errors.
A sympathetic native listener would understand immediately.

PRONUNCIATION

0 Not the appropriate quality for Higher Level.

1 Some non-native sounds but mispronunciation rarely leads to misunderstanding.

2 Few consistent or conspicuous mispronunciations; immediately comprehensible to a sympathetic native listener.

3 Near native accent. Good intonation.

C. NORM-REFERENCED CATEGORICAL SCHEME

For fluency, range of vocabulary and complexity of structures circle a mark on the scale 0 to 7.

No descriptions of performance are given.

Fluency	0	1	2	3	4	5	6	7
Range of vocabulary	0	1	2	3	4	5	6	7
Complexity of structures	0	1	2	3	4	5	6	7

REFERENCES

F. Chambers and B. J. Richards (1992). 'Communicative language ability and criteria for oral assessment in Modern Languages', *Language Learning Journal*, 6, 5–9.

F. Chambers and B. J. Richards (1993). 'Oral assessment: the views of language teachers', *Language Learning Journal*, 7, 22–26.

F. Chambers and B. J. Richards (1995). '"Free conversation" and the assessment of oral proficiency', *Language Learning Journal*, 11, 6–10.

D. Crystal (1991). *A Dictionary of Linguistics and Phonetics (3rd edn.)*, Oxford: Blackwell.

B. MacWhinney and C. Snow (1990). 'The Child Language Data Exchange System: an update', *Journal of Child Language*, 17, 457–472.