5. Moderating Assessment of Formal Speaking Performance[1]

Pip Neville-Barton

Abstract

A significant feature of the Diploma in English programme at UNITEC Institute of Technology is the student-organised conference. Each student is required to present a 15-minute paper which is assessed as part of their final mark. Although this is a worthwhile activity on many grounds, it raises significant pedagogical problems. Two particular issues are those of establishing performance criteria, and moderating assessment. Both of these have been largely resolved by the use of video. Students are understandably concerned about their public presentation in a foreign language, and about the requirements for gaining a pass. The use of video footage of previous conferences makes it feasible to model the performance expected, to clarify performance criteria, and to demonstrate that this task is possible for anxious students. Time restrictions and student numbers necessitate at least six parallel sessions, hence there is a need for several markers. Assessment of oral presentations on a topic of the presenter's choice is a highly subjective situation, thus effective moderation between markers is essential. The use of video footage in pre-conference moderation sessions has been shown to be both effective as a moderation tool, and efficient in terms of time.

Introduction

Formal oral presentations are a significant component of most English for Academic Purposes programmes. Not only do they challenge and exercise a wide range of linguistic skills, but they also test the study skills required of tertiary students who are frequently operating in a culturally unfamiliar academic environment. According to Brown and Hudson (1998:662) the validity of a well designed formal oral presentation task is enhanced if it measures the student's ability to respond to a real life language task (such as required at university), if it estimates the student's true language ability (at the discourse level rather than discrete item testing), and if it can predict the student's future performance in real life situations.

The importance of making satisfactory formal oral presentations in universities has been well documented in literature (Boyle,1996; Brown and Hudson, 1998). Ferris and Tagg (1996) carried out a survey amongst university students and their professors which indicated that the professors, in the majority of disciplines, ranked formal speaking as the most important of the academic oral skills. Significantly, students perceived formal oral presentations in the academic context to be the most difficult of the oral skills required.

Having established the place of formal oral presentations in English for academic purposes courses, this article addresses two main issues. The first relates to the preparation and performance of the presentation, and the second relates to inter-marker reliability in the judgment of the performance. Both these issues will be discussed within the context of a particular programme that prepares students for the academic world or for entry to the professional workforce. The results of an action-research study will be presented and the use of video technology, in a very accessible, straight-forward way, will be shown to be an effective tool in the modelling and moderating processes. It should be noted here that the study involves the analysis of that data which was available from a regular, pre-event moderation session and should be viewed only as a pilot for further formal research.

Institutional background

The School of English and Applied Linguistics at UNITEC Institute of Technology, Auckland offers a variety of programmes of intensive English at all levels to a mix of permanent residents and international students. The School also offers teacher training and administers the biggest IELTS centre in New Zealand. There are 450 students currently enrolled in the School of English and Applied Linguistics.

The Diploma in English Programme

The research took place within the Diploma in English programme which is aimed at the highest English language proficiency level in the School. The entry level for this programme is minimum Band 5.0 on the IELTS scale or equivalent. It is a one-year full time programme which offers five different one-semester (16-week) certificate courses as follows:

General English

- English (Upper Intermediate),
- English (Advanced)

English for Specific Purposes

- English for Professional Employment,
- English for Business and Computing Studies
- English for Academic Studies.

To gain the Diploma, students must successfully complete two of the one-semester courses.

In most semesters, student numbers total approximately 100, and although a number of these are overseas students on study visas, the majority are new immigrants who qualify for some government subsidies. The ethnic mix comprises students mainly from PR China, Taiwan, Hong Kong, Korea, Russia, and Eastern Europe.

The Context of the Research: Student Conference

Background

A significant feature of the Diploma in English is the student conference which is held as a day-long event at the end of each semester. The conference is organised by the students with the help of an elected student committee and some staff assistance. A conference theme, related in a general way to life in New Zealand, is chosen by all the students. With the exception of the Upper-Intermediate class, each student is required to present a 15-minute paper on an approved topic, which is assessed as part of their final mark. Since its inception five years ago, the conference has become the highlight of the programme and its success as an event can be evidenced by the enthusiasm of the students and by the audience the conference attracts from within and outside of the School. The organising committee is encouraged to invite an opening and closing speaker from the wider community. These speakers have included Members of Parliament and business leaders.

The Assessment

The main requirements of the assessment are:

- the presentation should be 15 minutes in length followed by 5 minutes question time;
- the topic must be related to the conference theme and to the New Zealand context;

- students are expected to fulfil the requirements of the performance criteria in the major areas of preparation and content, presentation style, language production, and the use of visual aids.

In preparation for the conference students are required to write an abstract of their presentation which must be approved by their teacher. These abstracts are published by the students in a conference booklet.

Each of the ten performance criteria on the marking schedule is rated on a 0 - 5 point scale and timing penalties may be applied. Results are recorded in three categories: **merit pass (42-50), pass (32-41), fail (0-31)**. This is an important assessment which must be passed in order for students to pass the whole course. Because the conference is held in the last week of the semester, there is insufficient time for students to resubmit a failed performance. Hence it is essential that students understand what the criteria mean in practice and are given every chance to pass the assessment at the conference.

Modelling Performance

As students begin to prepare their presentation, they are understandably concerned about their ability to present in a foreign language and are anxious about the requirements for gaining a pass. This presentation may not be the first presentation they have done in class but it is the most important and performed before a much larger audience. Through previous practice, the students have been introduced to the type of performance criteria expected (and in some instances have contributed to the development of the criteria).

Although students are expected to do the majority of work for their speeches outside the classroom, considerable time is also spent with the teachers preparing for this assessment. This time is well spent. The task demonstrates the positive characteristics for alternative assessments referred to in Brown and Hudson (1998:654) in that it requires students to create, produce, perform, and to focus on the process as well as the product. Brown and Hudson (1998:662) also claim that "well designed performance assessments can provide positive washback effects especially if they are linked to a particular curriculum". In other words, the preparation process is valuable as it revises and reinforces much of the completed course-related work which now has to be applied to a coherent piece of spoken academic discourse.

The use of video footage of previous conference presentations has been found to be one of the most valuable aspects of the preparation process. It complements the teaching content in a very practical way thus saving valuable classroom time and giving students a chance to view for themselves how other students have coped with the same assessment. Many students panic in anticipation of talking in English for 15 minutes to a large audience. By using the videos as models they can see that the task is within their ability. The videos also demonstrate the depth of the task required and the commitment necessary to produce a satisfactory performance in this final course assessment. Although ethical issues prevent students from viewing poor or failed performances, students develop a critical awareness of the level of performance required to satisfactorily fulfil the performance criteria. Student response to the usefulness of this type of video modelling has been extremely positive.

Moderating Assessment

Because of the time restrictions and the number of students taking part in the day-long conference, it is necessary to run approximately eight parallel sessions. The sessions are marked by teachers from within and outside of the programme and every effort is made to ensure that teachers do not mark their own students. To facilitate this requirement and to spread the marking load amongst busy teachers with other timetable commitments, it is necessary to have at least twelve markers on the schedule. The number of markers involved and the importance to the students of this assessment mean that markers must be familiar with the performance criteria and agree with the standards required. Thus inter-marker reliability becomes a prime objective of all pre-conference moderation sessions.

Inter-marker reliability

> Inter-rater reliability refers to the degree of similarity between different examiners; can two or more examiners, without influencing one another, give the same marks to the same set of scripts or oral performances? (Alderson, Clapham Wall 1995:129)

Alderson et al. (1995) say that it is unrealistic to expect this to happen all the time. Marking oral assessments is a subjective situation and inter-marker reliability is difficult to achieve.

This is illustrated to some extent in the debate in the January *ELT Journal* between Saville and Hargreaves (1999) and Foot (1999). Saville and Hargreaves support the practice of double marking to increase reliability, the usefulness of which Foot (1999:53) queries. Double marking is also advocated by Alderson et al. (1995). However, as in the case under discussion, double marking is not always a practical option.

Given this situation, other assurances of reliability are required. There seems to be accord amongst the researchers that taking part in regular moderation sessions is necessary in order to develop a common understanding of criteria and to attempt to arrive at a common standard (Alderson et al., 1995; Brown and Hudson, 1998; Brindley, 1998; Saville and Hargreaves, 1999).

> The successful functioning of a speaking test (...) relies heavily on a system for training and standardising the oral examiners so that they rate accurately and consistently. (Saville and Hargreaves, 1999:49)

The Study

The study took place during the pre-event moderation sessions for the markers involved. There were three sessions: one prior to the Semester One conference (Session One) and the other two (Session Two and Session Three) prior to the end-of-year conference. Most of the seven Session One markers, were also present in either Session Two (6 markers) or Session Three (6 markers). As part of a process of informal, action-research type development, it was decided to record data at these sessions. It was hoped that this would give some indication of the extent of any reliability problems and indicate areas for improvement.

Procedure

- At each session the criteria on the marking schedule were discussed (see Appendix One). Most of the markers were already familiar with the criteria.
- Two student presentations, one good and one mediocre, videoed from previous conferences, were played at each Session. (The Session One videos were different from those of Sessions Two and Three). In Sessions Two and Three the same videos were used but they were presented in reverse order. This gave a total of six view-

ings from which data was available.

- Markers were asked to assess the first video without discussion and to hand in their marking schedule.
- This was followed by a lengthy discussion which addressed each criterion in turn. These lively interactions forced markers to re-evaluate the performance criteria.
- Markers assessed the videos again and handed in a second marking schedule.
- The procedure was repeated for the second video.

Results

Overall Reliability of Final Mark

The final mark is out of a possible 50. From the point of view of the students it is the Pass/Fail (Pass 32+) and the Pass/Merit (Merit 42+) boundaries that are important. Although the students receive a copy of the marked assessment schedule, the actual mark within each category is not reported on their final academic record.

Before the Discussion

By examining the standard deviation of each group of markers after they watched each video and prior to their discussion, it was possible to assess the extent of the differences between their marking. A standard deviation of less than 1.5 was (arbitrarily) regarded as acceptable. Three of the six viewings produced larger values than this: 5.14, 3.74, 3.25. These all occurred on the three viewings of mediocre performances. The viewings of the good performances produced acceptable standard deviation values (1.40, 1.21, 0.75).

Effect of Discussion

All standard deviations reduced after discussion, and the unacceptable ones came within or close to the acceptable range (1.14, 1.25 ,1.70 respectively).

Reliability of Performance Criteria

As the same videos were used in Sessions Two and Three, enough data was available to indicate the reliability of markers on each performance criterion separately. There were 12 markers for each video. Possible marks on each criterion ranged from 0 – 5 and half marks could be awarded.

Before Discussion

The standard deviations of the marking of each criterion prior to discussion, showed that the reliability of markers depended upon the level of student performance. When marking the good performance, the largest variation occurred with the criteria 'Delivery', 'Pronunciation' and 'Use of at least one other type of Visual Aid'. (These standard deviations were between 0.4 and 0.5). The smallest variation in marking occurred in 'Organisation' and 'Questions' (complete agreement).

The mediocre performance proved much harder to judge on all criteria; every standard deviation was greater than 0.4. The largest variation was 'Language Appropriate to Audience and Topic' (s.d. = 1.11). This criterion was not a problem with the good performance (s.d. = 0.29). The smallest variations occurred in "Choice of Topic', 'Organisation', 'Main and supporting Ideas', 'Delivery' and 'Pronunciation'.

Effect of Discussion

As expected, the variation on most criteria reduced significantly after discussion. However, 'Main and Supporting Ideas' did not improve in either the good or the mediocre performances (all standard deviations in the range 0.31 - 0.49).
In the good performance, little improvement in standardisation was also shown in 'Delivery' (0.5 to 0.42), and in 'Pronunciation' (0.49 to 0.33).
In the mediocre performance, little improvement was also shown in 'Organisation' (0.47 to 0.41) and 'Use of OHT' (0.86 to 0.73).

Discussion

It is clear that moderation sessions are effective. All standard deviations reduced and inter-marker reliability improved to generally acceptable levels.

Despite having previous experience, wide variations amongst all markers still occurred when first viewing the video during each moderation session. This indicates, therefore, that it is necessary to hold moderation sessions on a regular basis immediately prior to formal oral presentation assessments and to expect all markers to attend. Saville and Hargreaves (1999:50) talk about "on-going commitment to validation involving data collection, monitoring and evaluation". This level of commitment should be a goal in all professional and assessment development programmes.

A major point emerging from the data is that the situation is different when evaluating different standards of performance. The inter-marker reliability problem is only significant with mediocre performance. This, and the results from the performance criteria analysis, indicate that marking mediocre performance is more difficult than marking good performance. Therefore, future moderation sessions in this programme should focus mainly on videos of mediocre performances. In this particular context, it would also be useful to use videos showing performances on the Pass/Fail, Pass/Merit boundaries rather than videos that are clearly within one of these three categories.

With respect to particular criteria, one of the conventional wisdoms is that grammar is the most difficult category to evaluate in an oral performance. It is interesting to note that there was acceptable agreement amongst the markers on this item for both the good and the mediocre performances. 'Pronunciation' and 'Delivery' on the other hand, caused problems. This indicates that more research would be helpful in this area. Furthermore, there is a need to review and develop several performance criteria and to clarify their interpretation. This is part of the "evolutionary process of change and further revision ..." referred to by Saville and Hargreaves (1999:50).

Summary

This paper has described both the context of a formal oral presentation as a major assessment in an EAP course, and the moderation process that has been initiated to increase the consistency of marking. Analysis of the data gathered from three moderation sessions has been presented and some useful conclusions have been made in relation to inter-marker reliability and to the individual performance criteria.

We have seen that the use of video as a model in the classroom enhances the learning process. It has proven helpful to students in the preparation stages by increasing their confidence and promoting their understanding of the criteria and awareness of the standards expected of them. We have also seen the usefulness of video in the pre-event moderation process to facilitate inter-marker reliability.

Formal oral presentations are frequently required at university and in the professional workplace. There is little doubt that a performance-based assessment such as this has considerable merit in an English for Academic Purposes course. There is also little doubt that to ensure consistency of marking, regular, vigorous moderation practices must be in place and the use of video technology in a normal, low-key way is both sensible and effective.

Postscript

As a result of this study, marker attendance at pre-conference moderation sessions is now scheduled at the beginning of each semester and is no longer optional. In addition, the performance criteria have been amended to clarify each criterion and reduce possible ambiguities (see Appendix Two). A further study is about to be undertaken to gauge the effects of these changes.

Since this study was completed, a poster presentation has also been introduced to provide an alternative task for students who have achieved the required standard of presentation at a previous conference and would like to try a new challenge.

Pip Neville-Barton, MA Applied (Hons), is a senior lecturer in English for Academic Purposes in the School of English and Applied Linguistics at UNITEC Institute of Technology, Auckland, New Zealand. In 1993 she was appointed by the New Zealand government as leader of a teaching team to Guizhou University, China. As a programme leader in UNITEC, her research interests developed in moderation practices and reflective teaching practice. Her current research interest is in the area of immigration and the relationship between English language learning and the lives of new immigrants.

Note

[1] This paper is an amended version of a paper that was presented at the British Association for Lecturers of English for Academic Purposes conference "Issues in EAP Learning Technologies" at the University of Leeds, UK, in 1999.

References

Alderson, C., C. Clapham, D. Wall. *Language Test Construction and Evaluation.* Cambridge:CUP, 1995.

Boyle, R. "Modelling oral presentation" *ELT Journal*, Vol. 50 No 2, 1996, pp.115-126.

Brindley, G. "Assessment in the AMEP: Current trends and future directions" *Prospect*, Vol.13 No 3, 1998, pp. 59-71.

Brown, J.D., T. Hudson. "The Alternatives in Language Assessment" *TESOL Quarterly*, Vol. 32 No 4, 1986, pp.53-675.

Ferris, D., T. Tagg. "Academic Listening/Speaking Tasks for ESL Students: Problems, Suggestions, and Implications" *TESOL Quarterly*, Vol. 30 No 2, 1996, pp.297-320.

Foote, M.C. "Reply to Saville and Hargreaves" *ELT Journal*, Vol. 53 No1, 1995, pp.52-53.

Saville, N., P. Hargreaves. "Assessing speaking in the Revised FCE" *ELT Journal*, Vol. 53 No 1, 1995, pp. 42-51.

Upshur, J.A., C. E. Turner. "Constructing Rating scales for second Language Tests" *ELT Journal*, Vol. 49 No 1, 1995, pp. 3-12.

Wharton, S. "Teaching Language Testing on a Pre-Service TEFL course" *ELT Journal*, Vol. 52 No 2, 1998, pp.127-132.

Appendices overleaf

Appendix One *Conference Presentation Assessment Criteria (1)*

Name of presenter ______________ Name of marker ______________

Class ______________ Date ______________

	Yes, definitely		Yes, to some extent		Not really	No
A. PREPARATION AND CONTENT						
1. **Choice of topic:** Well researched, interesting, informative topic made relevant to the audience and related to NZ	5	4	3	2	1	0
2. **Organisation:** Clear introduction, Clear conclusion, Well-structured and cohesive, Good use of 'signpost' words	5	4	3	2	1	0
3. **Main and supporting ideas:** Main ideas/points clearly explained? Good supporting statements? Enough examples, details	5	4	3	2	1	0
B. PRESENTATION STYLE						
1. **Delivery:** Good use eye contact/body language? Voice — audible and varied tone? Good use of notes? (not read)	5	4	3	2	1	0
2. **Language:**						
a) Grammar accurate?	5	4	3	2	1	0
b)Pronunciation clear?	5	4	3	2	1	0
c) Language appropriate to audience and topic?	5	4	3	2	1	0
3. **Questions:** Questions from the audience effectively dealt with? Asked for clarification if question not understood?	5	4	3	2	1	0
C. USE OF VISUAL AIDS						
1. **Use of Overhead Transparencies:** OHP used effectively? OHP well prepared — easy to read/language correct?	5	4	3	2	1	0
2. **Use of at least one other type of visual aid:** Visual aid(s) relevant/appropriate and language correct?	5	4	3	2	1	0

Start time __________ **Finish time** __________

Overall comment:

Sub Total: ______________

Timing Penalties
One mark per
minute deducted if
talk under 13 mins
or over 17 mins: ______________

Pass: 32/50   *Merit Pass: 42/50*

Total Marks: ______________ **/50**

Appendix Two *Conference Presentation Assessment Criteria (2)*

Name of presenter ______________ Name of marker ______________

Class ______________ Date ______________

	Yes, definitely		Yes, to some extent		Not really	No
A. PREPARATION AND CONTENT						
1. Choice of topic: Well researched, informative, made relevant and accessible to the audience, related to NZ	5	4	3	2	1	0
3. Organisation:						
a. Clear, effective introduction, Clear effective conclusion	5	4	3	2	1	0
b. Well structured and cohesive, Good use of 'signpost' worlds	5	4	3	2	1	0
c. Main and supporting ideas, Main ideas/points clearly explained? Good supporting statements? Enough examples, details	5	4	3	2	1	0
D. PRESENTATION STYLE						
1. Delivery: Good use eye contact/body language? Voice — audible and varied? Good use of notes? (not read) Well-paced?	5	4	3	2	1	0
4. Language:						
a) Grammar accurate?	5	4	3	2	1	0
b)Pronunciation clear?	5	4	3	2	1	0
5. Questions: Questions from the audience effectively dealt with? Asked for clarification if question not understood?	5	4	3	2	1	0
E. USE OF VISUAL AIDS						
1. Use of Overhead Transparencies: OHTs/slides well-prepared, easy to read, language correct? Technology used effectively?	5	4	3	2	1	0
3. Use of at least one other type of visual aid: Helpful/effective?	5	4	3	2	1	0

Start time __________ **Finish time** __________

Overall comment:

Sub Total: ______________

Timing Penalties
One mark per
minute deducted if
talk under 13 mins
or over 17 mins: ______________

Pass: 32/50 *Merit Pass: 42/50* **Total Marks:** ______________ **/50**