Diagnosing Students' Proficiency on a Spoken Performance Assessment

Paul Anthony Marshall
paulanthony.marshall@gmail.com

Abstract

The aim of this study was to diagnose specific gaps between current student proficiency and a target standard of proficiency in presenting a daily bulletin, in order to make an informed decision about what I can do to help students to span these gaps. After much trial and error with a variety of diagnostic tools, the study uses a thematic chart to successfully identify gaps in student ability on the performance assessment. Here, I begin by outlining the methodology, which has been broken down into three stages: definition of performance criteria, rubric development, and rubric operationalization. I then go on to reflect on the successes and shortcomings of the process and the decisions made.

Introduction

I was recently teaching English for Specific Purposes to eighty Laotian nationals at an Australian-managed gold and copper mine in southern Lao P.D.R. The management of the Training Department decided that the "Professional English" course should switch to using the Australian vocational performance assessment system of competency-based assessment which essentially meant assessing students on practical work-related tasks such as meetings or presentations. This decision was made part-way through the course which had already been fully planned and partly delivered so I needed to develop an effective approach, and quickly. Based on articles that I had read on formative assessment, I saw it as a possible vehicle to drive my students to success on performance assessments.

The first stage would be to build up a clear picture of my students' current levels of proficiency and of a realistic target level of proficiency. In order to do this, it would be necessary to carry out a formative assessment of students completing the task. As assessment criteria did not yet exist, I first set out to determine appropriate criteria. Ideally, these would be criterion-referenced in order to assess students according to an external, standardised set of criteria, which have been tried and tested.

Sadler (1989, p. 119) focused on "the nature and function of formative assessment in the development of expertise" where "student outcomes are appraised qualitatively using multiple criteria" and discusses the benefits and drawbacks of qualitative judgments, the use of descriptors, fuzzy, as opposed to sharp, criteria, and metacriteria, the criteria for using criteria. It provided guidance for many of the micro decisions made in this study. Black & Wiliam (1998) provided excellent procedural input for the implementation of formative assessment in the classroom which I used while planning the initial diagnostic stages. Huhta (2008) deals with the nuances between the definitions and functions of a variety of assessment types, and also introduces the idea of diagnostic competence which led me to use video to record student presentations. Biehler and Snowman (1997) contributed understanding of the importance of measurement and evaluation in the process of performance testing and during the analysis of test results. Davison and Leung (2009) supplied an insightful exploration of possibilities for using assessment for learning in the classroom.

While all of these articles provided inspiration and methodological input on utilising formative assessment to improve student competence on performance assessments, this study focuses only on the initial step; namely that of diagnosing areas of weakness for potential focus for formative assessment techniques. My research into the diagnostic evaluation of student presentations also consisted of

collecting assessment rubrics and an instructional article by Simkins (1999), both of which I utilised to select the most suitable assessment criteria, write descriptors, and design rubrics for the specific task of presenting a daily bulletin in my specific context.

Method

Participants

Before testing out the criteria, I needed to select a manageable set of performance assessments to try them on. I took a number of factors into consideration when choosing three students to represent the entire population of eighty Intermediate Professional English students. I had been teaching most of the members of this course for almost three years, and I was confident that the three students were representative of the entire Professional English population in terms of gender, the range of ages, backgrounds, professions, and the range of competence in English fluency, comprehension, and presentation skills. I felt that three students was a sufficient number for a small-scale study, and choosing an odd number avoided the possibility of split results. I sat with all three students and explained to them what I was asking of them.

Instrument Development

In order to formatively assess students' performance assessments comprehensively, I first had to select or create some appropriate criteria. The most effective method of assessment I had experience of was IELTS speaking and writing examinations which use a nine-band rubric. IELTS Examiners attend standardisation training in order to make sure they are all interpreting the criteria in the same way. However, by personally assessing my students against criteria, the results of the data generation would hinge on my own concept of the target standard, and were based largely on my own independent evaluations of student performances. Assessing student performance against multiple criteria and based on a target standard determined only by the teacher is by definition subjective, "the teacher must possess a concept of quality appropriate to the task and be able to judge the student's work in relation to that concept" (Sadler, 1989, p. 121).

The process of designing a suitable instrument consisted of a great deal of trial and error. Before experimenting with a group of existing oral presentation skills rubrics I had gathered to assess videos of my students' presentations (McCullen, 1997; NCTE/IRA, 2004; Swinton, 2012), I excluded irrelevant or inappropriate criteria from them such as those related to presentation slides. Where similar criteria existed on more than one of the original rubrics, I selected those that I judged to be most relevant to my students in their context. While I favoured the idea of an IELTS-style rubric with comprehensive descriptors, I decided to use universal descriptors, knowing that I would rewrite the rubric in a substantial way after this initial trial run.

This process resulted in Rubric A, shown in Figure 1, which combined the most suitable success criteria from a range of oral presentation skills rubrics. However, after viewing the three videoed presentations numerous times using Rubric A, I felt that the universal descriptors were unsuitable for the task, and the criteria needed reviewing. I went on to try out several more rubrics which had a variety of formats and some alternative, but similar criteria. I hand-wrote notes onto these rubrics about their strengths, weaknesses, and suitability in order to further refine the rubric. Following Simkins (1999), I limited the number of criteria to four because this forces the designer to prioritise which are the most important. I grouped together similar criteria, and incorporated criteria-specific descriptors for the groups to create Rubric B, shown in Figure 2. Again following Simkins (1999, p. 23), I created four levels of descriptor for each criterion because three levels does not provide sufficient discrimination but more than four leads to splitting hairs.

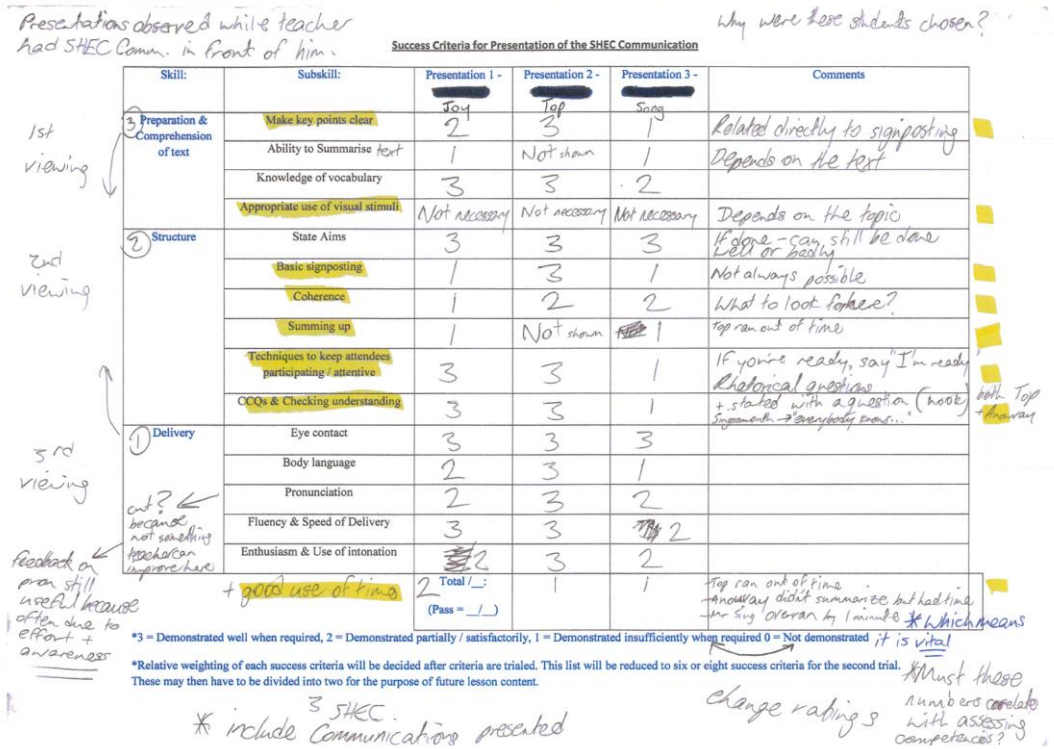

Figure 1. Rubric A

FINAL RUBRIC	1	2	3	4
Structure Coherence Basic signposting	Clear intro-content-summary / conclusion. Coherent and basic signposting used effectively.	Logical structure OR coherence / signposting.	Some structure OR insufficient but present coherence / signposting.	No identifiable structure. Incoherent. No signposting.
Getting Key points across Checking understanding	Made key points abundantly clear and made certain that they had been understood. Good awareness of audience understanding.	Made key points mostly clear, some checking of understanding OR one of the two points in box 1. (Awareness.) * Some awareness.....	Key points mostly unclear OR insufficient checking of understanding / awareness of audience understanding.	Key points unclear. No checking of understanding. No awareness of audience understanding.
Good use of time	Time used effectively for the purpose of content delivery.	Time used effectively but marginally ran out / overran / missed an opportunity.	Time used ineffectively – some content missed.	Important content missed.
Techniques to keep interest	A combination of rhetorical question, visual stimuli and ~~animated~~ personal style kept audience attentive.	Mostly kept audience attention through one or two techniques but could have been executed more effectively.	Kept audience interest some of the time but insufficiently.	Kept little or no audience interest.

(animated personal style = engaging style")

Not explains of main points
3 Little structure
3 coherent
3 No signposting
3 NO CCQs
3 No summary
3 basically reading with good intonation
no visuals fingers
1 signposting
2 clear not enough CCQs
2 No summary
as ✓ If you're ready, say "I'm ready!"
3
2/3 large void here
3 No summary
3
difficult and dull subject matter which was not chosen or organised by presenters.

Figure 2. Rubric B

Peer feedback

I discussed Rubric B with a colleague and received some brief feedback on it which can be seen hand-written onto it in Figure 2. I then used Rubric B to assess the three videoed presentations during repeated viewings, and hand-wrote very brief notes on student performance onto the rubric. This trial of Rubric B allowed me to evaluate the strengths and weaknesses of grouping criteria together, the descriptors I had written, and of my students' performances. I came to the conclusion that the grouping of criteria made assessment more difficult because frequently students would achieve one criterion but not the other in the same group. The descriptors did not allow for this eventuality. I also realised that limiting the number of criteria to four was completely unnecessary in this case because my purpose for the use of criteria was diagnostic and not to provide feedback or report progress.

Data collection

All of the evaluations were done by watching pre-recorded videos of student presentations. As a reaction to the results of trialling Rubric B, and after having started reading into data analysis and interpretation, I decided to alter the data generation process to evaluate the videoed presentations more thoroughly. While I was trying to decide how best to present my data, I considered presenting the comments in paragraphs by presenter, or in paragraphs by criterion but essentially I was searching for a method of presentation which, following Spencer, Ritchie, and O'Connor (2003, p. 210), allows searches to identify thematic categories and patterns and shown associations between phenomena within persons and between persons or groups of persons. As a result, I decided that it would be logical and easy to reference if this data could be searched by both presenter and criterion on one table, leading to the thematic chart shown in Table 1.

The thematic chart was not pre-planned; it was a contingency which I feel considerably improved the descriptive quality of the data gathered, which in turn facilitated my analysis of the data. The additional column for general comments about each presenter, and the additional row for general comments about each criteria meant that the data was not limited to my preconceived categories. I eventually prepared and processed my data and presented it in different formats to aid with analysis and interpretation, and to ensure it could be easily accessed and referred to.

I viewed the videos numerous more times while writing evaluative comments into the thematic chart for easy reference by criteria and by presenter. I was becoming very familiar with my students' presentations by this time which in itself meant that I could evaluate them in much more detail. I also included examples of actual presenter monologue where possible. Sub-dividing comments and monologue by specific criteria meant that I could specifically diagnose what students need training on, but it also served the additional purpose of categorising the data in preparation for analysis and interpretation.

On completion of the thematic chart, I assigned criteria to what I perceived to be the most enlightening four classes at a higher level of abstraction; questioning, emphasis, audience understanding, and time. Following this, I created an extra column at the end, and an extra row and at the bottom of the rubric. I used these to write a brief summary of the information included about each criterion, and about each presenter. This process aided both the analysis, and the interpretation of data.

One of the most useful and revelatory patterns that resulted from sorting and categorising my data was a possible insight into the thinking of the presenters. I discerned from the data that the presenters did not appear to assume responsibility for audience understanding. This can be implemented through asking questions to check understanding, emphasising key points, personalising, and concluding. The identification of this pattern will enable me to further observe this phenomenon, and to plan future lesson content based on this need.

Table 1

Thematic chart displaying assessment observations

Criteria: Students:	**1a Introduction: stating topic, activating schemata, creating interest**	**1b Rhetorical questions**	**1c Questioning the audience**	**1d Emphasis through repetition**	**1e Emphasis through stress**	**1f Emphasis through visual aids**	**1g Awareness of audience understanding and interest**	**1h Checking understanding of key points**	**1j Personalising / contextual-ising the content**	**1k ~~Summarising~~ Concluding**	**1m Good use of time**	**Comments on each presenter**
Joy	Joy stated the topic, then used a rhetorical question as a sort of hook to introduce the topic**. "So do you know how fires start from welding? OK, I can tell you now."** More would have been better.	Joy used one rhetorical question at the start**. "So do you know how fires start from welding? OK, I can tell you now."** More would have been better.	No other questions were asked. The opportunity to check understanding and / or contextualise the content was missed.	The key points were not repeated. This could have been an effective way of making sure the audience understood what the key points were.	Joy used intonation very effectively to keep audience interest, and to emphasise the key points.	No visual aids were used, but the content didn't necessitate the use of visual aids.	Very little awareness of audience understandings hown other than monitoring and maintaining interest by making eye contact.	This was not done despite finishing early. A missed opportunity.	This was not done. Joy could have asked the audience for personal experiences related to the topic.	This was not done. A missed opportunity to emphasise the key points through repetition, personalisation, or to check understandingof key points.	The missed opportunity to summarise or personalise the content (despite finishing early) was one of the main weaknesses of Joy's presentation.	Joy uses intonation, eye contact, and body language effectively but could benefit a great deal from using the other techniques listed here.
Top	Top introduced himself, stated the topic, & used a rhetorical question to spark interest. The question could have been more effective. He signposted **"today I'm going to talk about six ways..."**	Some rhetorical questioning. More would have been better. Top kept checking audience agreement with the points he was making by asking **"Yes?"**	This was done only briefly at the start**: "Do you think accidents are a kind of luck?" "Do you think that accidents can be prevented?"**	Top's checking of audience agreement was a method of repetition and was used to highlight the topic but not the key points.	Intonation was used effectively to keep audience interest and to emphasise the meaning of the topic, although the key points were not emphasised.	The only visual aids used were fingers to show the number of the several points. This was sufficient for the topic.	Top effectively maintained interest with the phrase: **"If you're ready, say I'm ready!"** Top stopped using any techniques to maintain audience interest during the content phase. This may have been due to time constraints.	Top kept checking audience agreement with the points he was making by asking **"Yes?"** but this did not check audience understanding. The opportunity to assess and treat this was missed due to running out of time.	Top's presentation would have benefited if he had related the topic to the audience in their working context.	This was not done although I am certain Top would have concluded if he hadn't run out of time. He is an experience d and trained presenter.	Top ran out of time which indicates that either the content was too great, or that the content should have been more effectively summarised throughout.	Top basically started off very well and got worse. This is an unfair reflection in some ways because I think this was mostly caused by the tight time limit. I am in no doubt that Top would have maintained the same professionalism throughout if he had not been caught out by the time limit.

Song	The introduction was basically just stating the topic. This could have been used to excite the audience about what is a relatively exciting topic.	No rhetorical questioning was used.	No other use of questioning was used.	This was not used but could have been used to highlight the key points.	Song's intonation was much like his usual spoken style. His presentation would have benefited from a 'performer' personality.	This was not used but could have been effective in getting the key points across.	Little interest shown. No observable techniques used.	This was not done.	This was not done. Song's presentation would have benefited if he had related the topic to the audience in their working context.	This was not done. A missed opportunity to repeat the key points, personalise, or check understanding	Song overran the time limit significantly. Content could have been better summarised.	Song's ability to present the SHEC Communication seems limited by his level of English fluency. Song could definitely benefit from utilising some of the techniques listed here.

Categor-ising	Understanding	Questioning	Emphasis	Audience Understanding	Time	
Comments and action by criterion	Introductions are very important and would be a great focus for a workshop. All presenters could benefit from some training on hooks and the need to plan these beforehand.	All presenters could benefit from some training on rhetorical questions and the need to plan these beforehand. I would also like to encourage the use of questioning throughout the presentations and at the end as a method of checking audience understanding.	All presenters need some work on this. This should be connected to the work I want to do on preparation of the key points – highlighting the key points on the SHEC Communication document. All participants could do with some focus on the identification of affordances for visual aids use, the variety of visual aids possible, common mistakes with visual aids, preparation of visual aids at the planning stage, and effective use of visual aids.	Training on this will require a change of mindset. A lot of students have a adopted the 'lecture' approach whereby the presenter only has to present the information and it is up to the audience to understand it or not. I would like to design a kind of workshop which incorporates skills practice but also encourages presenters to take on the responsibility of audience understanding.	This is not a skill which I think students require particular training on. It's a matter of practice – practice that they will receive while practising the other techniques listed here.	Overall I have identified some very useful areas of weakness which I can use to design future lesson content.

Conclusions, Reflections, and Future Directions

There are various aspects of my assessment instrument that I feel could still be improved. I approached this study with the ideal that my performance assessments would be criterion-referenced in order that I would be empowering my students to reach an actual, measurable standard of competence. In practice, I soon realised that due to the uniqueness of the task, my diagnosis of gaps in student competence would have to be based only on my own conception of a realistic target competence for my students because no external standard exists. This also meant that evaluations were norm-referenced in the sense that I was judging students' performances based on my notion of what they are capable of, which is "inappropriate for formative assessment because it legitimates the notion of a standards baseline which is subject to existential determination" (Sadler, 1989, p. 127). To counteract the norm-referenced orientation of assessing students against my own concept of a reasonable target standard of competence, I would have ideally preferred to include at least one more assessor to increase the objectivity of the generated data and achieve triangulation, as Allwright and Bailey (2004, p. 73) advised, "at least two perspectives are necessary if an accurate picture of a particular phenomenon is to be obtained." Unfortunately this was not possible in this instance. An additional weakness of the data generation was that starting the evaluations with a list of predetermined criteria meant that I was not receptive to aspects of students' strengths and weaknesses which were not included on the list. Ordinarily this would not be desirable for assessing student presentations but it may have been useful for diagnostic purposes.

Despite the criticisms mentioned above, there are aspects of the data generation that I am content with. I feel that the specificity of the criteria, basing the initial assessments on descriptors, and the repeated viewings of videoed presentations meant a thorough diagnosis of the gaps in each student's competence. I also feel that the thematic chart approach meant that more descriptive data was collected which led to more effective analysis and interpretation, and more specific diagnosis. Also, utilising the thematic chart during the ultimate stage of the data generation addressed concerns about norm-referencing to some extent, because the data became a great deal more descriptive and therefore more transparent. Comments, even if they are somewhat subjective, by nature provide the reader or analyst with more information than grades or band scores.

The most important conclusion I have drawn from this study is that teachers can work independently to diagnose their students' needs before tackling the task of addressing those needs. A thorough diagnosis increases the likelihood that the teacher can meet the students' specific requirements. I wanted to ensure that this study was informed by a basis of established research, and conducted in a manner which was as objective as possible. I conducted this research in a pragmatic manner, in essence just tackling each stage in order with very little ability to foresee the subsequent stage. Of utmost significance is the fact that I take data and conclusions away from this research that I will use to begin an action research project into using formative assessment to improve my students' proficiency on performance assessments. The areas of weakness identified here, will dictate the focus of future lessons and projects.

References

Allwright, D., & Bailey, K. M. (2004). *Focus on the language classroom: An introduction to classroom research for language teachers*. Cambridge: Cambridge University Press.

Biehler, R. F., & Snowman, J. (1997). *Psychology applied to teaching* (8th ed.). Boston: Houghton Mifflin Harcourt.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74. doi: 10.1080/0969595980050102

Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly, 43*, 393-415.

Huhta, A. (2008). Diagnostic and formative assessment. In B. Spolsky & F. M. Hult (Eds.), *The Handbook of Educational Linguistics* (pp. 469-482). Malden: Blackwell.

McCullen, C. (1997). Presentation rubric Retrieved 13 January, 2012, from http://www.ncsu.edu/midlink/rub.pres.html

NCTE/IRA. (2004). Oral presentation rubric Retrieved 13 January, 2012, from http://www.readwritethink.org/files/resources/lesson_images/lesson416/OralRubric.pdf

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119-144. doi: 10.1007/bf00117714

Simkins, M. (1999). Designing great rubrics Retrieved 13 January, 2012, from http://www.registereastconn.org/sblceastconn/greatrubrics.pdf

Spencer, L., Ritchie, J., & O'Connor, W. (2003). Analysis: Practices, principles and processes. In J. Ritchie & J. Lewis (Eds.), *Qualitative research practice: A guide for social science students and researchers* (pp. 199-218). London: Sage.

Swinton, L. (2012). Oral presentation rubric: How to get great presentation grades Retrieved 13 January, 2012, from http://www.mftrou.com/support-files/oral-presentation-rubric.pdf