ORAL PRESENTATION ASSESSMENT OF SIMULATION PORTFOLIOS

Amparo García Carbonell
agarciac@upvnet.upv.es

Frances Watts
fwatts@upvnet.upv.es

Begoña Montero
bmontero@idm.upv.es

Universidad Politécnica Valencia

Arcadio Gotor
Universitat de València

Beverly Rising
brising@chs.upcomillas.es

Universidad Pontificia Comillas

Abstract:

One of the greatest problems found in using simulation in education is assessment. In the field of language learning, oral skills assessment is particularly arduous because of facilitator subjectivity in rating. The lack of standardized criteria for assessing speaking is another obstacle. A debriefing activity after participating in an international telematic simulation (IDEELS) is the oral presentation by the students on how their portfolios are taking shape. These portfolios contain information from every phase of the simulation by reflecting on their preparation, lectures and discussions, and learning outcomes. This paper presents our first attempt at ascertaining the reliability of our assessment of oral presentations of simulation portfolios.

The assessment criteria used were taken from Matthews and Marino (1990) and included the broad categories of delivery, content, organization and language, which were each broken down into five specific criteria. Our efforts in oral assessment have first been directed at attaining reliable results in the application of these criteria. The effect of the rating criteria in the assessment of oral presentations was measured using four independent judges in rating students of English as a Foreign Language.

The coefficients found through analysis of variance for the overall categories showed good levels of reliability in content, organization and language. Reliability dropped in the assessment of delivery. Further examination of the individual criteria bore varying results.
**Introduction**

The use of portfolios as a means of assessing student progress after participating in a telematic simulation activity ties in very well with current thinking in the field of language acquisition, in which the emphasis is on the critical and effective assessment of performance. Adams and Hamm (1998: 30) describe the following advantages of a portfolio:

Portfolios represent the cutting edge of a more authentic and meaningful assessment process. They are a powerful performance assessment tool that requires students to select, collect, and reflect on what they are creating and accomplishing. Selecting the academic work that represents their thinking and interest provides another opportunity for students to use critical thinking skills. As students think about evidence that they collected and decide what it means, they are participating in a meaningful critical thinking experience.

Students present a dossier containing an assortment of information that documents experience in terms of learning and achievements. The portfolio displays how the students perform in the language and also offers built-in reliability of measurement due to the variety of the evidence, as the authors have argued elsewhere (in Geurts et al. 1998). An additional occasion for assessing student progress is the oral presentation of the contents of the portfolio. This paper presents our first attempt at ascertaining the reliability of our assessment of oral presentations of simulation portfolios.

Students of English as a foreign language from the Telecommunications Engineering School of the Universidad Politécnica de Valencia have participated in telematic simulation since the academic year 1992-93. The first experience was with Project IDEALS, followed by participation in Project ICONS and finally as a partner institution in Project IDEELS, funded under the Socrates program of the European Union. The partners are educators from the Universities of Bremen (Germany), Nice (France), Bergen and Nord Trondelag (Norway) and the Universidad Politécnica de Valencia (Spain). The project team so far has tested three different scenarios using Eutropia as the backdrop that mirrors Europe's rich cultural diversity. Although student participants come from many areas of study, the students in Valencia all study English as a Foreign Language.

The foreign language used in simulations is a means to the end of communicating. At the same time, the students discover and master patterns in language as discourse, patterns beyond the sentence, which is the boundary of grammar teaching. They are expected to enlarge their passive and active linguistic repertoire. They increase passive knowledge through the reading of varied types of texts from a diversity of topics related to the simulation, and expand their active skills through the use of new lexical items and structures that they come across in their readings, in their written assignments and in their oral interventions in class.
The students write a portfolio summarizing their experience of participation in the simulation. The portfolio contains information from every phase of the process by reflecting on their preparation, lectures and discussions, and learning outcomes. The emphasis on learning by doing, the total learning philosophy based on experiential learning as a circular process (Kolb, 1984, Joplin, 1995, Cowan, 1996, Pfeiffer, 1995), is fundamental in simulation (Lederman 1992). Therefore, making the portfolio is a logical follow-up to a learning experience involving simulation (García-Carbonell, Watts & Rising in Geurts, 1998: 333) and is part of the simulation debriefing process. An additional debriefing activity is the oral presentation by the students on how their portfolios are taking shape. The assessment of the oral presentation activity is the main focus of this paper.

During the simulation the students have had practice in the skills of small group interaction. As students are guided through the process of interacting effectively, they are able to develop communicative strategies that carry over into real-life scenarios. Presenting the portfolio in class orally helps the student to develop communicative skills that will be useful tools in their future professional lives. As professionals, they may be called on to explain research, discuss company policies, analyze problems, offer recommendations or give on-the-job instructions. In addition, they may be requested to give oral presentations to accompany more formal written reports. There is an obvious need for them to acquire communication skills as in the near future they may find themselves speaking at an informal gathering, giving a formal presentation at a conference or talking to experts or non-experts (Matthews and Marino, 1990:1).

The oral presentation of the simulation portfolios before their classmates is the last step in the simulation process. In the presentation, the students use their own criteria to select the items which they consider most relevant to their learning experience, which produces a comprehensive profile of their learning (García-Carbonell, Watts & Rising 1998: 335).

Students consider the oral presentations one of the high points of their English classes, as testimonies in the portfolios have repeatedly shown. Apparently, students find it enlightening to be able to witness how classmates say that they are dealing with the struggle (because for many learners of English it is a struggle) to harness all the information that has been offered thus far and mold it into something coherent to a reader. The oral presentations help both the speakers and the listeners to clarify their ideas. At the same time, they all find it refreshing to monitor their classmates’ progress so that they can monitor their own to some extent.

Students have to plan, organize, develop and deliver the presentation by following clear, step-by-step guidelines given at the beginning of the course and discussed in class. During the simulation process functions such as expressing facts and opinions, describing situations and processes, comparing and contrasting, expressing feelings, speculating, arguing for and against, agreeing and disagreeing or introducing and concluding have been practiced. Thus,
forms and processes of oral discourse have also been studied. However, students concentrate sometimes so hard on putting across the content of their presentation in a way that their audience can easily understand that they forget to achieve linguistic accuracy.

Apart from the communications skills mentioned above, the language objectives to be achieved are:
1. English phonology practice
2. Intensive oral comprehension and expression practice
3. Fostering of the students’ confidence in the use of oral English.

Assessing is a complex process that involves not only giving a reaction to the first impression but also means seeing the production of the student in a much wider context in which a whole range of different aspects have to be considered. We are interested in making sure that our assessment of the student’s oral performance is fair and we also want to make sure that the challenge and skills required in the task place neither too great nor too little demand on the learner, so as to keep the student motivated. Many students see their product –the oral presentation of their portfolio- as the result of much deliberation, many drafts, many corrections. In short, behind each presentation there is much hard work, and therefore students are particularly sensitive towards what they may perceive as negative response on the part of their peers or their teachers. Our response to the oral presentation must be delivered with the greatest respect for the efforts made by the students.

The rating process
In assessing oral skills most instructors favor a general impression mark since this allows the performance to be judged as a whole. However, there is a need to fine-tune the assessment of student performance in the oral presentations of their portfolios. Several categories can be borne in mind. We have chosen to follow Matthews and Marino’s (1990: 204-206) criteria for rating presentations, which include the categories of delivery, content, organization and language.

According to these authors, assessing delivery should contemplate volume, rate of speech, eye contact, natural delivery and speaker mannerisms. The volume and speed at which a speaker presents are essential for the listeners to understand clearly and easily. Eye contact with the audience insures keeping the listeners’ attention. Speakers can also see whether or not their message is being followed. Natural delivery, i.e., not reading or reciting from memory, further guarantees adapting to the audience; a presentation given naturally will always be more successful. The speaker’s posture or body language also contributes to the listeners’ paying attention to the message, and not focusing distractedly on the speaker.
The criteria for assessing the content of the presentations include the central idea, the suitability of the topic for the audience and time available, argument support and meeting time requirements. The presentation must center on a clear central idea, which must be stated to the listeners. The topic should not be too technical or too well-known to the audience and should be developed with relevant details, facts and examples that provide strong support of the central idea. In the portfolio presentations, the raters who are facilitators make an assessment of the creative work of the students as well as their skills at making explicit connections between their creative work and the simulation content.

In assessing the organization, raters evaluate the ability of the student to organize information and ideas coherently and to use clear language markers to indicate the various stages of the presentation. There should be a clear introduction and clear transition from stage to stage. The main points should be clearly stated. The development of ideas should be logical and easy to follow. The conclusion should be clear and consistent with the presentation.

The criteria for assessing language include evaluating the accuracy of communication through the clarity of ideas presented by the speaker. The vocabulary used should be appropriate for the audience in that it should not be too difficult or contain technical words that are not explained. Pronunciation should be clear and intonation should be correct. The speaker should use a variety of expression and avoid inappropriate jargon or informal register. Finally, there should be effective control of grammar and vocabulary to achieve precision in communication.

While assessment criteria or scales are developed to obtain systematic descriptions and appraisals of performances, it has been noted that they do often fail in this respect (Wolf 1990: 496). The failure may be due to a lack of a uniform standard or reference, but many problems can probably be attributed to the variable of the individual rater. One of the most frequent objections to oral testing is subjectivity in rating and potential unreliability. Studies have shown that ratings given by even competent, independent raters can vary. What is important for one rater may be less important for another. Bachman (1990:180) remarks that raters may differ in their interpretation of a rating scale. For example, if raters are asked to rate on a scale of one to ten, one rater's nine may be another's five. In borderline case where performance falls at the intersection of two levels of performance, raters may be forced into “either/or” judgements. One rater will be lenient, another not. McNamara (1996: 124-125) adds that the discrete rating categories of a scale may appear to have equal intervals, but are rarely equal in practice.

Regarding rater behavior, certain specific patterns have been identified:

- The generosity error or tendency to rate individuals at a higher level than is warranted
- The central tendency error or the inclination to rate in the middle of the scale and avoid its extremes
- The halo effect or the tendency for a rater’s general impression of a ratee to influence the ratings of specific traits
- The logical error or the tendency to give similar ratings on traits that the rater thinks may go together but may not (Wolf 1990: 496).

Rodriguez Neira et al. (1993: 72) assert that the competence of the rater is crucial to the quality of the results. However, there are other factors such as fatigue, haste, mood and fading sensitivity which all impair rater reliability.

Notwithstanding all the potential pitfalls in oral assessment when considering rating scales and rater behavior, we still assume that a task may be considered effective if it stimulates learning. In our case, the oral presentation is an activity that motivates the students in their study of the language because after intensive involvement in the simulation process, finally being able to present their experiences to the whole group is a challenge with positive and surprising outcomes. The oral presentations increase the students’ self-confidence as they gain experience in speaking before a group which is good preparation for future professionals who will have to communicative effectively in English. Despite the pitfalls, in the end, assessment feeds into course redesign and assists in the subsequent implementation of our simulations.

The benefits for the students seem clear but the oral presentations pose a problem for the facilitator-teacher who has to award a numerical grade. Assessment of written work has been the subject of much research. It is now understood that the consistent application of well-defined grading criteria produces reliable marking (common sense but it has been scientifically proven). The criteria for assessing writing competence has also been set in a scientific manner. But assessment of oral performance is an ongoing concern for researchers. It is a much more crucial matter for the teacher pressed to decide in an unrepeatable moment what criteria the student meets.

Our efforts in oral assessment have first been directed at attaining reliable results in the application of criteria. We have started conservatively, for there is already evidence that oral assessment can be reliable. However, in our specific context no previous experiments had been carried out.

**Materials and method**

The effect of the rating criteria in the assessment of oral presentations was measured using four independent judges in the rating of 18 students of English as a Foreign Language. Two of the judges were experienced teachers who were native speakers of Spanish; two were experienced native teachers of English who were also fluent in Spanish. The subjects were all students attending the School of Telecommunications Engineering at the Universidad Politécnica de Valencia.
The assessment criteria were taken from Matthews and Marino (1990) and included the broad categories of delivery, content, organization and language, which were each broken down into five specific criteria. So that the students know what is expected on them, they are provided with Matthew’s rating criteria in advance. They are, therefore, familiar with the specific areas that we concentrate on in our assessment. The criteria used were the following:

I. Delivery
   A. Volume – loud enough to be heard clearly
   B. Eye contact with audience
   C. Natural delivery – not read or memorized
   D. Rate of speech – not too fast or too slow
   E. Posture/body movements – no distracting mannerisms

II. Content
   A. Clear central idea
   B. Topic suitable for time available – not too limited or too general
   C. Topic suitable for this audience – not too technical or well-known
   D. Topic developed with relevant details, facts, examples that provide strong support of central idea
   E. Presentation meets time requirements – not too long or too short

III. Organization
   A. Introduction
   B. Use of transition
   C. Main points clearly stated
   D. Development of ideas logical, easy to follow
   E. Conclusion

IV. Language
   A. Accuracy of communication – clarity of ideas
   B. Vocabulary appropriate for this audience – difficult or technical words explained
   C. Pronunciation/Intonation
   D. Use of a variety of expressions – no inappropriate jargon or informalities
   E. Precision in communication – effective control of grammar/vocabulary

The rating took place in a two-hour session in which each student was given five minutes to make his presentation. The raters followed the criteria as presented above by awarding 2 points in each category for excellent performance, one point for satisfactory performance and 0 to indicate the need for improvement. For the subsequent data analysis the 2-1-0 values were transformed to a five-point scale to allow for the half values that the raters actually gave in judging the subjects.

Results
The coefficients of reliability found through analysis of variance for the overall categories are shown in the table below.

<table>
<thead>
<tr>
<th></th>
<th>Delivery</th>
<th>Content</th>
<th>Organization</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.458</td>
<td>0.705</td>
<td>0.761</td>
<td>0.635</td>
</tr>
</tbody>
</table>
All of the findings are statistically significant, although the category Delivery presents a notably lower figure. In order to address the issue adequately, the raters’ behavior in the individual criteria was examined. The following table offers the reliability coefficients obtained through analysis of variance for the different criteria.

<table>
<thead>
<tr>
<th></th>
<th>Delivery</th>
<th>Content</th>
<th>Organization</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.364</td>
<td>0.369</td>
<td>0.527</td>
<td>0.357</td>
</tr>
<tr>
<td></td>
<td>NS</td>
<td>NS</td>
<td>p&lt;0.05</td>
<td>NS</td>
</tr>
<tr>
<td>B</td>
<td>0.25</td>
<td>0.464</td>
<td>0.59</td>
<td>0.716</td>
</tr>
<tr>
<td></td>
<td>p&lt;0.05</td>
<td>p&lt;0.02</td>
<td>p&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.528</td>
<td>0.463</td>
<td>0.483</td>
<td>0.454</td>
</tr>
<tr>
<td></td>
<td>p&lt;0.05</td>
<td>p&lt;0.05</td>
<td>p&lt;0.05</td>
<td>p&lt;0.05</td>
</tr>
<tr>
<td>D</td>
<td>0.638</td>
<td>0.5</td>
<td>0.687</td>
<td>0.431</td>
</tr>
<tr>
<td></td>
<td>p&lt;0.005</td>
<td>p&lt;0.05</td>
<td>p&lt;0.005</td>
<td>p&lt;0.05</td>
</tr>
<tr>
<td>E</td>
<td>0.588</td>
<td>0.834</td>
<td>0.556</td>
<td>0.389</td>
</tr>
<tr>
<td></td>
<td>p&lt;0.02</td>
<td>p&lt;0.001</td>
<td>p&lt;0.02</td>
<td>NS</td>
</tr>
</tbody>
</table>

The results referring to delivery show an acceptably moderate level of interjudge agreement in items C, D and E, whereas the results for criteria A and B are nonsignificant. With regard to presentation content, the coefficients shown in the table imply agreement among judges, which is especially high in criteria E. Criteria A, however, is nonsignificant. In reference to organization of the information in the presentation, results all show moderate interjudge agreement, although Criteria D is somewhat higher. Finally, in the category referring to language usage and accuracy, the coefficients vary according to the criteria. In criteria A and E, agreement is nonsignificant, which contrasts with criteria B in which agreement among judges is highly significant. The results for criteria C and D show moderate agreement.

**Discussion**

Considering the general categories, there was very high agreement among the raters when judging the organization of the students’ presentations and slightly less, but also high, agreement on judging content. On rating the language used by the students, raters still showed a good degree of agreement. However, agreement dipped in the category of Delivery, which is comprised by the seemingly objective criteria of voice volume, eye contact, not reading or reciting from memory, speed and posture.

When we examine rater behavior in the individual criteria, we see that agreement on appreciating voice volume and eye contact with the audience was not at all significant. This could be on account of varying degrees of rater hearing ability in the case of volume, cultural differences in rating what proper eye contact is or simply inexperience in rating performance in these areas; in any case, the first two criteria are subjective for some reason. In judging natural delivery, raters were able to discern adequately if the speeches were read or recited. In judging the rate or speed of speech, discernment was even better. In the last criterion in the category in
which the posture or body movements of the speaker are judged, raters also proved to discriminate well. These three criteria, therefore, can be considered objective and easy to follow.

In the category Content, the first criterion evaluates the existence of a clear central idea, on which raters did not agree, as the nonsignificant finding shows. Perhaps this owes to the fact that the subject of the student speeches was actually a summary of the different phases of a work process, instead of a topic of their choice. Theoretically, all of the speeches had the same central idea. The next three criteria all show only marginally significant reliability. The suitability of the topic for the audience and time available is apparently a vague aspect to rate and/or could be irrelevant for our purposes. Criterion D, in which supportive evidence presented in developing the topic is judged, appears to be a somewhat more objective criterion, as shown by the higher coefficient. And last, the criterion which rates if the presentation meets time requirements showed very high rater agreement, which seems a logical finding when something is objectively quantifiable.

When we look at the individual findings for the category of Organization, we observe that the introduction to the presentations seems to be an aspect that produces only moderate agreement in grading, which could be due to the fact that the introduction is given very quickly and the raters are, therefore, constrained by time. When judging the use of transition, raters showed higher consensus than in rating the introduction. It would be interesting to study if the raters were guided by organizational or content transition or if they were influenced by the use of language transition words. The next criterion, main points clearly stated, surprisingly only produced marginally significant agreement. The explanation may be the interference of organizational aspects with language aspects or the order of appearance of the criterion in its category, which may have distracted the raters. Criterion D, the development of ideas, proved to be accommodating to raters and can thus be considered objective. Finally, the conclusion to the speech also produced consensus among raters and can be determined objective.

All the criteria in the last category come under the heading of Language. The first aspect rated was accuracy of communication or clarity of ideas, on which the agreement was not statistically significant. Here the cause for the lack of reliability may have been interference between the content and language criteria. On the appropriateness of the vocabulary used, in contrast, the reliability coefficient is highly significant. This finding corroborates what has been scientifically proven regarding the objectivity of the assessment of vocabulary in language testing. Criterion C rates the pronunciation and intonation of the speaker; the findings showing only marginal significance, which concurs with Orr (2000) who states that non-native judges pay more attention to pronunciation and intonation than native ones. The following criterion which assesses the use of a variety of expressions shows only a tendency towards statistical significance. There could be some overlapping with the vocabulary that has already been rated.
Finally, precision in communication displayed through the effective control of grammar/vocabulary proved to be a criterion which raters interpreted subjectively, as the non-significant result shows. This finding could be a result of the fact that vocabulary had already been assessed by criterion B and suggested in criterion D. Or there may have been confusion between the first and last criteria with the terms accuracy in communication and precision in communication, the explanations after the dashes having been ineffectual.

Conclusion

It seems clear that teachers, at least the teachers in our study, are used to rating, and rate well, aspects of content, organization and language usage even if individual criteria are ambiguous. However, rating the delivery of a presentation requires more benchmarking in order to reach statistically significant agreement.

The categories we have used in our assessment are more suitable for assessing a set speech, such as the portfolio presentation, than an interactive speech. When assessing oral skills, it is one thing to assess a formal standardized oral presentation, a prepared speech, but quite another to assess “the subtle and dynamic oral exchanges which occur in simulations, role play and non-taught exercises” (Jones 1995: 126) which is beyond the scope of this research and will be the topic of another paper.

References


