AN INTRODUCTION TO PEER ASSESSMENT OF ORAL PRESENTATIONS

David James Townsend

Abstract: Presentations are an area of second or foreign language education that has great potential for transfer of skills from the classroom to real world applications. Presentation skills practiced in the classroom are the same skills needed to succeed in a wide range of business environments. They are essential for effective communication and valuable to students in both their native and second languages. One way to optimize the time spent in the classroom is to have all students assess their classmate’s presentations. Instead of being an idle audience, they are actively engaged in reviewing the presentation skills of their peers, focusing on those points that are most important for making a presentation successful. A valid concern of many teachers in allowing students to participate in the very important task of assessment is the reliability of their scores. This presentation will detail the findings of a one-year study of Japanese second-year university students assessing each other’s presentations. The results focus on the reliability of students’ assessments of their peers.

Introduction

Peer assessment is defined by Davies (2006) as a process that involves students grading and/or giving feedback on their peers’ work, and being judged for the quality of the appraisals they made. While I have had some experiences having students observe each others performances and provide feedback based on their observations, this would mark the first time that I would have students be responsible for assigning actual scores for other students’ performances, scores that would go towards their final course grade.

This paper will explain as concisely as possible my initial efforts to implement peer assessment. It should be noted that this was done prior to any substantial research being conducted, and with only some informal consultation with the other teachers overseeing similar classes. Then, I will detail some of the major findings of my research. Initially, I will detail the potential benefits of peer assessment and ways it might prove useful in the language classroom. Following this, I will detail the problems that other researchers have found implementing peer assessment in their classrooms. Finally, I will offer up some ideas of how my initial attempts could be improved upon in the future as well as areas for future study and/or analysis.

Procedure

The class, comprised of second year university English majors, focuses on oral discussion and presentation skills. It is in the department of practical English, and meets twice a week for a total of 32 90-minute classes a semester. It is a mandatory class taught by four different teachers, and as such is subject to some amount of standardization. It was explained to me that while I would be granted some measure of latitude, it was strongly desired that peer assessment be included in the mix.

While this class focuses on both discussion and presentation skills peer assessment was utilized exclusively in the presentation skills component. The students were required to perform three presentations on an assigned topic per semester, for a total of six presentations over the course of the academic year. Each presentation accounted for 15% towards each semester’s final grade. For each presentation, the cumulative score from their peers would amount to 30% of the final score for the presentation, with the remaining 70% coming from the teacher. The students were required to speak for approximately 3 to 5 minutes.

Initially, the students were introduced to the basic organization of a presentation which followed a structure that included an introduction, a body and a conclusion. Following this, they viewed a supplementary video containing two sample presentations with the first being an example of a well done presentation, and the second a poorly conducted presentation (see details in Harrington & LeBeau, 2009). After viewing the sample presentations, the students worked in small groups comparing the two videos with a focus on identifying what factors contributed to the first presentation being more successful than the second. From this the students came up with a core set of five basic presentation skills as follows; (1) posture, (2) eye contact, (3) voice, (4) gestures, and (5) easy to understand. These would form the basis for assessment.

Finally, over the course of three 90-minute lessons, the students received a series of instructional inputs where both good and poor examples of each presentation skill item to be assessed were introduced. This was followed by the students undergoing practical training with each skill through structured pair work activities.

Limited rater training was conducted before every set of presentations, with the teacher explaining what constituted a poor, fair or an excellent score. It should be noted at this time, that the primary goal of peer assessment in this class was to assist in the identification of strengths and weaknesses in the hopes this knowledge would help the students improve on their performance.
The students were randomly divided up into small groups of 5-6, and took turns presenting in front of their peers. The items to be assessed were listed, and scoring was done with a 10-point Likert scale, with 1 being the lowest score and 10 being the highest. The presentations were conducted in sets of three, and while they were underway, the teacher would continually rotate among them taking extensive notes and assigning a grade based on a holistic interpretation of each performance.

**Potential Benefits**

Perhaps the biggest benefit of peer assessment for students is its potential for positive ‘washback’. Hughes (2003) defines washback as the effects of testing on teaching and learning (p. 1), and as suggested, it can be either harmful or beneficial.

Negative washback can occur when the role of tests and assessment is principally to measure the outcomes of learning (often through formal externally-derived tests and examinations) in order to maintain standards or to compare student populations or educational institutions. Used in this way, tests can be seen as a means of monitoring and controlling teaching and learning. In such instances, there can be strong motivation for the teacher to conform their classroom instruction to match what will be tested, all too often at odds with the real language needs of the students.

While no-one would deny that outcomes are of a major importance in language programs, many language testers would argue that to focus only on the products of learning as measured by final tests is to downplay the important role that assessment can play in providing feedback to learners. Receiving a final score at the end of a course of instruction in the form of a simple letter or numerical grade does not provide the necessary information students require to make informed decisions that will influence their future learning strategies.

According to Brown & Hudson (1998), positive washback can occur when the tests measure the same types of materials and skills that are described in the objectives of the class and taught in the courses (p. 668). They argue that tests and other assessment procedures, in addition to measuring learning outcomes, can be integrated into the instructional process to provide diagnostic information that can be used to improve teaching and learning. Shohamy (1992) expands on this point by explaining that connecting testing with learning in the classroom makes assessment a vital part of the instructional process. Assessment is continuous, formative, and diagnostic, and leads to repair and improvement for both students and teachers (p. 15).

Clearly feedback is important in diagnostic and achievement testing; particularly in objectives based testing (Brown, 1990, 1996). Students can understand not only how they did on a particular test in general, but how they did on different aspects of the test.

**Potential Problems**

My initial hesitation with instituting peer assessment in my own classroom stemmed from a reticence to allow students a say in the most sacred of teacher tasks: assigning grades. Put simply, I did not think they were all capable of handling this responsibility. I am not alone in having such feelings. In his research on peer assessment, Saito (2003) found that many instructors in EFL/ESL classrooms hesitate to employ peer assessment because they believe student ratings will be unreliable.

Unreliability as it relates to testing can be described as the notion that equivalent test performances are accorded significantly different scores. For example, the same presentation may be given a very different score by different markers (or even the same marker on different occasions). It is perhaps worth making explicit something about the relationship between scorer reliability and test reliability. If the scoring of a test is not reliable, then the test results cannot be reliable either (Hughes, 2003: 43). When we give a test to someone it is ideal that their results will be similar if the test is re-administered on a different occasion.

There are ways in which we can measure the reliability of assessments. In the case of subjectively-rated tasks, such as samples of speaking or writing which are judged according to a rating scale, determining reliability usually takes the form of calculating an inter-rater reliability (IRR) coefficient which indicates the extent to which raters agree on a given individual’s performance.

Hughes (2003) explains that when scoring requires no judgment, and could in principle or in practice be carried out by a computer, the test is said to be objective. In such cases, like a multiple choice test, the scorer reliability coefficient would be 1. But when a degree of judgment is called for on the part of the scorer, as in the scoring of performance in an interview, perfect consistency is not to be expected. Such subjective tests will have inter-rater reliability coefficients of less than 1 (p. 43).

While the perfect reliability of objective tests is not obtainable in subjective tests, there are ways of making it sufficiently high for the results to be valuable. It is possible, for instance, to obtain scorer reliability coefficients of over 0.9 for the scoring of compositions.

While there is no generally agreed on measure of significance, Landis and Koch (1977) provide the following table as a useful guideline for interpreting $\kappa$ values.
Analysis and Reflections

The inter-rater reliability coefficient was calculated for the student assessed scores of all six groups of presentations. The Online Kappa Calculator (2008) was used to calculate \( \kappa \) – a chance-adjusted measure of agreement – for any number of cases, categories, or raters. This calculates the degree of agreement in classification over that which would be expected by chance and is scored as a number between 0 and 1. The \( \kappa \) values for the 6 groups of presentations are as follows:

<table>
<thead>
<tr>
<th>Presentation #</th>
<th>(IRR) Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.28</td>
</tr>
<tr>
<td>2</td>
<td>0.33</td>
</tr>
<tr>
<td>3</td>
<td>0.28</td>
</tr>
<tr>
<td>4</td>
<td>0.42</td>
</tr>
<tr>
<td>5</td>
<td>0.27</td>
</tr>
<tr>
<td>6</td>
<td>0.31</td>
</tr>
</tbody>
</table>

According to Landis and Koch’s guidelines, all the sets of presentations except for presentation 4 would be considered to have only fair agreement. This would suggest that the raters are not sufficiently agreeing on how to interpret the assessment scales. This would lead to the conclusion that additional rater training is needed for the students to be more reliable. It also shows that at least in the confines of this brief study, that repeated exposure to peer assessment coupled with the continued but minimal rater training had a limited effect on improving inter-rater reliability.

An additional factor that might be adversely affecting the inter-rater reliability coefficients in this study is the number of items on the rating scale. Unlike a more traditional 5-point scale, this study utilized a scale from 1 to 10. It stands to reason that allowing the students a greater range of options in rating performances would in turn lead to less agreement among them.

Attempts at improving rater-reliability through rater training have come up with some promising results. One study by Saito (2008) concluded that rater training improved inter-rater reliability by reducing the possibility of a misfit in the data. That is, after rater training it was less likely for one student to have a very different score from the rest of his or her peers (p. 574).

As noted earlier, the goal of rating peers on a number of distinct criteria is twofold. First, it draws explicit attention to the skills that the course content considers important. Second, it allows the students being assessed to know where they need to make improvements in the future. In regards to the latter, however, this can only be done if the information they receive through the assessment scores from their peers demonstrate their strong and weak points in a consistent manner. To this end, more substantial rater training would surely be advisable. As well, it must be considered if limiting the number of options on the assessment scale would improve inter-rater reliability.

Finally, while this first attempt at implementing peer assessment uncovered some areas were improvement is needed, it also highlighted some important benefits. Peer assessment provides a useful, engaging activity to undertake while the other students are presenting. Otherwise, they might be totally disengaged while student after student stands in front of them and presents on similar topics. Focusing on the explicit points being assessed underscores the importance of these skills for a successful presentation. As well, getting multiple sources of feedback on their presentations has the potential to provide the students with a better understanding of how they could improve going forward.
References


