Judges self-analysis of speech contest rating

Graham Robson[*]

ABSTRACT

This paper is one of a series of papers that looks at how teachers deal with the process of rating a speech contest held at Toyo University. To date there have been two speech contests. After the first contest analysis showed that teachers have different approaches to rating scores (Robson, 2007). After the second speech contest this paper, with data collected through a questionnaire of the raters, looked at five main areas of how teachers approach the rating process. The results showed that retraining had helped maintain consistency with raters, but that the rating process is complex and subject to many different factors.

Keyword: speech contest, measurement, rating approach

INTRODUCTION

The Regional Development Faculty has organized an English speech contest for its students for the past two years. This event, it is hoped, would give motivated students a chance to perform in front of an audience, leading ultimately to producing higher levels of motivation in students. The students that take part are chosen by teachers involved in the speech contest. The teachers feel such students would benefit from taking part in this type of language event.

Speech contest 2007

In the run up to the first speech contest that was carried out in 2007 judges who would assess the speech contest agreed on a number of criteria with which to assess the 30 speech contestants. These criteria were divided into criteria that would affect the performance, or delivery on the day of the speech contest, and those criteria that would have been practiced before the speech contest, the content. Then results from the speech contest were analysed through FACETS software (Linacre, 2007) to

[*] Faculty of Regional Development Studies, Toyo University, Japan

determine whether the rubric we had employed to assess the speech contest had been reliable, and what kind of bias the individual criteria and judges had produced using the rubric. The rubric used had six criteria, which were body language, eye contact, pronunciation and voice. These criteria were thought to be the most difficult to score because of performance anxiety and nerves on the day. There were also organization and interest, which it was thought would be easier to score as contestants had presumably practiced and been coached on polishing these criteria before the speech contest. Each of these criterion were assigned a score between one and five points, with a description that fitted each point being a different stage of ability in one of the sub constructs.

The original paper that analysed that speech contest (Robson, 2008), found some bias with the criterion and judges themselves. It also hypothesized the delivery criteria being scored more harshly than the content criteria. Part of the actual results from the Robson study pertaining to the criterion can be seen in table one. Interest, which should have been screened by the assigned helping teacher turned out to be the most difficult to judge at .51 logits, and voice, a delivery item, was the easiest at –0.07 logits. Incidentally logits are measurements that have been converted from an ordinal scale from using a rubric to an interval scale, Enabling comparison of all measurement criteria of judges and the rubric to be measured together through an algorithm conversion.

Next table two shows the bias among the raters for criteria for the speech contest. There were a total of nine significantly biased interactions (four judges and six criteria). Table two shows that at the top judge one was extremely harsh judging organization, compared to other criteria, with a score of 4.11 z-scores less than the

Table one - Rater measurement report for six criteria in the speech contest

Criteria	Severity(logits)	Error	In-fit(Mean square)
Interest	0.51	0.17	0.97
Body Language	0.36	0.17	0.88
Eye contact	0.03	0.15	0.89
Organization	- 0.07	0.16	1.05
Pronunciation	-0.16	0.16	1.15
Voice	-0.67	0.18	1.01
M	0.00	0.17	0.99
SD	0.38	0.00	0.09

expected value. At the bottom of the table judge six judged organization significantly more leniently than other criteria. On the whole, however, the speech contest results showed consistency among all judges within acceptable limits.

Table two - Rater bias towards speech contest criteria

Student	Judge	Observed	Expected			In-fit
	(rater)	Score	Score	Error	z-score	Mean Sq.
Organization	One	45	56.0	0.40	-4.11	0.5
Pronunciation	One	58	52.6	0.42	2.10	0.9
Eye contact	One	53	46.9	0.36	2.20	0.8
Interest	One	46	52.3	0.42	-2.64	0.8
Body Language	Three	36	42.7	0.45	-2.68	0.8
Organization	Three	60	53.5	0.40	2.47	0.7
Pronunciation	Five	56	62.8	0.40	-3.17	0.8
Voice	Six	53	58.0	0.44	-2.21	0.6
Organization	Six	69	61.3	0.50	2.92	0.8

The original hypothesis of differing scores for content and delivery criteria was not proved. There are a number of possible reasons why this hypothesis was not proven. One reason could have been because of the small n-size of 30 that may have affected the reliability and validity of the results. Also, the criteria themselves could have been picked incorrectly. Indeed, when compiling the original rubric there was a paucity of studies that related specifically to speech contest (monologues) in a second language. Therefore, the literature review that went into making up the individual parts of the rubric could have been flawed. Whatever the reason the scale in some way needed revision.

Along with the quantitative analysis of the 2007 speech contest, qualitative analysis was carried out through informal unstructured interviews with three of the judges who had shown bias in their results. Those results revealed that judges seemed to have different approaches to both using the rubric and the approaches they used to score the speech contestants.

Speech contest 2008

The following year the second speech contest took place. Before this speech contest teachers involved in the judging (the same as last year) analysed the results of the previous year, and set about reviewing the bias that they had been individually indicated through FACETS printouts. The author also organized a short training session with those teachers, so that they could learn about their recorded bias, as well as retrain using the rubric through watching and scoring of videos of the 2007 speech contest. Along with teachers tempering their harsh or lenient bias to build in higher levels of self-consistency when scoring, the rubric itself was updated to reflect views from all the teachers, most of whom claimed that the previous rubric had been too long. The number of criteria in the 2008 speech contest rubric was dropped to three, which were voice features, including pronunciation, pausing and volume; capturing interest, which looked for the content and enthusiasm of delivery; and body and eye control, which dealt with eye contact and gestures.

Teacher self-analysis

The aim of his paper and those to follow will focus on how individual teachers viewed their scoring during the speech contest and what approaches they employed in using the rubric for the speech contest. As previously said, there were only a few studies that have dealt with L2 speech contest, and naturally the number of studies that focus on how teachers dealt with rating in that situation is even less. One study by Upshur & Turner (1999) showed that depending on how teachers rate for success or failure could actually affect how the scores are given. Indeed the Robson (2008) study also concluded that the judges who looked for success, rather than failure, in criterion judged contestants more leniently than those judges looking for failure in the criterion. However, this study and the Upshur and Turner study are still not enough to throw serious light on how the judges rated in speech contest.

Unfortunately, the number of students who took part in the speech contest was only 20 in 2008, so FACETS could not be used to judge the student performance – 30 is the minimum statistically stable number of 30 observations for each criterion of using FACETS, (Linacre, 2007). Therefore, this paper, focused only on how teachers dealt with the task of judging the 2008 speech contest.

METHOD

The participants for this study were the six judges who were full-time and male inside the Regional Studies department at Toyo University. These judges rated the speech contest in groups of three in two different rooms, 10 students each. Three students from the each group of 10 were chosen to go through to the final. After the whole of the speech contest had finished the judges were asked to fill in the questionnaire in Appendix A as soon as possible, so that their memory of how they scored would still be fresh. The questionnaire itself was divided into five sections of statements, including how raters used the descriptors that appeared in the rubrics; how they timed their scoring for each contestant; what they felt about their own strictness or leniency of the rating; how they weighted different types of criterion; and their overall approach to scoring the speech contestants. Each one of these statements required the rater to choose between four levels of agreement from disagree strongly to agree strongly. The original version given to the judges had more space than the Appendix A version to write down comments for each question.

RESULTS

Owing to the small n-size of six raters statistical analysis would be flawed, so the individual scores for each question were put into the chart in Appendix B. This results section will address answers that were given as a numerical choice from Appendix A, and will also include comments that were added for each question by the teachers (judges).

Use of descriptors of rubric

Most of the teachers referred to the rubric when making their choice of scores, except judge number one. When teachers used the rubric, teachers did not generally apportion an equal time to looking at each criterion, apart from judge six. Judge six listened holistically, and was, it seemed, aware of all the criteria at once. Of those judges who apportioned unequal time for the scoring, judge two said that content took a very conscious effort to evaluate, and judge three claimed that performance in one area of the rubric might have had repercussions for other criterion. For example, weak voice features could have impacted negatively on the interest score. In this way, it would have taken less time to evaluate the next criteria, after the weak first criteria. Statement three was basically the opposite of statement two. It came as no surprise that all judges except, again, judge six, agreed with this statement. Question four saw some split in ideas over which criteria were judged in which particular order. Judges one, two and four used an order for judging, all three judges choosing voice features the first to judge. The other judges, however, did not rate one criterion first over another. The last statement in this section checked whether judges had made their own mental scale of one to five points for the criterion, instead of using the written description from the rubric. All except judge two followed the rubric when deciding the difference between, say, two or three points for a criterion. Judge two claimed that the rubric itself was a little inadequate. Indeed, elsewhere he wrote that the rubric did not match the reality of the speeches in some places.

Timing of rating

When asked about when judges had made their decision for points on the rubric, judges one, two, four and six had made their decisions before the end of each speech. Judge two had explained this was the case when confronted by a student who was giving a weak performance. It can be said that there is little that can happen to bring the speech back into a more positive frame of light for the judges, once the performance turns bad. In seems that judge three, however, was changing his scores until the end. Maybe he thought that there was a possibility that some part of the score may change before the speech had finished. Statement seven was the opposite of statement six. The only point to mention here was judge three's claim that even if he waited until the end to score, he was also scoring as the speech went on, but would make a final decision at the end. That is why he has disagreed for these two opposing statements, six and seven. Finally, all judges stated that they had not judged the first few entrants differently from other entrants of the speech contest. The judges wrote their reason for disagreeing with statement as having already received training on how to score meant that such bias had been avoided. Also, another judge wrote that the first few entrants tend to be a benchmark for scoring, so they would not have been scored differently from others.

Strictness / leniency of rating

All judges believed that they were neither too strict nor lenient when judging the speech contest. Judge four, it seems, tried to make sure this did not happen as he did not want to appear to be too different from the scoring of other judges. Judge three had a high expectation for the speeches, so thought that high scores were possible, and he also tried to maintain consistency with himself, without worrying about how other judges had scored.

Weighting of criterion

This section dealt with how judges may have been influenced by one type of criterion more than another. The first statement, number 11, was asking about whether voice could have been the most influential criterion. All judges, apart from judge six, did not consider voice features to be the most influential criterion. Judge four, however, did concede that pronunciation among voice features is the first thing that judges hear, and this can sometimes have an immediate impact. Judge three further suggested that only in cases of severe pronunciations problems might this become an issue. Judges were divided in opinion regarding the influence of speech interest. Judges one, two and three agreed this was most influential when decision making. Judge three reiterated the point that interest is not independent by itself, but can cover body language, and voice features that may enhance the interest of a speech. Judges four, five and six did not see interest as the most influential criterion. Finally, as for body language, only judge three was most influenced by body language during the speech. Judge two saw body language as the "cosmetic icing on the case", as reference to a recognition of other criteria first, over body language. Judge three claimed that this criterion was the weakest as displayed by the students in the speech contest, but he claimed that good body language features are "tell-tale" signs of a well-executed speech.

Approach to scoring

This last section of the questionnaire addressed two approaches that judges might have had for scoring the students. The first statement asked if teachers had an image of a perfect speech, and had taken points from this perfect score as errors happened. None of the judges used this approach. Teacher four said that he did not take points off from a perfect score per se, but did need to readjust scores in cases where students relied too heavily on notes. As for the last statement, all teachers agreed that they had no particular scoring approach, but looked for example of good speech behaviour. Judge three further added that he looked for, not only good speech behaviour, but also bad speech behaviour.

CONCLUSION

From the results it is clear that most of the judges in the speech contest used the rubric for scoring, but did not generally apportion an equal time to looking at each criterion. This result is slightly different from results in the Robson study interviews. At that time one judge said that he did not really look at the individual descriptions of the categories because he said that there was a lot to look at. It may well have changed this year because the rubric itself was shortened to three criteria. It was easier to look at criteria on the rubric. Some judges claimed it took a very conscious effort to evaluate interest, so this criterion would need to be listened to, and evaluated for longer. One consequence of this should be the rubric displays higher scores for interest / content because it is that much harder to judge. Further, weak performance in one criterion may have impacted negatively on other criteria. It may be that judges cannot separate so easily different criteria in their mind when scoring. It also could be that a particular speech did not have an all-round good performance. Whereas content takes longer, voice features are judged first, and quicker than other criteria. The score weightings of voice features on the rubric may need to have less score attached. Body language, the other criterion may need to be researched further. In this study judges agreed that it was not the most influential criterion. In fact, the point was made in the original Robson study that body language, as a criterion may be inappropriate for the speech contest. Further research should focus on how students perceive body language, and what kinds of body language students are employing during the speech contest.

The timing of the scoring showed that nearly all judges made their decisions before the end of the speech. However, there is belief by one judge that part of the score for one criterion may change before the end. This shows that some judges need less time to make up their minds about scoring than other judges. It would be interesting for a future study to see at what point within the study that the judge gave the score for each criterion. This would throw some light on how much importance is given to criteria, as well as show how much time is realistically needed to judge criteria effectively, and whether this score is changed before the end of the speech. None of the judges felt that they had judged the first two speeches differently from the rest. This result is different from a result in the Robson study in which a judge admitted that he needed time to think about how he would proceed with the judging after he had calibrated the scores he had given for the first two entrants. The change in not judging the first two entrants differently this time, and the belief by all judges that they were neither too lenient or too strict, may have come about because of the training that helped the judges to be aware of the fact that there could be a difference. The necessity of some kind of retraining as advocated by (Kondo-Brown, 2002; Bonk & Ockey, 2003) may have improved the internal consistency of all judges scoring.

In the final category, approach to scoring, none of the judges claimed that they started with a perfect score and took point off where mistakes were displayed. This goes against results from the Robson study that showed one of the judges did actually use this approach. Again, training may have changed the scoring strategy of this judge. All judges, however, agreed that they had no particular scoring approach, but looked for example of good and bad speech behaviour. This is different from results in the original Robson study in which some judges scored an entrant based on the previous entrant. In that first study there was a clear dichotomy that appeared to show evaluation depended on an aversion to judge primarily for either success or failure may inhibit scoring. More work needs to be done to establish how much judges change their approaches or consistency from one speech contest to another and identify the factors, which influence their decisions.

This study tried to find out more information about what judges do when approaching the rating of the speech contest. Some of the answers were conclusive, but some seemed to conflict answers from the previous Robson study. It is clear that the data from six judges is not enough to make generalizations about judging in general, and that in-depth interviews with all judges next time may yield more information.

REFERENCES

Bonk, W.J. & Ockey, G.J. (2003). A many-facet Rasch analysis of the second language group discussion task. *Language Learning*, 20, 1, pp. 89-110.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19, 1, pp. 3-31.

Linacre, J.M. (2007). *A Users Guide to FACETS*, Chicago: Winsteps.com, p. 185.

Linacre, J. M. (2007). Facets Rasch measurement computer program. Chicago: Winsteps.com.

Robson, G. (2008). Applying rasch measurement to judged ratings from a speech contest at a Japanese University. *JACET Journal*, 47, p. 51-66.

Upshur, J.A. & Turner, C.E. (1999). Systematic effects in rating of second-language speaking ability: test method and learner discourse. *Language Testing*, 16, 1, pp. 82-111.

Appendix A

Thanks to your help last year and over the past while, I have been able to undertake some research into speech contests. Again, this year I am trying to continue some research into how we make speech contest judging decisions. While your actions are still fresh in your mind would you mind filling out the following (ring a score and where possible adding a comment in the box under) as honestly and thoroughly as possible.

THANK YOU

1 Disagree strongly 2 Disagree a little 3 Agree a little 4 Agree strongly

Use of descriptors of rubric

1. I often referred to the individual descriptions on the rubric to make my scoring decisions 1 2 3 4
2. I apportioned an equal time to each criteria as I listened to the speech 1 2 3 4
3. I did not really apportion a set time, but made my decision as I noticed certain speech features 1 2 3 4
4. I judged criteria in a particular order 1 2 3 4

(please indicate with numbers if necessary 1st 2nd 3rd voice____capture interest____ body and eye______

5. I formed my own image of what each number on rubric should conform to, referring little to rubric 1 2 3 4

Timing of rating

6. I made a decision on scoring before the speech had finished 1 2 3 4
7. I made my scoring decisions at the end of each speech 1 2 3 4
8. I think I judged the first few entrants differently from subsequent entrants 1 2 3 4

Strictness / leniency of rating

9. I felt I was not too lenient in scoring 1 2 3 4
10. I felt I was not too strict in scoring 1 2 3 4

Weighting of criterion

11. Of all criteria I was most influenced by voice features of the speech entrant 1 2 3 4
12. Of all criteria I was most influenced by my perceived interest of the speech 1 2 3 4
13. Of all criteria I was most influenced by body and eye control of entrant 1 2 3 4

Approach to scoring

14. I started with each student having a good score, then knocked points off for errors 1 2 3 4
15. I started with no score idea, and looked for examples of good speech behavior 1 2 3 4

Appendix B

Results of Self Analysis Survey of Raters

Question	Judges					
	1	2	3	4	5	6
Use of descriptors of rubric						
1	1	4	3	4	4	3
2	2	1	2	1	1	4
3	4	4	4	4	4	2
4	4	4	2	3	1	1
	V1,I2,B3	V1,B2,I3	/	V1	/	/
5	2	3	2	2	1	2
Timing of rating						
6	4	3	2	4	1	3
7	1	2	2	1	3	3
8	2	1	2	2	2	2
Strictness / leniency of rating						
9	3	4	4	3	4	3
10	3	4	3	3	3	3
Weighting of criterion						
11	2	2	2	2	2	3
12	3	3	3	2	2	1
13	2	1	3	2	2	2
Approach to scoring						
14	1	1	2	1	1	2
15	4	4	3	3	3	3

スピーチコンテスト採点方法に関する 審査員の自己分析

要旨

本論文は東洋大学で行われたスピーチコンテストにおける審査員の採点方法に関する調査の一部である。これまでに2回のスピーチコンテストが開催されている。第一回のコンテストの後の調査（Robson, 2007）では、採点方法は審査員の間で異なっていた。第2回のコンテストの後で、ここでは、審査員を対象に質問紙によるアンケートを行い、主要5項目における審査基準を調査した。評価方法の再訓練は審査員の採点の仕方に一貫性を持たせる一助となるが、その評価方法は複雑で、多くの異なった要因に影響されることが示された。

キーワード：スピーチコンテスト、測定、評価方法