Mr Romulo Guedez-Fernandez

The University of the West Indies, St Augustine Campus

Faculty of Humanities and Education
Department of Modern Languages & Linguistics
St. Augustine, Trinidad & Tobago, W.I.

Tel: (868) 662-2002 | extension: 84047
Mobile: (868) 351-2873

Email: Romulo.Guedez@sta.uwi.edu & rguedez@gmail.com

ASSESSING FOREIGN LANGUAGE INTERACTIVE SPEAKING PERFORMANCE (FLISP): THE CASE OF UNDERGRADUATE STUDENTS OF SPANISH

ABSTRACT

This study seeks to evaluate the consistency of the current rating scales used for the assessment of peer-to-peer Foreign Language Interactive Speaking Performance (FLISP) for the Spanish Language Programme at The University of the West Indies (UWI), St Augustine campus. The sample population consisted of an intact class of fifty seven first year Spanish Majors/Minors. The data collection instruments include students' tests scores and recorded speaking performances, questionnaires, semi-structured interviews, students' journals as well as a focus group session with raters. The analysis guiding the construction of the rating scale draws from the measurement-driven approach. The data collected allowed for the identification of drawbacks and gaps in the current rating scales from both the students' and raters' perspectives. The analysis of both quantitative and qualitative data suggested that there be changes to the band descriptors in order to tailor rating scales to the specific teaching context and course content/objectives. This investigation highlights the importance of students' comments on their performance during the test; as well as feedback and self-assessment as contributing elements to raising students' awareness of and engagement in their learning process. The findings of this study have implications for the reconstruction of an appropriate rating scale for the assessment of FLISP and therefore, the operationalization of the construct of interactional competence. Implications for pedagogy and classroom assessment are also addressed.

Keywords: classroom performance assessment, foreign language speaking rating scales, consistency, reliability.

1. INTRODUCTION

Over the last decade, the Spanish language programme at The UWI St Augustine campus has undergone several internal and external revisions. This has encouraged lecturers and instructors to improve the quality of teaching delivery and assessment in order to ensure that the programme provides students with a richer learning experience. The Spanish programme has moved from using a single (universal) rating scale for the assessment of Foreign Language Interactive Speaking Performance (FLISP) for all three levels of the programme to now incorporating the Common European Framework of Reference for Language, Teaching and Assessment (CEFR), as a guiding instrument for learning, teaching and assessment (Council of Europe, 2001).

As a consequence of the implementation of the CEFR, two distinct rating scales were incorporated per level. More recently, academic staff appreciated the need for ensuring more validity, reliability and fairness in the assessment of this interactive skill; therefore, individual rating scales were implemented per semester. Band descriptors for these rating scales are based on the specifications provided by the CEFR. However, in order to adapt band descriptors to match The UWI's teaching context, minor modifications to the wording of these descriptors were undertaken. After the implementation of the six newly revised rating scales, it was found that more specific descriptors for the assessment of FLISP should be formulated. The present study provides empirical evidence to support this claim.

This paper is based on a pilot study carried out with Level I undergraduate students and their instructors/raters of the Level I Spanish programme. The study seeks to evaluate the consistency of the current rating scales used for the assessment of FLISP for the undergraduate Spanish Language programme at The UWI, St Augustine campus. The Communicative Language Ability model which involves language knowledge as it is mediated in different language use contexts by strategic competence (Bachman, 1990, Bachman and Palmer, 1996), together with the CEFR, comprise the framework that informs this research.

1.1 Classroom peer-to-peer assessment

Most of the second/foreign language research on speaking assessment focused on large scale proficiency tests, while there has been less research focusing on language assessment in the classroom. In recent years, there has been an increasing volume of research on assessment involving peer-to-peer interaction in both large-scale and classroom settings (Taylor and Wigglesworth, 2009). Some studies have focused on the criteria used by raters while assessing candidates (May 2006), or on the construction of empirically-based rating scales to assess interaction of a peer-to-peer L2 beginner level Spanish performance (Ducasse, 2009), or essentially what raters focus on when rating paired candidate interaction (Ducasse and Brown, 2009). Other studies, however, have focused on the interaction of test-takers in oral proficiency tests (Brooks, 2009), or on the definition and operationalization of interactional competence in speaking tests through the analysis of candidate discourse combined with raters' notes (May 2009). In the above studies the focus is mainly on raters input on the designing and rating of assessment tasks.

Elder and McNamara (2002) investigated the impact of performance conditions on perceptions of task difficulty in a test of spoken language. They formulated the hypothesis that "posited differences in task complexity would be reflected in actual differences in task performance", however, this was not confirmed (358). They therefore suggest that test-takers be consulted at the early stages of test development. This recommendation guided this research, as test development involves the designing of the rating scale to be utilised.

In previous studies, there has been little attention paid to adult candidate's input on the process of generating new band descriptors. The present study differs from previous studies as it focuses on both raters' and test-takers' insights on the descriptors of the rating scale and the assessment of peer-to-peer FLISP in a classroom setting. This research is designed to explore from the raters' and students' perspectives the following research question: To what extent are the modified rating scale and descriptors consistent in catering for the speaking tests?

2. METHODOLOGY

A multi-method approach underpins this investigation which involved advanced-level foreign language learners of Spanish over a period of one semester. SPAN 1001 course content and objectives are aligned to the CEFR level B1. The two speaking tests administered accounted for ten-percent of students' final course grade per semester.

2.1 Participants

An intact class of fifty seven year-one Spanish Majors/Minors (female=50, male=7) participated in this research. The mean age of the participants was 22.3 years (SD=8.8). All of the students had been learning Spanish for more than six years. 80.6% of the students consider that there is a significant gap in conversational Spanish between the secondary school and university level and 82.9% identified overcoming emotion, fear, nervousness or embarrassment to speak as their main focus in conversation.

The four raters (female=1, male=3) who participated in this research were foreign language instructors and native speakers of Spanish. All were trained on the assessment of FLISP before the test. Two raters had previous experience on assessing FLISP while the other two were novice raters. The role of interlocutor-rater was assumed only by the instructor in charge of the respective class being assessed while the role of first rater/examiner was assumed only by the rater acting as assessor during the test.

2.2 Data collection procedures

The peer-to-peer interactive speaking Test 1 and Test 2 were administered during weeks six and eleven. A day after each speaking test, each student received an email with a shared link to the audio of their respective speaking test for them to access. Students were then asked to listen to their individual performance during the test, self-assess it and email their comments to the researcher.

Eighteen (female=17, male=1) students were interviewed by a Spanish level III undergraduate student (henceforth called the Research Assistant), after completing the second interactive speaking test. Interview questions were designed to tap into students' reactions on the conversation component and on their performance on the test. A semi-structured interview was conducted with students responding to it, some of the questions being posed as follows:

> How was your experience in the conversation component?
>
> How do you think you performed in the conversation test?
>
> How difficult has it been for you to understand the rating scale that is used to assess your oral performance during the exam?
>
> Is there anything, for instance: range, accuracy, fluency, interaction, or coherence that you think is not assessed properly?

The Research Assistant administered forty questionnaires (female=35, male=5) during the seventh week of the semester and subsequent to the first speaking test. This questionnaire made inquiries into similar issues as were asked in the interview.

What motivates you to participate in conversation class?
What prevents you from participating in conversation class?
What are your least/most challenging aspects in conversation class?
What type of activities in conversation do you find most/least effective and why?
Do you understand the rating scheme in the conversation component? Is there anything you would change?

The researcher and the research assistant discussed themes and coded the responses of the interviews; the questionnaires as well as examiners' deliberation after students completed their speaking tests. A focus group session with three raters was held.

The qualitative data consisted of students' journals and feedback after the test. Students were provided with the audio recording of their speaking test for them to comment on how they felt they performed during the test and these self-evaluations were also gathered for qualitative data analysis.

The quantitative data included: final scores from the two peer-to-peer interactive speaking tests administered to participants. Both of these tests were video-recorded for further analysis.

Raters provided detailed data consisting of marks assigned to every performance. The raters' marks for each performance were specified by rating criteria which allowed the researcher to examine inter-rater reliability.

2.3 Rating Scales

The process of generating new band descriptors could be conducted either using measurement-driven methods which draw from teachers' and examiners' expertise, or through performance data-based methods, which draw from data obtained from learners undertaking test tasks (Fulcher, 1996, 2012; Fulcher, Davidson and Kemp, 2011; Galaczi, 2010; Galaczi, ffrench, Hubbard and Green, 2011; Upshur and Turner, 1995). This study implements the measurement-driven method with a focus on the construction of rating scales for a criterion-referenced performance assessment. This approach provides support for the wording of a sound set of band descriptors for the various scoring criteria to be used in the assessment of this competence.

This study aimed to expand the existing analytic rating scale for the SPAN 1001 course by including more detailed band descriptors. The previous rating scale consisted of three bands (i.e., 0, 1 and 2 for the CEFR levels A2, B1 and B2, respectively) and the interactional construct was broken down into the five rating criteria: range, accuracy, fluency, interaction and coherence. The wording of the band descriptors was in Spanish (refer to Appendix 1). This previous rating scale was expanded and modified to seven bands, (i.e., 0, 1, 2, 3, 4, 5, 6 and 7 for the CEFR levels A2, A2+, B1 and B1+, respectively). English was selected as the language for the band descriptors in order to facilitate students' understanding of their content (refer to Appendix 2). This modification was implemented in order to ensure consistency and fairness to students, as well as to allow instructors to provide students with a more detailed feedback of their FLISP in the classroom and of their speaking tests. Descriptors were selected from the EAQUALS[1] bank of descriptors (2008) and from the CEFR. The same five criteria used by the CEFR rating scales were used: range, accuracy, fluency, interaction and coherence; as these criteria capture most of the students' development in their oral production in the classroom. Based on teachers experience and teaching context (measurement-driven methods), some CEFR and EAQUALS band descriptors were tailored to the course content and objectives. The performance data-based method was not implemented. The rating scale was used over the course of seven weeks as part

[1] Evaluation & Accreditation of Quality in Language Services (EAQUALS)

of students' self-assessment of their individual participation in conversation in the classroom (see Glover, 2011). The latter was intended to instruct students on the rating scale usage and understanding as well as for them to get acquainted with the wording of the band descriptors for the conversation component.

2.4 Speaking tasks

The two interactive speaking tests reported in the present study are comprised of two types of tasks: individual and interactive. The individual task of Test 1 requires one student at a time to describe a photo using simple sentences. For the individual task in Test 2, each candidate is given a photo and asked to create a brief story about the content of the photo, subsequent to which they are given two minutes for the individual task. The interactive task for both Test 1 and Test 2 is similar. It consists of a peer-to-peer conversation about the videos selected and previously watched/viewed by the candidates. The time period allotted for this task is five to seven minutes. These videos were previously recommended by the instructor or used as instructional listening material in the classroom. Students can watch these videos outside of class on their electronic devices and at their convenience.

2.5 Assessment procedures

The speaking test was administered by two raters (interlocutor and assessor). The interlocutor-rater conducted the test with minimal interaction with the candidates. The interlocutor-rater provided test instructions to candidates and guided them on the sequence of the test tasks. Students chose their partner for the test. The FLISP of the tests is assessed by two trained raters using the modified analytic rating scale. The first part of the test was guided by prompts (colour images). For this part, raters were asked to ignore the descriptors for the criterion of interaction. The second part of the test consisted of a conversation between test-takers and was guided by the interlocutor through the use of short questions; the second assessor did not intervene. The overall score awarded to the examinee/candidate's performance was the average of scores for task 1 and task 2. The test was also video recorded. Subsequently, the examiners/raters provided detailed feedback to test-takers, with said feedback focusing on the specific criteria of the rating scale.

2.6 Data analysis

This investigation was carried out using a mixed-method approach of data analysis. The final scores from Test 1 and 2 were analysed using descriptive statistics in order to estimate inter-rater reliability between the first and second rater. Detailed data derived from scores assigned by individual raters based on each criterion were analysed using Paired t-test analyses of the difference in the means of the scores. Open ended questions from the questionnaire and semi-structured interviews as well as students' journal entries and their comments on their performance in the speaking tests were qualitatively analysed by theme.

3. FINDINGS

Data from the first and second examiners, who provided their individual assessment for each criterion, (i.e., range, accuracy, fluency, interaction and coherence), was statically examined using correlation analyses in order to determine the inter-rater reliability coefficient for each test. Table 1 shows this coefficient for each test and respective criterion. The mean value for the inter-reliability coefficient for Test 1 was .940 and for Test 2 was .894. These results suggest a high level of consistency among raters.

Table 1

Test 1 and Test 2 Inter-rater reliability coefficients for first and second raters

Pair	Criteria	N	Inter-rater reliability	
			Test 1	Test 2
1	Range	57	.937	.891
2	Accuracy	57	.935	.877
3	Fluency	57	.919	.855
4	Interaction	57	.947	.916
5	Coherence	57	.963	.932

The raters' scores for the 114 students' Test 1 and Test 2 performances were analysed using the Paired t-test. This analysis was undertaken by pairing each criterion (range, accuracy, etc.) of the modified rating scale with the respective mark awarded for each performance by first and second raters. The null hypothesis tested whether there was a difference in the means of the score awarded for each criterion by the two raters. Results from the Paired t-test analyses of the difference in mean scores indicate that there was a failure to reject the null hypothesis; hence there was insufficient evidence to conclude that there was a significant difference in the means of scores awarded by both raters. In other words, these results indicate that there was a relatively high level of consistency in the marks awarded by each rater, as the difference in the mean scores for the respective criterion/category was not significant at α=.01, as illustrated in Tables 2 and 3.

Table 2

Results from the Paired t-test analysis for Test 1

Pair	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	99% CI of the Difference Lower	Upper			
Range	-.04982	1.57109	.20810	-.60472	.50507	-.239	56	.812
Accuracy	-.00018	1.61864	.21439	-.57186	.57151	.000	56	.999
Fluency	-.30088	1.84478	.24435	-.95243	.35068	-1.231	56	.223
Interaction	-.19982	1.60540	.21264	-.76683	.36718	-.940	56	.351
Coherence	-.45053	1.29772	.17189	-.90887	.00781	-2.621	56	.011

Note. CI = Confidence interval. Pair indicates difference in the means of marks awarded by rater 1 and rater 2 for the respective criteria, i.e. range, accuracy, etc. df = Degrees of freedom.

Table 3

Results from the Paired t-test analysis for Test 2

Pair	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	99% CI of the Difference Lower	Upper			
Range	.25018	1.45571	.19281	-.26397	.76432	1.297	56	.200
Accuracy	-.25070	1.73126	.22931	-.86216	.36076	-1.093	56	.279
Fluency	-.25053	2.04004	.27021	-.97105	.46999	-.927	56	.358
Interaction	.15088	1.36806	.18120	-.33231	.63406	.833	56	.409
Coherence	.15000	1.25687	.16648	-.29391	.59391	.901	56	.371

Note. CI = Confidence interval. Pair indicates difference in the means of marks awarded by rater 1 and rater 2 for the respective criteria, i.e. range, accuracy, etc. df = Degrees of freedom.

Figures 1 and 2 are graphs which have been arranged by criteria and test to present the distribution of the mean scores awarded by raters, and serve to illustrate the high level of agreement between raters. They also demonstrate the students' progress in/across all the criteria. As depicted, students' weakest area of performance was that of fluency followed by accuracy.

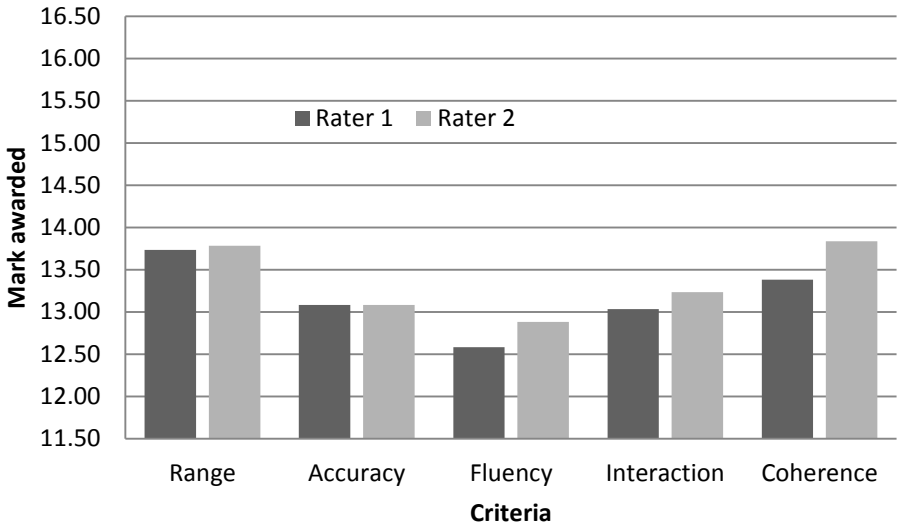

Figure 1. Overall students' performance in the conversation component in Test 1 has been broken down to show the criterion and mean mark awarded by each rater. The maximum score for each criterion is 20 marks.

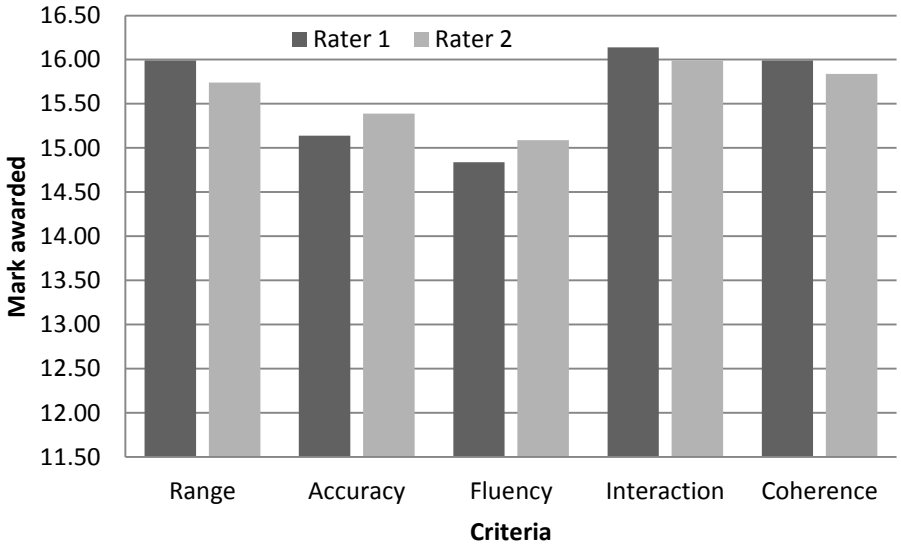

Figure 2. Overall students' performance in the conversation component in Test 2 has been broken down to show the criterion and mean mark awarded by each rater. The maximum score for each criterion is 20 marks.

4. DISCUSSION

In the present study the researcher investigated the consistency of an expanded rating scale designed to assess classroom context FLISP. Although, sociolinguistic and pragmatic competences play an important role in the assessment of FLISP some considerations on this ability are not feasible given the scope of the paper and its assigned length.

4.1 Inter-rater reliability

The inter-rater reliability coefficient for this study was relatively high, however, it can be observed from the quantitative analyses that the inter-rater reliability coefficient for Test 2 is slightly lower than for Test 1. In Test 2, the first task was more demanding than in Test 1. The mixed difference revealed by the inter-rater reliability coefficient cannot be attributed to the fact that raters were less lenient in the second test given that students obtained a higher score for all the criteria. On the other hand, raters preferred the modified rating scale to the former scale, as it allowed them to provide more specific feedback to students.

4.2 Consistency from the raters' perspective

Based on the raters' deliberation after students completed their speaking tests, together with comments obtained from the focus group session with raters, it can be observed that raters focused more on accuracy and fluency than on the other rating criteria. Raters' assessment for these two criteria was less lenient than for range, interaction and coherence. Descriptors for the rating criteria of range and coherence are seen by these raters as less defined/clear as it may lead to disagreement when determining the quantity and quality of vocabulary, expressions, linking words and content used by students in the test. This was already observed by Huhta et al. (2002) and Weir (2005) as they pointed out the CEFR does not provide examples of language, lexical resources or structures in its descriptors. It can be inferred that this gap allowed for inconsistencies in the interpretation of rating descriptors and therefore, generated distinct points of view among raters.

4.3 Consistency from the students' perspective

This section presents comments made by students after listening to their peer-to peer speaking tests and completing their self-assessment. The self-assessment exercise was intended to raise students' level of awareness of their performance in the test and therefore devise solutions that could help them to improve their command of this skill. Comments made by students indicated that this objective was met. Moreover, their comments confirmed findings from the quantitative analysis in terms of their agreement with the score assigned to their speaking test by the raters. These comments also indicated that students had a good level of understanding of and familiarisation with the rating scheme. Participants provided detailed insights into their self-assessment and evaluation of the rating scale:

> Yes, I think it [the rating scale] did capture my performance. I think that the rating scale for each category is very clear in terms of distinguishing the difference between what is classified as "1" compared to "7" and even includes "2", "4" and "6" which takes into consideration situations where a particular aspect might not strictly fit one specific description and might fall between two ratings. With this type of structure, I think a proper assessment can be made. Because I found it very detailed and explanatory, when I was listening to my audio, I was actually able to pinpoint a general range for where my speech could have fallen without any confusion. I think the rating scale is comprehensive

> and detailed which makes it really helpful in terms of knowing exactly what is being tested so that I can always know what my weakest areas are so that I can improve them which would then help with my overall competency in the language. (Participant 10, Test 2)

> Overall, I believe that I have improved in my conversational skills in comparison to the 1st conversation exam. My fluency has improved a little but it was affected by my lack of vocabulary in the language. Thus my range was also impacted due to lack of vocabulary (it was a little difficult to express my ideas.) I think that my accuracy was good but I still seem to have some minor issues with it. My interaction and coherence was a fairly good but there is still room for improvement. […] I graded as I did because I thought that the examination was pretty well done. My range was good as I thought that I was able to speak and describe stuff as I liked. Accuracy was also good. I thought that the interaction between my colleague and myself was great. I think that my fluency was great as well as my colleague and I was able to keep our conversation flowing without unnecessary pauses to search for words. Finally, coherence was also good. (Participant 44, Test 2)

It can be inferred from the previous students' reflections that the rating scale displayed an adequate level of consistency in capturing students' performance under the different rating criteria. However, Participant 22 considers that the rating scale does not provide a clear description for the assessment of spontaneity. This comment suggests an inconsistency in the rating scale as this participant sees this feature as key to the internalisation of the second/foreign language being learned at this level B1.

> Even though spontaneity is not a big issue, I think it should be included in the rating scale because that is a major part in being able to speak a language and knowing that that person has a good grasp of the language. (Participant 22, Test 2)

As a matter of fact, spontaneity is generally referred to in levels B2, C1 and C2 of the CEFR. However, a quick search in the CEFR reports this ability as only being assessed in one instance at the level B1, under the section of spoken interaction, 'interviewing and being interviewed': "Can use a prepared questionnaire to carry out a structured interview, with some spontaneous follow up questions" (p. 82). Why is this ability assessed here? The CEFR does not offer an explanation for it and other such issues of inconsistency in the CEFR do arise.

The modified rating scale fails to offer a proper description for interaction, naturalness and the use of nonverbal communication strategies. Participant 12 also views the non-inclusion of critical discussion of topics in the rating scale as part of the assessment of interaction.

> For the second part of the exam, I think I have considerably improved as I was familiar with the video we spoke about, and as a result, I was able to speak freely with sufficient vocabulary and linking phrases at a very good pace which demonstrated fluency and my control over the language. The interaction between Participant 16, Participant 35 and I was very natural, fluent and demonstrated coherence, which showed the extent of our language knowledge in this segment of the exam. The communication between us implored the use of non-verbal communication strategies and the ability to communicate accurately. Most importantly, I was able to approach this part of the exam fearlessly and confidently. (Participant 12, Test 2)

> I feel as though my interaction level is high however the situations given [in the modified scale] do not permit for critical discussion of topics. (Participant 12, Test 2)

One of the main criteria identified by Ducasse (2009) as crucial for successful communication in the beginner level of Spanish interaction was non-verbal interpersonal communication. Similarly, in the present study, non-verbal communication is seen by students as salient feature of interaction. However, in the CEFR, non-verbal cues in interaction are assessed at the level C2 (p. 28), the highest level of the CEFR. The user of the CEFR is left with little information: "Users of the Framework may wish to consider and where appropriate state: how skilled learners will need/be equipped/be required to be in matching actions to words and vice-versa" (p. 89). Students have raised the question: Should non-verbal communication be assessed at the level B1? Yes, it should be assessed because it is an essential feature of the interactional construct as Ducasse (2009) has demonstrated.

Other factors identified by students that have impacted on their performance include: task complexity, anxiety and prior knowledge. Elder and McNamara (2002) found that test complexity may not necessarily reflect students' actual performance; however, their study did not involve peer-to-peer speaking interaction. In the present study most of the participants reported task complexity as having impacted on their performance. The first task of Test 2 was more complex than the second task with participants reporting anxiety in this task and it having impacted on their actual performance. On the other hand, the second task seems to have restored their confidence and therefore they performed much better than in the first task. The following excerpts illustrate this impact of task complexity.

> I think that for the first part of the exam, I was very nervous and I made very simple errors that I would not normally make under different circumstances […] With respect to the second part of the exam, I spoke with much more fluidity and was more comfortable with the topic. Furthermore, I was able to relax and enjoy talking about the subject because of the prior knowledge that I had gained on the topic itself. I do recognize that I still made grammatical errors. (Participant 29, Test 1)

> I think that I did much better in this exam but there is still the great need to improve. […] The part with describing the picture was very difficult and I am always nervous so that can affect my mark a lot because I cannot function well when I am nervous. I plan to always improve so in the exams to come I can do very good and even get a mark in the 80's. (Participant 20, Test 2)

In short, this study has demonstrated some inconsistencies in the modified rating scale, which we argue could have been inherited from the CEFR. Both student's and raters' perspectives have contributed to the identification of gaps in the rating scale which will guide further wording of descriptors.

5. CONCLUSION

This investigation is a response to the need for a valid, reliable, fair and consistent instrument for the assessment of FLISP in the classroom within the context of the UWI St Augustine campus. Raters training on the use of the rating scale has shown a positive impact on raters' inter-reliability.

The newly reconstructed/developed rating scale displayed a satisfactory level of consistency, however, from both the raters' and students' perspectives; it was determined that this rating scale did not cater for some areas such as spontaneity, content and certain features of interaction, e. g., non-verbal communication. It has also shown the effect of self-assessment and assessment on learning and on student's progress in FLISP. Moreover, students demonstrated their familiarity

with the rating scale and the expected outcomes, and it is this knowledge of the rating scheme that has allowed them not only prepare for the test but furthermore, to identify their strengths and weaknesses in the respective areas. Detailed feedback on specific areas, which was provided by examiners after the test, allowed students to tackle those areas and progress in their speaking performance. Results suggest that both students' knowledge and use of the rating scales have contributed to: a) a greater awareness of their actual level of performance, b) engagement in their learning process, c) identifying their expected outcome, and d) motivating them to continue to develop their speaking skills. These findings have implications for best practice in teaching and classroom assessment.

6. Further Research

There is a need for the development of descriptors that define a more accurate interactive speaking construct at the various levels and both students' and raters' input will certainly enhance this developmental process. It is quite evident that the insight provided by conversational analysis in further analysing peer-to-peer FLISP can greatly contribute to this rating scale being more suitably tailored to the specific teaching context outlined in this study. The student's perspective has proven to be significant in the examination and reconstruction of rating scales as it offers an invaluable perspective which is often taken for granted.

References

Bachman, L. (1990) *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. and Palmer, A. (1996) *Language Testing in Practice*. Oxford: Oxford University Press.

Brooks L. (2009) Interacting in Pairs in a Test of Oral Proficiency: Co-constructing a better performance. *Language Testing*, 26 (3), 341-366.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching and assessment*. http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Ducasse, A.M. (2009) Raters as scale makers for an L2 Spanish speaking test: Using paired test discourse to develop a rating scale for communicative interaction. In Brown, A. and Hill, K. (eds.) *Tasks and Criteria in Performance Assessment: Proceedings of the 28th Annual Language Testing Research Colloquium. Language Testing and Evaluation Series, Vol. 13*, p.1-22. Peter Lang, Frankfurt.

Ducasse, A.M. and Brown, A. (2009) Assessing Paired Orals: Raters' Orientation to Interaction. *Language Testing*, 26 (3), 423-443.

EAQUALS. Bank of Descriptors. EAQUALS/ALTE Portfolio Descriptor Revision Project (2008) http://clients.squareeye.net/uploads/eaquals2011/EAQUALS_Bank_as_levels.pdf

Elder, C. and McNamara T. (2002) Estimating the difficulty of Oral Proficiency Tasks: What does the Test-Taker Have to Offer? *Language Testing*, 19 (4), 347-368.

Fulcher, G. (2012) Scoring performance tests. In Fulcher, G. and Davidson, F. (eds.) *The Routledge Handbook of Language Testing*, pp.378-392. Routledge: Abingdon, UK.

Fulcher, G (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, 13 (2), 208-238.

Fulcher, G., Davidson, F. and Kemp, J. (2011) Effective Rating Scale Development for Speaking tests: Performance Decision Trees. *Language Testing*, 28 (1), 5-9.

Galaczi, E.D. (2010) Paired Speaking Tests: An Approach Grounded in Theory and Practice. In Mader, J. and Ürkün, Z. (eds.) *Recent Approaches to Teaching and Assessing Speaking*. IATEFL TEA SIG Conference proceedings. Canterbury, UK: IATEFL Publications.

Galaczi, E.D., ffrench, A., Hubbard, C. and Green, A. (2011) Developing Assessment Scales for Large-scale Speaking Tests: A Multiple-method Approach, *Assessment in Education* 18, 3: 217-237.

Glover, P. (2011) Using the CEFR Level Descriptors to Raise University Students' Awareness of their Speaking Skills. *Language Awareness*, Vol. 20, no. 2, 121-133.

Huhta, A., Luoma, S., Oscarson, M., Sajavaara, K., Takala, S. and Teasdale, A. (2002) A Diagnostic Language Assessment System for Adult learners. In J. C. Alderson (Ed.), *Common European framework of reference for languages: Learning, teaching, assessment. Case studies* (pp. 130-145). Strasbourg: Council of Europe.

May, L. (2006) An Examination of Raters Orientations on a Paired Candidate Discussion Task through Stimulated Recall. *Melbourne Papers in Language Testing*, 11(1): 29-51.

Taylor, L. and Wigglesworth, G. (2009) Are Two Heads Better than One? Pair Work in L2 Assessment Contexts. *Language Testing*, 26 (3), 325-339.

Upshur, J. A. and Turner, C. E. (1995) Constructing rating scales for second language tests. *ELT Journal*, 49 (1): 3-12.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 22 (3), 281-300.

APPENDIX 1 — Previous Rating scale for SPAN 1001 - Spanish Language IA course

CRITERIO	0	1	2	TOTAL
ALCANCE	Utiliza oraciones básicas con expresiones, grupos de pocas palabras y fórmulas memorizadas con el fin de comunicar una información limitada.	Tiene un repertorio lingüístico lo bastante amplio como para desenvolverse y un vocabulario adecuado para expresarse, aunque un tanto dubitativamente y con circunloquios, sobre los temas del curso.	Tiene un nivel de lengua lo bastante amplio como para poder ofrecer descripciones claras y expresar puntos de vista sobre temas generales sin evidenciar la búsqueda de palabras y sabe utilizar oraciones complejas para conseguirlo.	
CORRECCIÓN	Utiliza algunas estructuras sencillas correctamente, pero todavía comete sistemáticamente errores básicos.	Utiliza con razonable corrección un repertorio de fórmulas y estructuras de uso habitual y asociadas a situaciones predecibles.	Demuestra un control gramatical relativamente alto. No comete errores que provoquen la incomprensión y corrige casi todas sus incorrecciones.	
FLUIDEZ	Se hace entender con expresiones muy breves, aunque resultan muy evidentes las pausas, las dudas iniciales y la reformulación.	Puede continuar hablando de forma comprensible, aunque sean evidentes sus pausas para realizar una planificación gramatical y léxica y una corrección, sobre todo en largos periodos de expresión libre.	Es capaz de producir fragmentos de discurso con un ritmo bastante uniforme; aunque puede dudar mientras busca estructuras o expresiones. Se observan pocas pausas largas.	
INTERACCIÓN	Sabe contestar preguntas y responder a afirmaciones sencillas. Sabe indicar cuándo comprende una conversación, pero apenas comprende lo suficiente para mantener una conversación por decisión propia.	Es capaz de iniciar, mantener y terminar conversaciones sencillas cara a cara sobre temas cotidianos de interés personal. Puede repetir parte de lo que alguien ha dicho para confirmar la comprensión mutua.	Inicia el discurso, toma su turno de palabra en el momento adecuado y finaliza una conversación cuando tiene que hacerlo, aunque puede que no lo haga siempre con elegancia. Colabora en debates que traten temas cotidianos confirmando su comprensión, invitando a los demás a participar, etc.	
COHERENCIA	Es capaz de enlazar grupos de palabras con conectores sencillos tales como «y», «pero» y «porque».	Es capaz de enlazar una serie de elementos breves, diferenciados y sencillos para formar una secuencia lineal de ideas relacionadas.	Utiliza un número limitado de mecanismos de cohesión para convertir sus frases en un discurso claro y coherente, aunque puede mostrar cierto «nerviosismo» si la intervención es larga.	
			TOTAL /10	

Descriptors for this rating scale were taken from: The Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR), (Council of Europe, 2001).

APPENDIX 2 — Modified rating scale for SPAN 1001 - Spanish Language IA course

SPAN 1001	0	A2 1	2	A2+ 3	4	B1 5	6	B1+ 7
Range		Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.		Knows enough vocabulary for familiar everyday situations and topics, but s/he needs to search for the words and sometimes must simplify what s/he says.		Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, and current events.		Has a sufficient range of language to describe unusual and predictable situations and to express her/his thoughts on abstract or cultural as well as everyday topics (such as music, films).
Accuracy		Uses some simple structures correctly, but still systematically makes basic mistakes. Can use correctly simple phrases s/he has learnt for specific situations, but s/he often makes basic mistakes – for example, mixing up tenses and forgetting to use the right endings.		Can generally communicate the main points of what s/he wants to say, though s/he sometimes has to simplify it. Can use some simple structures correctly in common everyday situations.		Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable and familiar situations. When s/he explains something, s/he can make the other person understand the points that are most important to her/him.		Can explain the main points relating to an idea, problem, or argument with reasonable precision. Can communicate with reasonable accuracy in familiar contexts, though with noticeable influences from her/his mother tongue.
Fluency		Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident. Can make her/himself understood with short, simple phrases, but s/he often needs to stop, try with different words – or repeat more clearly what s/he said.		Can participate in a longer conversation about familiar topics, but s/he often needs to stop and think or start again in a different way		Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.		Can express her/himself relatively easily when talking freely and keep the conversation going effectively without help, despite occasional pauses to plan and correct what s/he is saying.
Interaction		Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar topics and activities. Can handle very short social exchanges, even though s/he can't usually understand enough to keep the conversation going him/herself. Can ask and answer simple questions about likes, and dislikes.		Can ask and answer simple questions about things in the past. Can ask for and give opinions, agree and disagree, in a simple way. Can describe past activities, events and personal experiences. Can summarise simple stories s/he has read, relying on the language used in the story.		Can enter unprepared into conversation on topics that are familiar, of personal interest or pertinent to everyday life. Can start, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest. Can briefly give reasons and explanations for opinions and plans. Can narrate a story or relate the plot of a book/film and describe her/his reactions.		Can start a conversation on topics that are familiar or of personal interest and can help to keep it going by expressing and responding to suggestions, opinions, attitudes, advice, feelings, etc. Can develop an argument well enough to be followed without difficulty most of the time. Can compare and contrast alternatives, discuss what to do, where to go, etc.
Coherence		Can link groups of words with simple connectors like "and, "but" and "because". Can talk to people politely in short social exchanges, using everyday forms of greeting and address.		Can use the most important connecting words to tell a story (for example, "first", "then", "after", "later"). Can socialise simply but effectively using the simplest common expressions and routines.		Can link a series of short phrases into a connected, sequence of points. Can use simple expressions politely in a neutral way in everyday situations.		Can use connecting words to link sentences into a coherent sequence, though there may be some "jumps". Can use uncomplicated language to interact in a wide range of situations in a neutral way.
Total								Total /35

Descriptors for this rating scale were taken from: The Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR), (Council of Europe, 2001).
http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf EAQUALS Bank of Descriptors. EAQUALS/ALTE Portfolio Descriptor Revision Project (2008)
http://clients.squareeye.net/uploads/eaquals2011/EAQUALS_Bank_as_levels.pdf