Raters as scale makers for a L2 Spanish speaking test: Using paired test discourse to develop a rating scale for communicative interaction

Ana Maria Ducasse

This paper reports on the development of an evidence based rating scale to rate peer-peer L2 communicative interaction. The scale was based on experienced judges' comments on videoed student samples filmed during operational paired candidate tests of beginner level Spanish. Six trained and experienced raters generated criteria for communicative interaction which were incorporated into a tool for developing sample based rating scales, the Empirically-based, Binary-choice, Boundary-definition (EBB) method (Turner and Upshur, 1996), was adapted for the context. The findings reported on in this article examine the features of paired candidate interaction which raters used to define the boundary between performance levels. Three main criteria emerged as the boundaries used to define levels of interaction: non-verbal interpersonal communication, interactive listening and interactional management. These new notions are evidence of how peer-peer interaction can bee rated and also advance our understanding of the significant features of interaction in this rating context.

Introduction

Since the 1980s, paired and group orals tests have been increasingly common as a way of reflecting in testing the emphasis on Communicative Language Teaching in the classroom. Research into these new paired speaking tests originally concentrated on the effect on test scores caused by pairing candidates with different characteristics. Subsequently, the discourse produced in these paired tests was explored, and these studies have been followed more recently by rater verbal protocol studies to shed light on the process of rating pairs.

This paper focuses on rating criteria for paired speaking tasks, and more particularly how they are arrived at by scale makers. There has been a wide range of research into scale development in other contexts from various perspectives. Of particular relevance to this study is the use of student samples to derive empirically-based rating scales. Until now student samples have been used to develop criteria for writing, for monologic speaking tasks and for fluency scales, but not for paired tests involving peer-peer interaction samples.

This study was motivated by a practical need to comprehensively rate peer-peer interaction, in recognition of the fact that interaction among participants in a task plays a central role in generating discourse (Swain, 2001). If interaction is central, more research needs to be carried out into effectively incorporating it into rating scales. To identify the skills involved, the study looks at the point of intersection between the manner in which paired candidates manifest attributes of interaction and the way in which raters attend to those attributes.

The approach used is one that empirically derives scales by using teams of scale makers to define levels of performance by noting the salient differences between samples of paired L2 students performing a paired task in an oral test. Rating scales developed with teams of scale developers from student samples are not new. In a recent study Turner and Upshur (2002) use teams of raters to derive rating criteria from the same set of student samples.

Background to the study

Two strands of research provide the background to this study. One strand is on the development of rating scales, in particular data-based scales. The other strand concerns rating spoken interaction, in particular between peers.

Developing empirical rating scales

Rating scales usually mark out a series of levels, each of which is accompanied by descriptors that include characteristics of the performance expected at that level. The sample of candidate discourse used to assign a score is understood to derive from underlying language abilities or the construct being tested.

As reported in Turner and Upshur (2002), rating scales have been criticised for producing scores with low validity and reliability. Problems they cite involve:

the ordering of scale criteria may be inconsistent with the findings of second language acquisition (SLA)

criteria may be irrelevant to tasks and content

criteria may be incorrectly grouped at different levels

scales may lead to raters making false judgments because of relative wording

Improving the rating criteria could improve the problems with reliability listed above (Hamp-Lyons, 1991; North, 1995, 2003; North and Schneider, 1998). Scale development methods are basically divided in two types: intuitive and evidence based methods. Although the intuitive method is by far the most common way of arriving at rating scales using prior knowledge and consensus among experts, the evidence-based empirical method, which works *from* language output samples *towards* the descriptors, is the method chosen for this study. A rating scale based on what raters observe and notice during peer-peer interaction might address problems with reliability. It answers calls from the literature, such as that of Chalhoub-Deville (1997), who cautions that theory alone is insufficient to produce task specific scales, and Fulcher (2003), who directly calls for empirically developing rating scales.

The development of evidence-based scales for rating paired orals is further motivated by the fact that this format has been included comparatively recently into test batteries. There has been less time to research the peer-peer construct. It is difficult to gauge theoretically what features might be salient to raters in peer-peer interaction. It has been said that assessment that takes into account salient features of a task can improve measurement (Pollit and Hutchinson, 1987) but taking salient features into account can be difficult if such features have not been shown empirically to be salient from a rater perspective.

Rating paired orals

Different aspects of peer-peer interaction, in a group or in a pair, are interesting to testers. Features researched so far that have been empirically observed in paired discourse involve the number of functions produced (Lazaraton, 2002; Taylor, 2001) and conversation management skills (Dimitrova-Galaczi, 2004). These aspects have been qualitatively described and validated but have not been used as evidence to build data-based scales. There still remain, however, other unobserved, and until recently undescribed, features of interaction that make scoring interaction in groups or pairs difficult. Politt and Murray (1996) ask:

Should comprehension be assessed as part of oral proficiency?

Should a proficiency battery test language production or language interaction or both?

Should the oral test be one of communicative success or linguistic ability?

Comprehension, language production versus interaction, and communicative versus linguistic success are issues unexplored for the pair format from a *rater* perspective.

Of the studies carried out so far, a number have investigated the difficulty for scales and scale makers to adapt to the paired and group context: Nunn (2000) tackles the problem of designing rating scales for small group interaction during classroom activities as distinct from paired oral tests. The study acknowledges that for groups and rating scales "the considerable difficulties of reliability and validation need to be fully understood and the facile extrapolations about how students can perform in real life should be avoided" (Nunn 2000: 178). Nevertheless, despite the recognition of a difficult problem it is suggested that teachers recognise that "the question is not whether to do it but how to do it as fairly and efficiently as possible" (Nunn 2000: 178). The solution offered is to use the same scales for teaching, learning and assessment. How one develops these scales empirically still remains unresolved, regardless of the scope of the intended application.

In a more recent validity study on a university group oral test (Van Moere, 2006) the greatest variability was found in the person by occasion interaction: the people in a group are most likely to affect each others performance, which is expressed as " the more intangible interpersonal factors in the way group members react to each other" (Van Moere, 2006:436). The 'intangible' remains so far unexplained in the peer-peer testing context and these interpersonal factors need to be described and captured in a scale to reduce variability.

In contrast, Bonk and Ockey (2003) in a many facet Rasch analysis of a second language group oral discussion task found that "rater and scale reliability were achievable under real testing conditions even when the discourse was largely uncontrolled". We argue that rater training and scale relevance is the key to turning Van Moere's (2006) 'intangible interpersonal factors' which characterize paired oral communication into a "reliability…achievable under real testing conditions" (Bonk and Ockey, 2003). This can be achieved in two ways: by focusing empirical scale development on candidate output and by including features in scales that scale makers attend to while rating to facilitate rater training. These issues have been addressed in only a handful of studies so far, and these have focused on interviews not on peer-peer interaction.

Orr (2002) analyses verbal reports given by raters on the decision making process during the rating of the UCLES First Certificate of English (FCE). Thirty two raters completed verbal reports (Green, 1998) on two separate pairs of candidates performing the paired task from the FCE under test conditions. In that study Orr reports most compromising results. Raters were firstly found to apply different standards because they vary in severity, and secondly they were found to focus on rating criteria in different ways. (This has also reported been reported in Brown (2000) and Meiron (1998)) Lastly, raters were found to vary in the amount of non-criterion information they noticed for each candidate.

Included in the non-criterion information heeded while rating the paired interaction was the amount of non-verbal communication, for example eye contact and body language. The results have serious implications for the validity of the paired oral: the raters had varying perceptions of the performance but how the raters vary was not obvious in the scores. This makes it difficult to understand what FCE speaking test scores represent.

In a preliminary to the present study, raters of peer-peer interaction were also found to heed eye contact and non-verbal communication. The same verbal protocol methodology was used as in the Orr study. The participants comprised twelve language raters and severnteen pairs of beginner level candidates performing videoed paired tasks. The task consisted of maintaining a conversation for 10 minutes in response to three different topics. The topics were given on a card to each participant. The raters watched three different pairs on video and commented in English into a tape recorder about the peer-peer interaction. While observing the candidates' performance, raters commented on what made the interaction successful or not. A content analysis of the protocols suggested that the language experts oriented to three main features (*interactive listening, non verbal interpersonal communication* and *interactional management*) as salient features of interaction. Having identified features that language experts orient towards in interaction between peers, the next step was to build scales developed by raters in the role of scale makers, which is the focus of the present study. This fits with a longstanding call from the field for including in a scale "what raters attend to" (Politt and Murray, 1996).

In the light of the varying perceptions by scale makers and the varying severity that results in assessments, the difficulty of rating candidate pairs leads us to raise two very important questions. Firstly, we should consider whether the process of 'communicative interaction' (as it is called by Cambridge ESOL) is a construct that can be adequately operationalised in such a way that raters "understand the model of communicative ability on which rating scales are based" (Orr 2002: 153), and secondly, we need to investigate whether communicative interaction is scalable in the same manner that linguistic abilities have traditionally been scaled in band levels with accompanying descriptors.

Context of the study

The context of the study is a university Spanish foreign language program. As students are consistently taught within the framework of CLT, tasks that require pair and group work make up a high proportion of the available class time. Because of this, and because of the teaching focus on developing interactional skills, the decision was made to develop assessment tasks which reflected the tasks and the type of interactions students were accustomed to participate in, in the classroom. This resulted in a paired test task being developed and trialled. The development of a scale to assess these interactional skills was the next necessity.

Six trained and experienced teacher-assessors participated in the development of the scale for communicative interaction, basing their development on video-taped student samples filmed during operational paired candidate tests of beginner level classes. The candidates who participated in the study were beginner level students from the same year across two universities who volunteered to take part in exchange for an opportunity to watch their performance and obtain detailed specific feedback on it.

Methodology

The empirical method chosen for this study is known as the Empirically-based, Boundary Bound, Binary-choice (EBB) method (Turner and Upshur 1996). In this method, boundaries between levels are identified by requiring raters to focus on differences between levels.

The EBB scaling procedure

Upshur and Turner (1995) and follow up studies Turner and Upshur (1996) and Upshur and Turner (1999) describe a scaling procedure that "is *empirically* derived, requires *binary* choices by scale makers and defines the *boundaries* between score levels" 1999: 82). It leads to "a hierarchical sequence of attribute checks" (Turner and Upshur, 1996), requiring scale makers to make binary choices about the salient features of student performance. Upshur and Turner's scale development project was conducted in a French medium school in Montreal, Quebec and aimed to provide reliable assessments of ESL speaking ability. The scale development was based on a sample set of twelve performances on each of two tasks: a Story Retell and an Audio-Pal which involved a taped 'oral' letter. The participants were twelve teachers, as test developers and scale makers, and thirty-six grade 6 ESL students. .

It is important to bear in mind the points below taken from Turner and Upshur (1996: 61) for this type of scale development in which scale developers:

let actual performances tell what elements of the property space actually occur

do not assume what variables are important at different levels

let scale include only as many discriminable levels as raters can use reliably

assure that all levels are used

make explicit the procedures for constructing scales

ensure that scoring is efficient both in training time and rating time

incorporate knowledge and procedures followed by experts

Following the recommendations above, and using the EBB method of dividing the discourse sample into groups, scales were empirically derived. Turner and Upshur (2002:55) summarise the procedure as follows:

“A group of scale constructors, generally L2 teachers, is given a sample of writings or recorded oral performances. Working without a rating scale, the raters first arrive at a consensus on assignment of the sample performances into an identified number of levels and then identify and describe salient features that distinguish performance at adjacent levels”.

There are five tasks to develop the scale:

Task 1: Rank the candidate performances.

Task 2 Divide the sample into two groups: an upper level and a lower level.

Identify the most salient attribute of interest that divides the sample of collected data. Form a yes/no question about that attribute that divides the sample into those with or without that attribute. The question should refer to an observed difference that is relatively easy for teachers to recognise.

Task 3 Identify how many score levels the sample can be divided into.

Rank the upper level sample, *with* the salient feature, from task 2. Identify the most salient attribute of interest that divides the level. Divide the sample into two groups with or without that attribute. Form a yes/no question for that attribute.

Rank the lower level sample, *without* the salient feature, from task 2. Identify the most salient attribute of interest that divides the level. Divide the sample into those with or without that attribute. Form a yes/no question for that attribute.

Repeat until there are no more viable divisions.

Task 4 Set out the questions needed to sort the samples into score levels.

Task 5 Provide a score level description based on the salient features used to divide the sample into all the clusters, as identified in task 3 and set out in task 4.

In their subsequent research, Upshur and Turner (1999) identify three major concerns, which are addressed in this study by adapting the methodology. The first concern was that features that do not distinguish between different learner levels do not necessarily emerge. A second concern was that "when using empirical methods of scale construction the composition of construction teams and the make up of the samples of performances may have effects that deserve study" (Upshur and Turner, 1999: 107). Turner and Upshur addressed this issue themselves in their 2002 study for rating student writing, not speaking. Three teams of raters were provided with two samples of writing from which to build empirically derived scales. The researchers observed that the "scale development team had a minor effect on ratings" (2002: 65). Their final concern was whether these types of scale were task specific, described in Upshur and Turner (1999: 107) as the "tension between the need for accuracy in assessing a particular performance and the generalization to broader domains of language use".

In the case of the peer-peer interaction construct, as newly re-defined by the scale makers to include listening and non-verbal features, it would not be useful if the rating scale operationalising paired interaction only applied to the particular task performance. Performance on the task and the demonstrable skills that are rated based on the output need to be separable. The manner in which candidates can or cannot interact with a peer is deemed transferable to other peer-peer non test situations because interaction is a demonstrable skill: interaction is not *the task* in itself; it is a *result* of the task.

Scale development adaptations of EBB procedure

In order to make the scale more robust, the speaking samples were also very carefully chosen from the self-selected candidates in Phase 1 of the larger study to represent a range of performance types and candidate characteristics. The larger study has three phases as set out below:

Figure 1: Larger study

Phase 1	Phase 2	Phase 3
Define construct:	Devise scale:	Validate:
Content analysis of rater: 12 verbal protocols on 17 paired candidates	Phase 1 informs three teams of raters observing 8 paired candidates.	25 candidate verbal recalls validate construct and scale.

This paper reports on the team scale construction which is Phase 2. The construct findings from Phase 1 are transferred to Phase 2 to inform the scale development.

The EBB procedure as presented above was adapted to the context from the original in three ways. The EBB had been used by the researchers that developed the procedure to develop data-based scales for monologic spoken tasks (Turner and Upshur, 1996). The tasks used as input for those studies for this scale development procedure were of a different level of complexity when compared to the 10 minute Paired Oral Test used in this study. For this reason the EBB procedure was adapted in three ways:

1. The individual familiarization stage

This involved closely observing the 10 minute clips of peer-peer discourse and producing verbal protocols.

2. The provision of the reduced content analysis data

The data from the protocols were transcribed and analysed. The rater comments were reduced by the researcher and presented in tables on A4 sheets with a summary of the comments per pair of candidates made by three different raters.

3. Consensus moderation of the scales

There was a presentation by each team of their scale and a consensus as to which version to adopt for trial.

Adaptation 1: The individual familiarization stage

This adaptation was made in order to address the first concern raised by Upshur and Turner: features that did not distinguish levels not emerging as salient. In order to guard against it in the study reported on here, the 12 raters observed all the data alone before participating in scale development. This is an adaptation (1) of the EBB method. (All the adaptations are described in a section below.) The raters described all that they attended to and considered as scale makers to contribute to successful/unsuccessful peer-peer interaction. In the EBB procedure scale makers work alone first and *rank* performances. The focus then is ranking not describing what is noticeable about a performance.

In the adaptation, the scale makers spent two hours on their own, focusing specifically on interaction, prior to coming together with their colleagues for the scale development. In this time they provided verbal reports on the features of interaction they observed in the student performances presented on video. While they were evaluating the quality of the interaction they were not guided as to what features they should consider important enough for them to make comments on. As they were not focusing on distinguishing levels of performance, there was nothing to stop them commenting on aspects of performance which did not do this. In this way the first concern raised by Upshur and Turner was addressed before the scale was developed.

The pre-scale development task also presented raters with the opportunity to familiarize themselves with a sample of the range of performances that would ultimately be used in the scale development. This prepared them also to argue in support of their decisions because they had considered the issues at hand and the reasons for their ideas on interaction prior to the scale development workshop.

Adaptation 2: The provision of the reduced content analysis data

For the second adaptation, the content analysis data from the verbal protocols gathered in adaptation 1 were reduced and summarised by the researcher. This information was made available to all other scale makers before commencing the scale development procedure. It was presented in the format of an A4 table of data which was set out in columns per candidate pair, reduced from the content analysis of the verbal protocols of three different scale makers who had observed each candidate pair on video. The data reduction was placed onto separate laminated cards per pair of candidates with all the comments from the scale makers summarized and reduced onto a table with three columns of comments: one column for comments on the pair, and another for each of candidate *a* and candidate *b* (see Figure 2). This way all comments for the each of the pairs were visible at a glance.

The intention of this adaptation was to make other raters views manageable for scale makers to read and refer to during the scale making process. Apart from being a cohesive and experienced group of assessors in the context in question, more importantly, they would all see each other's opinions. The aim was to give the scale makers as much information as possible about each candidate before starting. The scale makers had already had three pairs each to comment on, with an overlap, so that each pair of candidates had comments recorded by three different scale makers. By providing this information, the issue regarding the effect of the specific scale makers on the scale was addressed.

Figure 2:

Pair 1 comments	on pair	left candidate	right candidate
By Rater X			

By Rater Y			
By Rater z			

Three main themes, each divided into two subcategories, had emerged from the findings of the content analysis of the verbal protocols made by twelve language teachers regarding salient features of successful paired interaction in Phase 1 of the study. The scale makers, a subset of the group of language teachers, were guided to discuss these three areas. The themes were subsequently used to guide the first step of the scale development. Raters were asked to discuss the importance/relevance of (a) interactional management (maintaining text cohesion by asking relevant questions or making relevant contributions or responses to the topic, and responding in turns fluently and evenly without excessively holding the floor), (b) interactive listening (being an engaged listener by using backchannel, and being mutually supportive as a listener in the interaction, e.g. by filling silences and gaps in language or by demonstrating comprehension), and (c) interpersonal non-verbal communication (using supportive gestures and maintaining eye contact).

Adaptation 3: Consensus moderation of the scales

The three teams developed separate but, as we shall see, similar scales. Through a process of discussion, consensus was reached as to which scale to trial.

Data collection

Participants in operational paired test videos

The preliminary stage of data collection involved selecting candidate performance samples on which to base the scale. In the study by Turner and Upshur (1996), in which they develop scales for monologic speaking, twelve individual performances are used. Taking into account that candidates perform together in this study, and that each sample comprises ten minutes of discourse, the number of performances was reduced to eight pairs, or sixteen individuals.

These eight pairs were chosen by the researcher from a total of seventeen pairs of candidates that had already taken part in Phase 1 of the study, referred to above, which explored raters' general orientation to interaction.

In Phase 1, where verbal protocols were elicited from the raters on successful interaction, it was evident from the transcriptions that some pairs attracted more comments. It was assumed that the greater the number of comments a pair had attracted, the more salient their performance had been to the raters. This was considered when selecting eight pairs for the study: four pairs with more comments on particular *individual* candidates and four pairs with more comments on the *pair* were selected. At first glance it appeared that candidates commented on individually were not interacting as well as pairs commented on as a pair. Also, of the eight pairs selected four were evenly matched for linguistic proficiency and four were not evenly matched. (The matching was based on a departmental 5-point rating scale from the candidates' end of year oral performance marked by trained departmental raters.) The performance had been scored for range and accuracy of grammar, vocabulary and pronunciation.

Participants in the rating scale workshop

Following Turner and Upshur, (1996: 61) who recommend between "four to eight members who are familiar with the aims of assessment" in their empirical rating scale development procedure, six scale makers participated in the scale devising workshop. The scale makers were experienced university Spanish language teachers, familiar with the task, the level and the rating context.

Scale development with the EBB procedure

The Turner and Upshur Method is a five step process which was followed to make the scale development process replicable. The scale makers followed a guide provided for them to ensure the scale development workshop followed the process step by step without the interference or influence of the researcher. What follows is the step by step process of developing a data-based rating scale based on a student sample. The five steps are expanded below with the attributes used to rank and divide the student sample in this study for the development of an evidence based scale for interaction using the EBB method.

Step 1: A single question for the top of the hierarchy

The aim of step one was to rank the performances and then to formulate a question. First, the teams watched and mentally ranked the performances. They did this by clicking on icons for the videos on a computer screen. (The raters had access to multiple computers to watch and compare performances of particular pairs as needed.)

Secondly, in their teams, the raters discussed which particular feature of successful interaction they observed in the performance which would enable them as raters to split the sample of candidate pairs into + or – a particular feature. (The + indicated a YES response to a question formed by the raters and the – indicated a NO response to a question formed by the raters.) The particular feature chosen was deemed to be the most salient attribute marking the boundary between two levels. The salient attribute was formed into a yes/no question. This question would be asked of every performance rated with the scale. The scale makers wrote their question into a text box marked Q1. The first question proposed by the three different pairs of scale maker teams was:

Figure 3: Question 1

Question 1 Are they supportive listeners?	Question 1 Are they mutually supportive visibly?	Question 1 Do they have supportive body language?
Rater team 1	Rater team 2	Rater team 3

As we can see in Figure 3, all three teams separately came to three questions which have the element of working together in common. These questions which mark a boundary between levels are known as criterial questions.

Finally to finish step 1 of the scale development, the rater teams had to reach an agreement on which performances belonged to the group for which the response to question one was YES and like wise for the group for which the response to their first question was NO.

Step 2: Questions for level 2 of the hierarchy

In step 2 the rater teams decided whether to work on the upper or the lower ranked part of each sample, i.e. the pairs grouped in the upper half with an answer YES to the question that divided the sample or the pairs grouped in the lower half with an answer NO to the question that divided the sample.

The scale makers ranked the performances in the section of the sample that they were working on. The scale makers wrote a question that divided the remaining performances in the sample then tested it against the candidates grouped at that level. The questions for level 2 of the hierarchy are shown below:

Figure 4: Question 2

Q1 answer Yes: Q2 Supports interaction	Q1 answer Yes: Q2 Supportive listener	Q1 answer Yes: Q2 Supportive listener?

with the body?	with back channel?	
Q1 answer No: Q2	Q1 answer No: Q2	Q1 answer No: Q2
Asks questions relevant to topic?	Asks adequate questions?	Asks relevant questions?
Rater team 1	Rater team 2	Rater team 3

The questions that follow on from a ‘yes’ answer on the first question from all three rater teams are either about listening or non-verbal support. Those that follow on from the ‘no’ answer all contained ‘question’ as an indicator of what would move the interaction on from this point to the next level. This seems to indicate that non-verbal or listening support, are of a higher order than asking questions, in scale makers’ orientation to successful interaction.

Step 3: Each cluster becomes a level.

The scale makers continue to rank and divide the pairs with questions that mark the boundary between levels. When the sample can no longer be divided the cluster becomes a level.

Step 4: Developing the EBB model

To conclude the session each team of raters completed an overhead of their EBB model. This involved writing up the questions they had used to divide up the sample. Each team presented their model on an overhead and defended it to the group. Two models were very similar. The three different EBB scale models developed in the scale development session are presented below:

Figure 5: Rater team 1

QUESTION 1 QUESTION 2 QUESTION 3 LEVEL

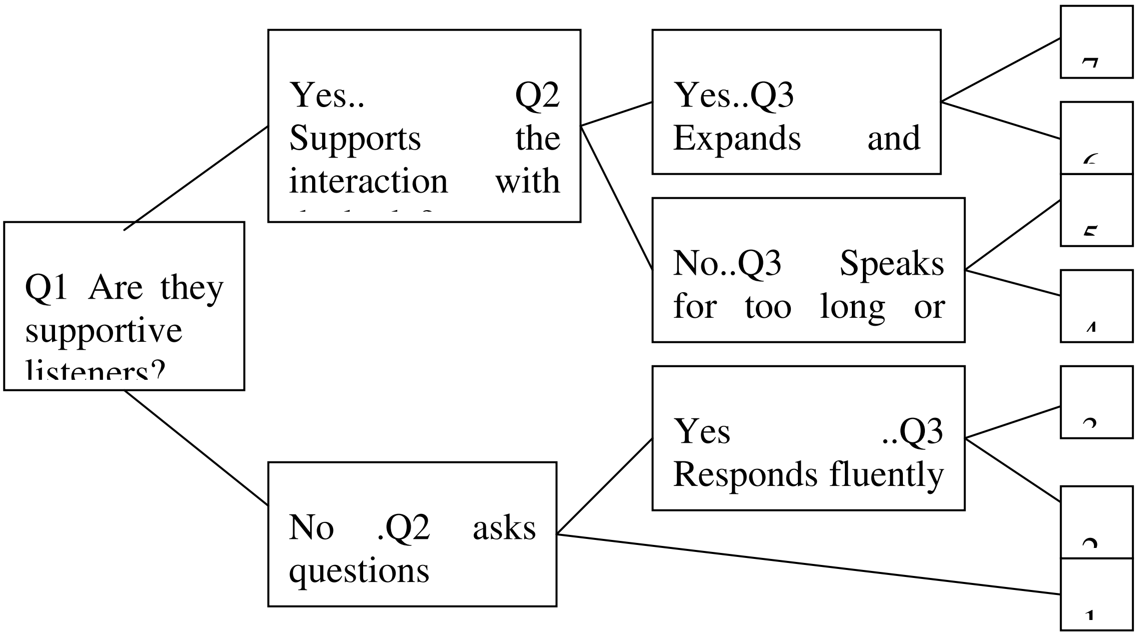

Q1 Are they supportive listeners?
Yes.. Q2 Supports the interaction with
No .Q2 asks questions
Yes..Q3 Expands and
No..Q3 Speaks for too long or
Yes ..Q3 Responds fluently

Figure 6: Rater team 2

QUESTION 1 QUESTION 2 QUESTION 3
LEVEL

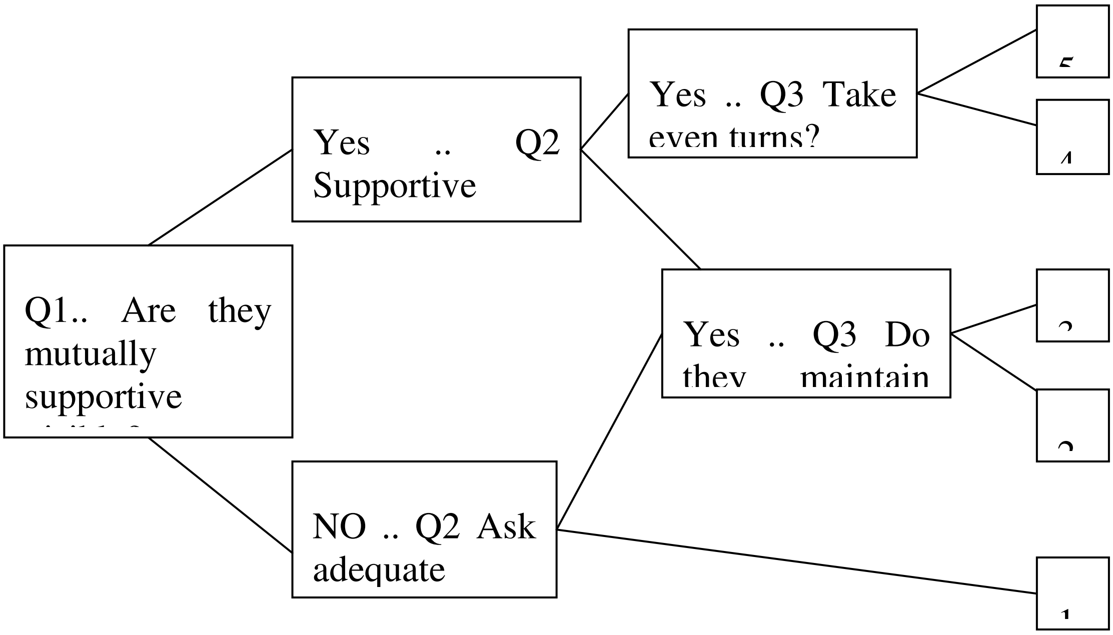
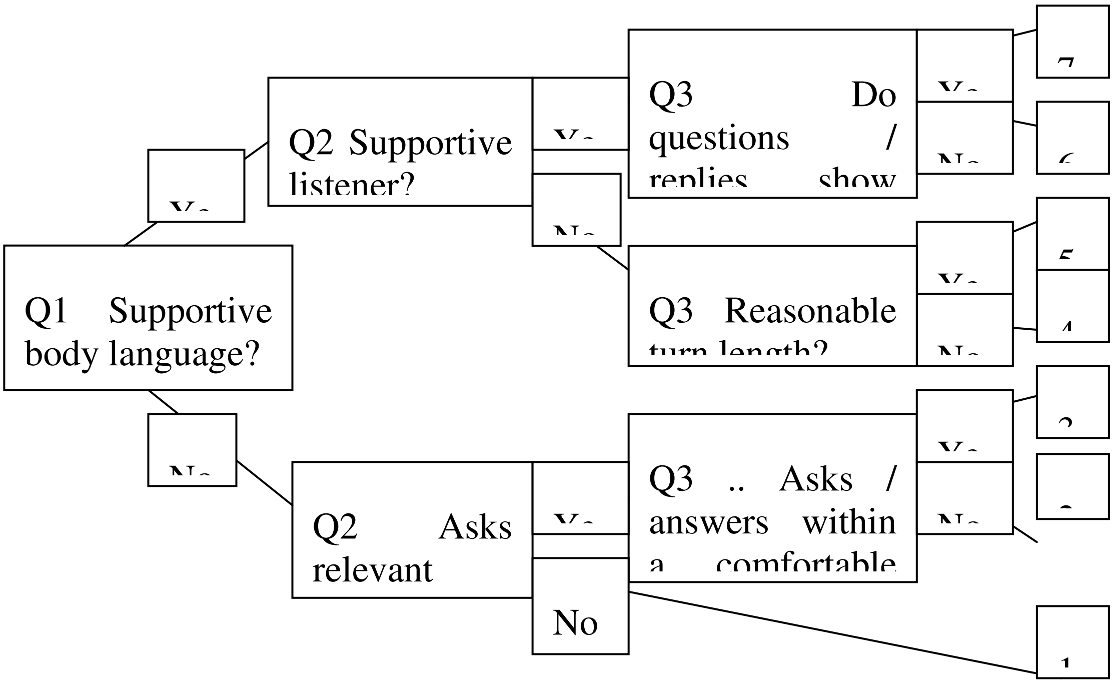

Q1.. Are they mutually supportive
Yes .. Q2 Supportive
NO .. Q2 Ask adequate
Yes .. Q3 Take even turns?
Yes .. Q3 Do they maintain

Figure 7: Rater team 3

QUESTION 1 QUESTION 2 QUESTION 3 LEVEL


Q1 Supportive body language?
Yes
No
Q2 Supportive listener?
Yes
No
Q3 Do questions / replies show
Yes
No
Q3 Reasonable turn length?
Yes
No
Q2 Asks relevant
Yes
No
Q3 .. Asks / answers within a comfortable
Yes
No
7
6
5
4
3
2
1

The EBB process had been adapted to develop scales in teams. Pairing the scale makers, in adaptation 3 above, resulted in three scales being completed by the end of the session. As the three figures show, the scales produced by rater teams 1 and 3 had seven levels, whereas that of rater team 2 only had five. What follows is the final step before arriving at a data-based scale.

It was believed that the final scale would be more robust if it was developed based on three scales made from the data that could be combined by the consensus of the scale developer/scale makers. Therefore, as part of the fourth step, each pair put up a transparency with their version of the scale for discussion and consensus.

By observing the similarities and differences over the three scales the scale makers in the study reported here reached a consensus as to which scale to trial. The chosen scale was 'tweaked' by way of consensus moderation by the scale makers before using it. The final version is below; note the differences in the point weighting for this scale to 5 points. Also different, is that the top row indicates the question order number to follow in the binary selection that channels the rater to a final rating.

The EBB starts with the first question: Supportive body language? 'yes' or 'no'? in which the 'visibility' of non-verbal communication is high in the hierarchy for a successful interaction in this rating tool for interaction. It means that for someone to successfully interact they need to look at the interlocutor and signal that they are listening for the interaction to be most successful.

It is followed by the choice of:

Yes Q2 supportive listener?

No Q2 Relevant questions/answers offered?

In the ‘yes’ Q2 and the ‘no’ ‘Q2’, listening and speaking are inseparable at both the higher and the lower level so the focus is on initiating and responding after non-verbal communication.

Figure 8: Final scale to trial

Question 1 →	Ans.	Question 2 →	Ans.	Question 3 →	Ans.	Rating
			Yes	Questions / replies mostly show cohesion b/n and within topics?	Yes	5
					No	4.5
Supportive body language?	Yes	Supportive listener?	No	Reasonable turn length?	Yes	4
					No	3.5
					Yes	3
	No	Relevant questions / answers are offered?	Yes	Asks / answers within a comfortable time?	No	2
			No			1

In the final step, Q3 determines the final mark by distinguishing the level of interactional management displayed in the performance.

Before leaving the session the scale makers discussed the importance of 'pairedness' in relation to the marks that could be attributed to performances. After their discussion they took the position that they would apply the boundary marking scale questions to the pair. While the response 'yes' or 'no' was held in common for both candidates, they would continue to ask one question of the pair. However, they would split the responses to yes or no for the individual candidates where candidates performance in interaction differed and one had a yes and the other a no for the response to a criterial question. In pairs where the difference lay in the first question they would be treated as individuals till all the questions were asked and the final mark was arrived at. In pairs where the answer to each of the three questions applied to the pair or both candidates then the same score would be awarded to both candidates in the pair. The scale makers therefore concluded that the scale would be able to offer individual marks or two identical marks depending on the performance of interaction.

Step 5: Writing a score level description

This last step involves writing a score level description to provide a picture of the trait being evaluated for score recipients such as other tutors, candidates, administrators or parents for example**.** Due to time constraints, the scale was compiled by the researcher, who produced descriptive statements from each of the three criterial questions. The statements were passed for comment to the course coordinator and to the teacher/scale makers for feedback. The proposed version of the scale (Figure 9) went forward to be trialled.

Figure 9: Trial scale

Level 5

Uses encouraging body language e.g. looks at speaker, smile, posture, hands head nodding

Is an audibly supportive listener, e.g. really? m mm, yes yes, shows interest while the other speaks

The moves within the interaction and the responses mostly show cohesion between and within topics

Level 4.5

Uses encouraging body language, e.g. looks at speaker, smile, posture, hands head nodding

Is an audibly supportive listener, e.g. really? m mm, yes yes, shows interest while the other speaks

The moves within the interaction and the responses do *not* always show cohesion between and within topics

Level 4

Uses encouraging body language, e.g. looks at speaker, smile, posture, hands head nodding

Is an audibly supportive listener, e.g. really? m mm, yes yes, shows interest while the other speaks

The turn length is balanced; it is neither too long or too short

Level 3.5

Uses encouraging body language, e.g. looks at speaker, smile, posture, hands head nodding

Is an audibly supportive listener, e.g. really? m mm, yes yes, shows interest while the other speaks

The turn length is *not* balanced it is either too long or too short

Level 3

Body language is *not* supportive; it tends towards visibly negative signals

Relevant questions and answers are given

Questions or answers are offered without too much hesitation

Level 2

Body language is *not* supportive; it tends towards visibly negative signals

Relevant questions and answers are given

Questions or answers are *not* offered without a lot of hesitation

Level 1

Body language is *not* supportive; it tends towards visibly negative signals

Relevant questions and answers are *not* given

Discussion

The findings of the study show that the scale maker teams focused on three interactional features. This is demonstrated by the hierarchy of Q1 through to Q3.

The feature at Q1, (*Supportive body language?*), that is most salient in dividing the candidates' paired performance first in the hierarchy, is the existence of outwardly visible signs of interaction: interpersonal non-verbal communication. This is the first area of focus in rating paired interaction.

At the Q2 level there are two options for the scale makers (Q1*Yes* > *Q2 supportive listener? and Q1No > Q2 Relevant Qs/answers offered?*). Here candidates who manifested signs of interactive listening are separated from those who failed to do so, which makes listening the second area of focus in rating paired interaction.

Finally at Q3 level, there are three pathways for the scale makers determined by the answer to each question (Q1*Yes* Q2*Yes* > questions /replies mostly show Cohesion between and within topics?; Q1*Yes* Q2 *No* > Reasonable Turn length? And Q1*No* Q2*No* > Asks/Answers within a comfortable time?) The element used to distinguish between levels is one of interactional management. At the highest level it is a question of cohesion, followed by turn length then fluency expressed in the time taken to respond. These three elements form part of interactional management which is the third and last area used to divide levels in rating paired interaction.

The scale makers' focus was on very fine details of peer-peer interaction in order to separate the last two levels. The salient details after non-verbal communication and interactive listening were features of the mutual support and signals of engagement between speakers which were demonstrated by observable interactional management skills.

Looking first at the lower end of the hierarchy, which leads to awards of between 1 and 3.5, at the lower Q2*No* there is an audible breakdown that could be hesitation, inefficient turn-taking, inappropriate response or initiation. However, if despite the communication problem Qs and answers are provided then 'yes' achieves a 3 on interaction if it is sustained sufficiently. If not then a 2 is awarded. If for Q1 there is *no* body language and there is *no* relevant initiation or response - just random offerings - for Q2, the rating is 1 for interaction.

An examination of the higher end of the hierarchy that leads to awards of 3.5, 4, 4.5 and 5, at the higher Q1*Yes* level the interlocutor is now audibly as well as visibly supportive, providing back-channelling, initiating and responding appropriately with ease. We have an engaged communicatively interactive interlocutor. If the answer is 'yes' to Q2 the candidate engages and contributes to the development of the discourse which moves to Q3 on cohesion, where the candidate is awarded 5 if cohesion between and within topics is sustained. However, if it is inconsistent, candidates are awarded 4.5.

If the answer is 'no' to Q2 there may be evidence of some discourse management problems, insufficient initiating or over length responses. In this instance, candidates are awarded 4 for observing turn taking conventions, but 3.5 if they are silent for too long or, conversely, speak for too long.

The problems in rating a paired oral test are caused by the interaction of many different factors such as listening, speaker engagement and non-verbal communication, which have been captured and represented on this scale. The results of the trial will demonstrate whether the EBB scale developed by the scale makers from candidate performances is sufficiently robust and sufficiently flexible to cope with variation in peer-peer interaction.

The most important findings of this evidence-based scale development are twofold: Firstly, the elements that were found to make up the construct in Phase 1 of the study mark separate levels in the scale. These are non-verbal communication, interactive listening and interactional management. Secondly, the order that the elements in the criterial questions are applied to tease out the levels suggests that the listening construct in mutually dependent interactive contexts needs to be explored. As was reported, while observing paired interaction the scale makers were aware of: the physical signals the partners emit, the listening and comprehension of the partners, and the reliance on each other's oral text cohesion and interactional management for the next thing they say. These all require further in depth exploration.

Most importantly, the findings call into question the effectiveness of other rating criteria for 'communicative interaction' and 'discourse management skills', at least as far as tests that include a collaborative discussion task are concerned. Raters may observe or attend to candidates in such tests manifesting non-verbal skills in peer-peer interaction or displaying skills in interactive listening. Hypothetically, subconscious orientation by raters in such contexts could inadvertently affect rating. If scale makers for the Spanish Beginner Paired Oral Test notice body language and how effective the listener is when rating pairs then possibly scale makers in other contexts may also attend to these factors, as was reported by Orr (2002) with regard to body language and eye contact.

The goal of empirically-based scale development is to improve the quality of the assessment by grounding it in student performance and the features that scale makers notice as being important to the performance. Unless there is a clear and common understanding of the construct, the rating system cannot work as it is intended to. In order to improve the validity of rating for this test discourse, where the pressures of university accountability are great, the staff adopted and supported the implementation of the new criteria by adopting the EBB scale for interaction, and using it in addition to already existing analytic scales for grammar, vocabulary and pronunciation.

In sum, the key point made above is that by including in a rating scale the features language experts orient to during peer-peer interaction, the result is that observable features of interaction are scaleable. This means that what was previously “intangible’ in interaction between peers has now been observed, described and placed on a functional scale.

Conclusion

This study was motivated by a practical need to comprehensively rate peer-peer interaction. The EBB scale developed by the scale makers, has built on Phase 1 of a larger study and on previous research on interactional management in peer-peer tasks. This was achieved by focusing on salient features of peer-peer interaction which included interpersonal non-verbal communication and interactive listening in addition to interactional management.

The scale development reported in this paper, based on a sample of paired candidate discourse, has avoided the problems with criteria encountered in other scale development methods that have low validity and reliability.

The criteria are relevant to the task and content

The criteria separate levels and group performances in clusters by moving from the large picture to the fine grain

The criteria do not included relative wording to differentiate between level boundaries

The findings show how scale makers developed a scale to incorporate what is salient to them. As a result the process has responded to Pollit and Murray’s (1996) questions:

Should comprehension be assessed as part of oral proficiency?

Yes, comprehension should be rated in a peer-peer paired task, because raters attend to candidates’ interactive listening skills.

Should a proficiency battery test language production or language interaction or both?

In a peer–peer task both production and interaction can now be tested and rated analytically.

Should the oral test be one of communicative success or linguistic ability?

In a peer-peer task communicative success can now be analytically rated separately from linguistic ability.

To conclude, the process of developing assessments of peer-peer interaction by 'defining the boundaries' with questions rather than describing the levels of interaction per se reveals the extent to which scale makers can determine what constitutes interaction based on student performance. In this study the raters operationalised the construct of 'interactive speaking and listening' in the construct, which included interactive listening during speaking, non-verbal interpersonal communication and demonstration of speaker engagement through interactional management. This has implications for the validity of oral assessment criteria currently being used. The lack of inclusion of these features in currently used scales could explain the differences in severity, inconsistency and the use of non-criteria observed by Orr 2002 or the 'intangible' in Van Moere 2006. It appears that the elements operationalised by the raters in this study do attract raters' attention, but that raters typically cannot find reference to them in the scales.

References

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

Brown, A. (2000) An Investigation Of The Rating Process In The IELTS Speaking Module, in Tulloh, R. (ed.) *IELTS Research Reports*, *Vol. 3*, ELICOS, Sydney.

Dimitrova-Galaczi, E. (2004) *Peer-peer interaction in a paired speaking test: the case of the First Certificate in English.* Unpublished PhD dissertation, Teachers College, Columbia University, New York.

Fulcher, G. (2003) *Testing second language speaking*. London and New York: Longman.

Green, A. (1998) Verbal protocol analysis in language testing research: A handbook. Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1991) Scoring procedures for ESL contexts. In Hamp-Lyons (ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Norwood, NJ: Ablex Publishing Corporation.

Lazaraton, A. (2002) *A qualitative approach to the validation of oral language tests*. Cambridge: UCLES/Cambridge University Press.

Meiron, B.E. (1998) *Rating oral proficiency tests: A triangulated study of rater thought processes.* Unpublished master's thesis. University of California, Los Angeles.

North, B. (1995) The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System, 23,4,* 445-465.

North, B. (2003) *Scales for rating language performance: Descriptive models, formulation styles, and presentation formats:* TOEFL Monograph 24. Princeton NJ Educational Testing Service

North, B. and Schneider, G. (1998) Scaling descriptors for language proficiency scales. *Language Testing, 15,2,* 217-263.

Nunn, R. (2000) Designing rating tasks for small group interaction. *ELT Journal, 54,2,* 169-178.

Orr, M. (2002) The FCE Speaking test: Using rater reports to help interpret test scores. *System, 30,2,* 143-152.

Politt, A. and Murray, N.L. (1996) What do raters really pay attention to? In Milanovic, M. and Saville, N (eds.) *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium* (pp. 74-91). Cambridge: Cambridge University Press.

Pollit, A. and Hutchinson, C. (1987) Calibrated graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing, 4*, 72-82.

Swain, M. (2001) Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*, 275-302.

Taylor, L. (2001) The paired Speaking test format: Recent studies. *UCLES Research Notes*, from http://www.cambridgeesol.org/rs_notes/rs_nts6.pdf

Turner, C. and Upshur, J. (1996) Developing rating scales for the assessment of second language performance. *Australian Review of Applied Linguistics, Series S*, *13,* 55-79.

Turner, C.E. and Upshur, J.A. (2002) Rating Scales derived From student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, *36,1,* 49-70.

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

Upshur, J.A. and Turner, C. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing, 16,1,* 84-111.

Van Moere, A. (2006) Validity evidence in a university group oral test. *Language Testing, 23,4,* 411-440.