Nvivo in qualitative language testing research: testing spoken interaction

Dr Ana Maria Ducasse
Lecturer Spanish
RMIT University
Melbourne 3001
Australia

Abstract

This paper reports on a mapping technique to find evidence in paired candidate discourse for what is said by candidates and raters in verbal protocols being used to validate components of the spoken interaction construct underpinning a rating scale for speaking.

Using N vivo to map the results from three previous studies onto each other, the aim is to investigate the possibility of demonstrating the correspondence between observable features in the candidates' speech sample and the features of 'engagement' that language experts claim to attend to while observing paired interaction which candidates also report on after performing a test.

The data in this preliminary study includes the speech sample of 2 beginner dyads in a Spanish foreign language context, 2 Spanish teacher raters Verbal Protocols and Stimulated Verbal Recall from 2 candidates. The analysis of the speech sample is guided by features that have been already identified in previous studies which lead to the findings of the study that show what the speech sample and the protocols have in common.

Studies on paired tasks in oral proficiency tests have largely focused on what affects candidate output and /or test outcomes; this study, however, validates the crucial mediating position between the output and outcomes. It is argued that at the point of intersection of what raters attend to and what candidates are monitoring lies an empirical basis for qualitative construct validation of the actual output which can be accessed with Nvivo,

Introduction

The empirical process that leads to identification of elements that comprise the construct underpinning a rating scale for paired interaction is a long one. At the end when the scale is adopted the stakeholders in the test, for example the candidates and the raters, would like to feel assured that the components of the scale are recognised as important features by the raters that mark the test and that the candidates that take the test also know what particular elements comprise what is being tested.

In these times of increasing accountability, if elements that make up the process of paired oral interaction, such as listening, turn taking, and topic change are identified by both rater and candidates and are empirically found to be in a transcription of original test data it is argued that this provides evidence that an element of the ‘construct interaction’ occurs during a test and could be included in a marking system by, for example, being incorporated into a scale.

Purpose of study

This paper reports on a possible method for validating the construct underpinning an empirical rating scale by using Nvivo, but is constrained by space to limiting the example discussed to listening within the context of paired interaction. This scale validation method involves a mapping technique where the discourse from several related studies is collated and searched across using a qualitative data analysis package to uncover similarities in the data. These similarities are considered as evidence offered in support of an argument for transparent methods of construct validation for empirical rating scale development.

Why argue for empirical and transparent methods in rating scale development? Scale development and scale use are processes that result in test outcomes: a ‘mark’. In this paper I move from a focus on the outcomes to a focus on process particularly the construct behind the scale development process. In the past, testing has been bound by traditional statistical methods to validate tests *outcomes* which have served their purpose but have nevertheless remained limited by psychometric constraints. Now with an increase in the number of more innovative approaches from a more qualitative stance stakeholders in language tests can hope for promising methodological changes that will illuminate test *processes* such as how part of a construct is operationalised and validated..

Particular criteria in rating scales for paired interaction are valid and useful tools to rate with in that context, if the elements of which they are comprised are empirically found in candidate paired test discourse. A feature of interaction is a demonstration of engagement by two participants by the way they manage the way they listen during an interaction. It is argued that if candidates produce elements of engagement and they are aware that these elements make up interaction and that raters also attend to these features when marking then there is an argument that because the features exist in the test discourse and if candidates are themselves aware that they are part of performances and raters themselves attend to these features while marking then these elements of speaker engagement are part of the operationalising of ‘interaction’

The research question asked here that encapsulates this is “What elements of ‘speaker engagement through listening’ in a paired interaction, as observed by candidates and raters, are evident in paired discourse data?”Conclusion will be drawn after searching through and connecting the three sets of data.

This research situated in the literature

Studies on paired tasks in oral proficiency tests have largely focused on what affects candidate output as in the ‘quality’ of the discourse (Lazaraton, 2002; Galaczi, 2004) or the combination of factors that effect test outcomes (van Lier, 1989 #103; Berry, 1998 ; Csepes, 2002 #23) . This study, however, validates the crucial mediating position between the output and outcomes. At this mediating position are the scales that are used to rate paired interaction. In this study a part of the construct that underpins them is validated by searching across three discourse studies for points in common.

The concern here is not *how* the pair has affected the test discourse or what *mark* will be assigned because of the particular pairing but rather working back from a position that accepts interaction during communication. This means taking oral interaction as an unavoidable given that occurs between two candidates as they perform a task in a pair. It also requires taking on board the Vygotskyan co-construction of meaning, different patterns of dominance in pairs (Storch, 2001) while validly rating individuals that are performing in a pair and displaying their abilities for test purposes. This could be attempted from a position that values part of an oral proficiency test taking part between a pair of peers. By working up from the multiple data this paper shows how this methodology can be used to validate scale constructs that incorporate issues surrounding the inclusion of elements of interaction such as listening as a criterion in an oral proficiency test.

The context

The setting for the test is an undergraduate Spanish Ab initio language program in a university with a long tradition of communicating language teaching the Spanish program. The program has an average of 250 students studying at the beginner level in the first semester. The students are taught by lecturers, teaching assistants and casual staff.

Before the development of the paired test, the tradition for oral testing in this setting was a mark for oral participation and performance in class for beginners and then teacher/student oral interviews for the language subjects and oral presentations in the content subjects. The innovation of introducing paired interaction brought along with it a need to develop and validate appropriate rating instruments.

In order to improve on the oral testing for the absolute beginner program the Spanish paired test was trialled then established with self selecting pairs whose oral proficiency was at a similar level because they had all started in the beginner's course. This is not to say that language learners are all identical when they start, as students bring different baggage with them, but the essential point is that the partners had never studied Spanish before. In sum, in this context the implementation of a paired oral and the subsequent rating scales development process to rate it was an innovation.

The participants

The participants are first year undergraduate beginner Non-Native Speakers (NNS) and they are partnered by classmates whose oral proficiency is similar in level, within beginners range. This is not to say that language learners are all identical at the commencement of a course, because students bring different language experience with them to a language class regardless of the language chosen to study, but the essential point is that they start Spanish ab initio.

The selection for the pairs was carried out once only, and when the process was completed the same set of paired videos was used for each of the three parts which are combined in this study.

Task type for the paired interaction

The language produced by the candidates in the task on which the construct and scale is based consists of elicitation and response as required by continually renewing context in the interaction. It was deemed sufficient for the purposes of an achievement test and in this context. The topics were taken directly from the topic taught on the course. The cards were written in Spanish in the test.

After they were given the card(table 1), the candidate started on the task as shown below on the sample task cards.

Table 1

Tiene 10 minutos en total
La familia
Los dias festivos
Los amigos

Tiene 10 minutos en total
Los fines de semana
Los pasatiempos
Las vacaciones de verano

They talked for ten minutes and meanwhile the raters listened and rated each of them individually for four criteria. By raters I mean the staff who had experience teaching the course: lectures, teaching assistants and casuals and had been trained and moderated to rate with the criteria.. Some have twenty years experience with beginners others have one year or less.

Why the interest in speaker engagement?

Speaker engagement could be said to be a sign given by the speaker that he/she is listening and also cooperating in a paired interaction. If we are to investigate speaker engagement it needs to be defined and narrowed for this context. When raters were asked to define peer-peer interaction in a previous study {Ducasse, 2006) listening during interaction was one of the categories they equated with successful interaction where speaker engagement was also evident through candidates demonstrating the right conversation management skills while they were listening to each other; these have been researched in the peer-peer context and among them are topic management and turn taking {Dimitrova-Galaczi, 2004) In a connected study (Ducasse, 2009) when candidates were asked to comment on the process of interaction while taking a paired test , they also commented on the effect of listening on the interaction.

Focus of the research

The element focused on here is listening and it has been selected from the coding that emerged from the data after transcribing two sets of protocols. From these two transcriptions two sets of coding grids were developed one for candidate Stimulated Verbal Reports and another for rater Verbal Protocols following the methods set out by {Green, 1998 #215}.

Where coding elements on the two grids are similar then evidence will be searched for using Nvivo to search and match elements in the candidate discourse to support what has been commented on in the two sets of protocols. In the interest of space I am presenting here the example from the listening category, that has emerged from content analysis of raters verbal protocols about candidates' test performance. It is used here to investigate whether candidates and raters have made any parallel observations that can bourn out by what the candidates *actually* said in the tests.

What type of research and why

The type of research presented below draws coherent conclusions from data that is displayed with qualitative data software Nvivo7. The relationship between stakeholders and output is clarified by listing examples and which contribute to a coherent understanding of a large quantity of discourse drawn from three different perspectives on one speech event..

Data

The table 2 below illustrates the quantity of data that has been uploaded on to Nvivo7 and that has been searched through to arrive at the conclusions drawn below. It is divided into two studies the first for developing the scale and the second for validating it.

Table 2

	Data	Data set	analysis
Study 1 Scale Part 1 Part 2	On 16 x 10 min. videos of operational paired orals Verbal reports by L2 specialists Criteria emerged from VR	12 Spanish L2 specialists each specialist observes and comments on 3 different pairs each pair is observed twice by two different Spanish L2 specialists 6 L2 specialists develop scales Empirically	Analysis of Verbal Reports Scale trial..
Study 2 Part 1 Part 2	Perceptions from VR Test discourse Candidate SVR	12 Spanish L2 specialists VR on test 16 pairs do test 25 candidates comment on their performances in that same test	**Compare VR to test discourse** **Compare SVR to test discourse**

Study 1 scale development

The definition of peer-peer interaction is examined from the point of view of trained Spanish speaking second language specialists: tutors and lecturers, who are trained to mark the speaking test used in this context. The interaction takes place between two candidates in a paired task in which there is no participation from an interviewer or rater. This is important to note because the most common kind of oral tests are interviews so the dynamic is different when a candidate speaks to a peer instead of an interviewer.

The definition of interaction is explored by having the language specialists watch video clips of students taking the test in pairs and asking them to simultaneously comment on the interaction between the pairs of candidate speaking to each other in the test. This method of data collection is called Verbal Protocols (Greene) and it is used here with the intention of illuminating the meaning that Spanish teachers attribute to the word 'interaction' between candidates in a test because they will mention features that are salient to them.

By asking them to comment on the performance they define and shape the meaning of interaction which is intrinsic to the context of testing in pairs. Delving into how candidates 'do' interaction from experienced language specialists' professional point of view is indispensable to this research because they are externalizing their professional received wisdom on the attributes of interaction between candidates that could otherwise effect raters' judgement of paired performance if these attributes were not present in rating criteria.

These comments on interaction were funnelled into empirical scale development but the point is that by having teacher input into scale development, the criteria for marking tests will be directly linked to observable features in actual performance. The function of these observable features in rating will be to replace the prevailing intuitive method of rating orals most commonly used in the context of assessing L2 Ab Initio learners. This paper contributes to research carried out in order to validate these empirically developed scales by handling a lot of data with Nvivo..

Study 2 scale validation

The data includes the speech sample of 17 beginner dyads in a Spanish foreign language context, 12 Spanish teacher raters Verbal Protocols and Stimulated Verbal Recall from 17 candidates. In the week after the paired oral, the candidates returned individually to watch their performance on video clips from their live oral test in return for personal feedback and to take part in the study. In this way I audio taped the SVR verbal protocols which were recorded while 24 individual candidates watched videos of their on-line test performance and commented on their performance of the interaction with their partner. In the sample there were 15 random and self selected pairs of students of which 9 pairs had both candidates participate and the other 6 pairs only had one candidate from the pair take part in observing and commenting on performance.

They were asked the following: ¨By observing your paired performance with another candidate comment on what you recall was 'happening' in the interaction. Please comment on all aspects of the performance including, but not restricted to, what or how something is said. Comment as you see examples of successful or unsuccessful interaction".

The candidates watched with a remote in hand paused the video and spoke into the tape recorder when stimulated by the video to recall some thing in the interaction on which they wished to comment." The SVR sessions lasted not much longer than 15 minutes and a limitation is that I was not able to train the candidates in giving feedback. Training time would have been longer than feedback time and I don't feel the data is in any way compromised. Needless to say, some candidates were much better at giving feedback.

The analysis of the speech sample was guided by features that had been already identified in the previous rater study and it pointed to elements that the both the candidate and the rater protocols had in common. This leads to the final mapping of the two verbal reports onto the actual test discourse to help illuminate the test taking and rating process and to validate the scale using all this as input during the development process.

Methodology

It is a particular methodology enabled by the searching power of Nvivo that allows the three studies to be mapped onto each other in an innovative qualitative validation. By using rigorous empirical methodology conclusions can be drawn about these scale developing, test rating and test taking processes.

Using N vivo to map the results from three previous studies on to each other, the aim is to demonstrate the correspondence between observable features in the candidates' speech sample and the features of 'engagement' that language experts claim to attend to while observing paired interaction which candidates also report on after performing a test.

Data analysis

The two types of protocols described in the studies above were audio taped then transcribed. All text was transcribed word for word, from sections delineated by the candidates pausing the tape to comment. I segmented the protocols by the naturally occurring pause then conducted a qualitative analysis of the transcripts resulting in my grouping the protocols by type of comment; I identified themes and categories relevant to interaction and developed a coding grid for coding all the data.

The data was converted from a text to a table in order to prepare for checking by another coder. Both sets of coded data, the Stimulated verbal recall and the verbal reporting on interaction, were then checked for rater reliability in the coding and were above 80%. As is clearly stated by Green (1998:12) the 'validity of codings is related to the reliability of codings'. So the categories that were defined had the larger number of comments and other themes were subsumed into larger groupings. Nvivo7 was very helpful in defining and clarifying the categories for the coding. The category explored from across the three sets of data to give an example of this technique is listening.

Findings

The attentiveness and the support of the candidate who is a listener helps advance the conversation between the two candidates. Not only do the raters comment on those but the candidates themselves also make comments on their 'listening'; behavior during the test. Below are examples of discourse that support this position by raters and candidates. This is followed by an excerpt from the test discourse that supports both the rater and the candidate observations.

Preliminary triangulated findings on listening

A peer-peer listener was defined by the raters as: A supportive listener in a pair showing *verbal* signs of comprehension or providing *audible* support to the speaker. This type of listening was divided into two subcategories: a different type of listening termed *interactive* listening and *comprehension*. Interactive listening was divided into two categories again: *audible support* and *verbal support*.

Firstly, the type of support offered through listening attentively is *audible* in interactive listening and provides feedback such as back-channelling while the other speaker maintains the floor. There are some features noted by the raters that would convey cooperation and interest for the speaker to continue but would not indicate to the speaker anything beyond interest, as in registering comprehension. Of the two types of supportive listening this first one did not require comprehension, just support.

Secondly, the *verbal* support in interactive listening provided by the listener involved filling a silence or providing the word the other partner was searching for but could not produced fast enough which enabled the interaction to continue. As is said by rater 5 about pair two: “The one on the right says the word the other is looking for”…rater 5:2. This is borne out by an example from pair two doing that.

Table 3

Line	pair	candidate	discourse
9	2:	L	no uh mi padre uh No yh my father uh
10	2:	R	trabajo? Works?
11	2:	L	si si si mi padre trabaje Yes yes yes mi father works

This second type of interactive listening required a demonstration of comprehension by the listener. This support in the interaction provided on the part of a listener that signals interest audibly or verbally by comprehension is a new conceptual category for analytic rating scales.

Comprehension per se as the other subcategory meant asking for help, as well as offering it, or asking for clarification in order to continue correctly. It meant comments raters made about candidates regarding comprehension or lack thereof. The quote below in table 4 shows how aware raters are of candidates’ demonstration of comprehension. Rater eight comments on pair eight in particular the candidate on the left of the pair.

Table 4

Rater	pair	candidate	rater content analysis
8:	8	L	“The questions say more than the answers. Are they related to what the other person says, r are they out of the blue and have nothing to do with the other person.

Following the comment above two different situations with candidate 8L and 8R are shown in the excerpt below in table 5 where they don’t listen to each other and change topic abruptly.

Table 5

Line	pair	candidate	Discourse
4	8	R	Muy bien. Y tu? Very well and you?
5	8	L	Primero, cuando eras niña uh que haces qu que haga Firstly , when you were a child what did you do what did you do cuando niña? as a child
6	8	R	cuando era niña um jugaba con mis amigos y caminaba When I was a child I placed with my friends and walked en el parque con mis abuelos. um yo nadaba en la pis piscina in the park with my grandparents um I swan in the p pool y si y tu? and yes and you?
7	8	L	uh Cuanto hace cuanto tiempo hace que tu ah tu tener uh how long is it how long ago that you ah you had tu primer beso? your first kiss
8	8	R	tiempo hace? hace doce años How long ago? Twelve years ago
9	8	L	i? con quien ? vho? With who?
10	8	R	nigo (risa). Trabajo? ny friend (laughter) you work?

Curiously in the candidate retrospective recall candidate 8L says
table 6

Line 3	8	L	I am thinking of the next question not really listening while she is talking.

And further down we can see there is an explanation for candidate 8 R to have asked an unrelated question:
Table 7

Line 5	8	L	I did that hand wave for her to ask the next question

The other partner in the pair 8R also comments on what was happening during those first moments:
Table 8

Line 2	8	r	What is he asking I am thinking I am afraid that I won't understand his question

So while one is not listening and the other is worried about comprehension the disourse that emerges from the interaction appears as is seen above in table 6: stilted and disjointed.

So rightly the rater, in the verbal protocol, comments that the question and what went before or the question and what follow holds together the interaction. The two first "and you?"(Lines 4 and 6) from the candidate on the right (R) are ignored and he recognises that he was not even listening. When the partner 8R says "Do you work?" out of the blue in the extract from the discourse, there is a reason and with the candidate's retrospective recall we know that this has an explanation.

Discussion

It is argued that at the point of intersection, of what raters attend to, what candidates are monitoring and the actual output, exists an empirical basis for qualitative construct and scale validation. The preliminary findings illustrated with single examples above show great potential for further use of this mapping technique to provide examples to support data triangulation.

In testing terms the features of interaction that are noticed by the teachers are what testers call the 'construct'. The construct being tested is like a three dimensional shape a polyhedron made up of a myriad of flat surfaces because the construct of interaction is multi faceted: one can't see all of the sides at once but because it is three dimensional you know it is a whole and hence there is something on the other sides. I argue that interaction has so much to it, that though it is all there, we are not able see it all at once so the features teachers see are only part of what makes up interaction, not interaction as a whole. One of the parts raters see is when the construct interaction is operationalised by them as a component of 'listening' during an interaction.

What have we learned discovered from this

In order to verify whether rater perception and candidate awareness are verifiable the verbal protocols are transcribed and analyzed. The comments, which emerged from the date on listening , are checked against the transcriptions and videos of the student performance. Checking the features salient to the raters against the student performance is a way of validating rater judgments along the lines of "this feature stand out and is perceived by the raters and it also appears in the language used by the students t" thus it can be concluded raters perceive features that are empirically observable in the data. It is taken one step further when candidates comment on the very aspect that raters have focused their attention on such as 'not paying attention to comprehension"

What are the implications?

We can use this Nvivo software to enable empirical development and validation of different criteria for different needs with this mapping method. Empirical criteria for pronunciation, for intonation, for details connected to speaking that are not as clear cut as 'errors' of semantics and syntax . Now more elusive features can be captured, observed and if need be have criteria developed to suit a particular context.

Further research

Now we have the tools to manage such large amounts of qualitative data, greater quantities of student writing taken from corpuses could be processed and matched against both candidate and test taker processes for very large scale tests such as TOEFL and IELTS. The same question can be asked 'Are students performing in ways their raters perceive them to behave in the oral tests? This could be answered empirically to inform the construct of speaking tests the vortex of the test development process.

References

Berry, V. 1998. Personality and oral test score variability. *TESOL conference*, Seattle ,WA.

Csepes, I. 2002. Measuring oral proficiency through paired performance. Budapest: Budapest.

Dimitrova-Galaczi, E. 2004. Peer-peer Interaction in a Paired speaking test: the case of the First Certificate in English. pp. 290. New York: Teachers College, Columbia University.

Ducasse, A. M. 2007 How do candidates view interaction in a paired oral. In Gitsaki, C. (Ed.), Language and languages : Global and Local tensions'. (pp. 184 – 200). Newcastle: Cambridge Scholars Publishing

Ducasse, A.M. 2010 Interaction in Paired Oral Proficiency Assessment in Spanish: Rater and Candidate Input into Evidence Based Scale Development and Construct Definition, Peter Lang, Frankfurt. A specialist series: "Language Testing and Evaluation".

Green, A. 1998. *Verbal Protocol analysis in language testing research: A handbook*. Cambridge, England.: Cambridge University Press.

Lazaraton, A. 2002. *A Qualitative Approach to the Validation of Oral Language Tests*. Cambridge: Cambridge University Press.

Storch, N. 2001. Role relationships in dyadic interactions and their effect on language uptake. *Department of Linguistics and Applied linguistics*, Melbourne: The University of Melbourne.

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

van Lier, L. 1989. Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly* 23: 489-508.