Using a Conversation Analytic Exemplar-based Rubric to Assess Engagement in a Paired EFL Test

Greer, Tim

(Citation)
神戸大学国際コミュニケーションセンター論集,16:1-24

(Issue Date)
2020-03-31

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/81011980

(URL)
https://hdl.handle.net/20.500.14094/81011980

『神戸大学大学教育推進機構国際コミュニケーションセンター論集』16 号（2020 年 3 月）
Journal of the School of Languages and Communication, Kobe University, Vol. 16 (March, 2020)
NCID=AA12182319

Using a Conversation Analytic Exemplar-based Rubric to Assess Engagement in a Paired EFL Test

Tim GREER
Kobe University, IPHE, SOLAC

Abstract

This study documents the piloting of a video-recorded exemplar-based rubric in relation to the construct of engagement within EFL interaction. The data consist of 73 video recordings of paired discussion tests between Japanese students in a first-year university EFL program, as well as recorded conversations between teacher-raters viewing those tests. The rubric highlights specific interactional practices that point to "engagement" within the test setting, and demonstrates how incorporating conversation analytic (CA)-style observations can facilitate an emically grounded assessment. The deeply descriptive rubric draws on segments of exemplar video-recordings to give account for how test-takers incorporate such resources into their talk. The study also conducts a qualitative content analysis of an inter-rater hermeneutic dialogue to consider the extent to which non-CA-specialist raters are able to use these exemplars to both operationalize and assess the concept of engagement.

Keywords
EFL discussion testing, Interventional Conversation Analysis, Test development, Engagement, Exemplar-based rubrics, Inter-rater hermeneutic dialogue

1. Introduction

One of the greatest challenges for teachers in English as a Foreign Language (EFL) contexts like those found in Japanese universities is how to assess a student's English interactional competence (IC). Traditionally, this has been done with one-on-one interviews between the teacher and each student, particularly in high-stakes contexts. However, this is not always feasible with large classes and as a result low-stakes tests, such as term-final speaking exams, are often carried out in pairs or with small groups of students. Video recording such tests leaves instructors free to carry on teaching the remainder of the class and allows them to view the test data in careful detail at a later time. Since the teacher does not participate in the conversation, it also means that the students must take a more active role in initiating topics, maintaining the talk and dealing with instances of interactional trouble (Gan, *et al.*, 2008). While this provides them with a greater opportunity to speak, in some cases it can lead to less communication when students are reluctant to speak or give only limited responses.

Paired student discussion tests in EFL contexts are commonly graded according to some form of rubric, typically one involving broad descriptors that have been simplified to fit an idealised case (Talandis, 2017). An alternative approach, however, constructs the rubric based on an extensive written description of a video-recorded exemplary case at each performance level. With its long history of deeply descriptive observation of unscripted interaction, Conversation Analysis (CA) (Sidnell & Stivers, 2012) is a highly suitable tool for constructing such exemplar-based rubrics.

This study documents the development and implementation of one such rubric, particularly in relation to the notion of "engagement". The data consist of 73 video-recordings of paired discussion tests between Japanese students in a first-year university EFL program, as well as video-recordings of a pair of teacher-raters watching a selection of the student videos in order to rate their interactional engagement according to the assessment rubric.

The research project views engagement as an emically reconceptualised form of the notion of “willingness to communicate” (Yashima, 2002), and evidence for a test-taker’s engagement can therefore be found in publicly available interactional practices, such as relevant post-expansions, stepwise topic shift, collaborative repair, and third-turn uptake (Schegloff, 2007). This study highlights such specific interactional practices as visibly available forms of engagement within the test setting, and thereby demonstrates how conversation analysis rubrics can help facilitate an emically grounded assessment. By externalizing willingness to communicate in this manner, participant orientations become accessible to test raters who observe the interaction. The deeply descriptive rubric (see Appendix) draws on segments of video-recordings to give a detailed account of how test-takers incorporate such resources into their talk. The paper will also discuss the extent to which teacher-raters who are non-CA specialists are able to use these exemplars to both operationalize and assess the concept of engagement.

2. Literature review

2.1 Background

The genesis of this study lay in a common enough observation in Japanese EFL classes: occasionally in my classes I would come across students who were not very proficient in

English, but still demonstrated a high interest in communicating with their partner, and this would often lead to the sort of conditions that foster opportunities for language use. Even when such learners had relatively low linguistic skills in terms of accuracy, fluency and complexity, it was still possible for them to take an active part in a conversation with the limited resources they had available to them. In short, they were highly engaged in the interaction, and this was therefore something that was worth incorporating into a discussion assessment instrument, since it could also have a positive washback effect, encouraging students to consider the importance of engagement for interactional competence. In this section, I will briefly consider some of the related research on engagement that has informed my study, particularly in regard to second language discussion testing.

2.2 Conversation analytic research on interactional competence

While traditional L2 acquisition research has adopted a largely psycho-cognitive stance that locates linguistic ability within the head of the individual, recent socio-constructionist research on *interactional competences* (IC) (e.g., Hall, Hellermann, & Pekarek Doheler, 2011) has offered an alternative approach that locates proficiency in the interaction itself. ICs are competences that are co-constructed by any and all participants in a conversation, not the L2 user alone (Young, 1999), and they encompass both linguistic and non-linguistic interactional resources. This position is grounded in Kramsch's early calls (1986) to view language as fundamentally social and to teach and assess proficiency in ways that pay attention to functions, performance and negotiation of meaning.

Early proposals were made for using CA as a resource for material development (Gardner, 1994) and as a means to understanding classroom learning processes (Markee, 1994). However, it was Firth and Wagner’s (1997) seminal paper that prompted SLA researchers to seriously re-consider the “mainstream” cognitivist approach to SLA. Firth and Wagner argued that SLA research should reconceptualize our view of language, learners, and acquisition, and treat L2 speakers' competence in terms of the way they use language in a locally contingent, practice-specific, co-constructed manner. CA has since become a prominent approach to investigating people’s use and learning of L2s, under the acronyms *CA-for-SLA* (Markee & Kasper, 2004) and *CA-SLA* (Pekarek Doehler, 2010). Sfard (1998) views learning as the increasing ability to participate by making use of the interactional practices that are routinely used in the target community, as demonstrated through the sequential details of episodes of actual talk. This notion of interactional competence as "knowing what to do next" is also captured in the CA notion of progressivity (Schegloff, 2007), and it is a fundamental assumption in interaction between proficient members of any community. Since CA aims to unveil things people regularly do competently, it is adept at revealing the details of naturally occurring talk, and this has led to its ongoing concern for issues of interactional competence.

2.3 Assessing interactional competence

Perhaps due to the importance and attention they are given, the vast majority of assessment studies on interaction have taken place in high stakes contexts, such as oral proficiency interviews (OPI) like ACTFL and IELTS. Such CA research has examined the OPI as a social event (Seedhouse & Egbert, 2006) and considered peculiarities of the test format that distinguish it from mundane conversation in fundamental ways. The asymmetrical nature of an interview, in which one person asks all the questions and the other responds without reciprocating, is a ubiquitous yet unavoidable part of this form of institutional interaction. In addition, the rules of the assessment often prevent the interviewer from initiating repair on something unclear the test-taker says, and the fact that there is no requirement to achieve intersubjectivity is at odds with the basic tenets of natural conversation (Seedhouse, 2013). On the other hand, the interviewer may pursue an apposite response by ignoring or reworking the response of the test-taker in order to achieve the institutional aims of the test (Okada & Greer, 2013).

Perhaps partly in response to such issues, some testers have begun to use alternative test-formats, such as paired (Brooks, 2009; Galaczi, 2008, 2014; Greer, 2019; Philp *et al.*, 2013) or group-based interaction (Gan *et al.*, 2008; Greer & Potter, 2008; Leyland *et al.*, 2016). Such formats are particularly appropriate in low stakes contexts where there is often a need to gather assessment information on a large number of students in a short amount of time and with limited resources; however, these practical matters are not the only advantages.

Galaczi (2008) identifies three global patterns of interaction in peer-to-peer test forms, which she terms collaborative, parallel and asymmetric. Taking into consideration the way the participants structure the organization of their turn-taking, sequencing and topic management, she notes that the "ideal" approach is the collaborative one in which test-takers listen to their interlocutor and base their next turn on what has just been said. However, a significant number of the pairs in Galaczi's study also adopted either an asymmetric style in which one partner dominated the conversation, or a parallel style in which both participants took long turns that were more or less unrelated to what the other person had just said. Rather than pointing to any deficiency in the paired test format, Galaczi's study in fact identifies the need for both test-takers and assessors to be aware of turn-taking management, topic development and listener support strategies as fundamental features of interaction. Galaczi's later work (2014) identifies these as distinguishing features across levels of interactional competence.

Due to the limitation of space, this review has been necessarily selective, but for a more comprehensive coverage of this topic, see Sandlund *et al* (2016). Although there has been considerable CA-informed work on candidate interaction in L2 speaking assessment contexts, there has been less research that takes an interventional CA approach (Antaki, 2011) by applying these findings to the rubric itself. In what follows, I report on a project that takes on that challenge.

3. The engagement rubric and its use by teacher-raters

3.1 An overview of the KTOP dataset

As outlined above, this study draws on two related datasets: (a) a collection of video-recorded paired student discussion tests known as the Kobe Test of Oral Proficiency (KTOP), and (b) an hour-long video-recording of two teacher-raters discussing a selection of those tests while referring to the exemplar-based rubric I devised. The KTOP dataset is part of a broader corpus of EFL oral proficiency tests video-recorded at Kobe University from June, 2015. The test-takers were all first-year Japanese students and the test was one of three assessment items for an oral English class, constituting 20% of the students' overall grade. At the point the data were collected the test-takers had participated in eight weeks of a course that focused on developing their spoken fluency through discussion. They discussed the six topics in class over six weeks (yourself, your extended family, travel, marriage, share-housing, and jobs), and these were also the topics given during the test.

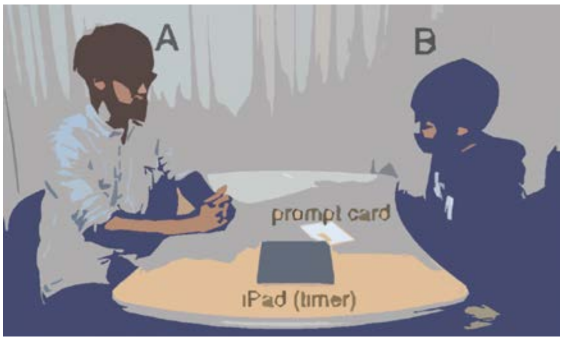

Figure 1. The seating configuration during the discussion test

At the beginning of the test, the test-takers were randomly assigned one of these topics by selecting a card with the topic written on it. They were then asked to talk freely about that topic for four minutes.

The test took place in a room near the students' regular classroom. Throughout the data transcripts, the student on the left is designated as A and the one on the right is B, as shown in Figure 1. There was also a camera operator in the room whose job it was to administer the test and record it for the instructor, who later graded the test-taker's conversation skills later in terms of fluency, accuracy and complexity. The camera operator (who will be called C) was not otherwise involved with the students' classes and was completely unknown to them prior to the test. Like the test-takers, C was Japanese, and her job involved proctoring the test by seating the students, checking their names, beginning the timer and managing the closing stages of the procedure (Greer, 2019).

On the table there was an iPad open at the Clock app, and this was used as a timer (Figure 1). The timer was set to four minutes and the camera operator pressed the start button to indicate the test had begun. Later the teacher watched the videos and assessed each student according to the following rubric (Figure 2).

Fluency Did you respond quickly without too much silence?		☺ 😐 ☹
Accuracy Was your grammar and pronunciation understandable?		☺ 😐 ☹
Complexity Did you challenge yourself and respond in detail?		☺ 😐 ☹
Engagement Did you actively interact with your partner?		☺ 😐 ☹

Figure 2. The simplified rubric used during grading

Each of the four focus areas (Fluency, Accuracy, Complexity, and Engagement) was awarded a score out of ten, with 6 as a pass, 7 as a B, 8 as an A (and rarely 9 or 10 as an A+ or S). The smiley faces were intended as an indication of the teacher's initial reaction, and were usually filled in before the grade was assigned toward the end of the test. The teacher also wrote a few short comments on each test-taker's performance. In the early recordings the students were only rated in terms of Accuracy, Fluency and Complexity. The engagement construct was only added after the initial data was collected, so the exemplars were later re-graded for this descriptor and became the basis for subsequent tests, which included all four items.

The first three criteria are reasonably straightforward, but the notion of *engagement* is one that needs further clarification. Although the version of the rubric above is a highly simplified one and was intended as a shorthand format that was accessible to the students themselves, the teacher also developed a more complex yet holistic understanding of what each of these criteria involves. The exemplar-based rubric was aimed at teachers and used a CA-style description of the test-takers engagement at each level (A, B, C, D) based on videos from the initial KTOP recordings as examples. A complete outline of these descriptions can be found in the Appendix.

The exemplar videos were chosen because they represented clear cases of engagement at each of the four levels. The descriptions were intended to be detailed, but still accessible to non-CA specialists. They often included extracts from the KTOP transcripts that illustrated key instances of interactional engagement within the video. They were designed to fit on one A4 page so that the teacher-raters were able to process them without too much difficulty. Each one was accompanied by the relevant video and assessors were asked to watch the video in conjunction with the descriptive rubric before later rating other videos based on the exemplars.

3.2 Rating the exemplars

Once the video exemplars had been finalised, two EFL teachers (who I will call "Bob" and "Jim" here) were asked to rate eight other pairs in order to ascertain if they could identify evidence of engagement in the interaction. The raters were both L1 speakers of English and were highly experienced in teaching Japanese learners. After they had completed their ratings individually, they discussed the ratings with each other, justifying any differences of opinions in a form of inter-rater hermeneutic dialogue (Walters, 2007). Each rater was therefore asked to:

1. watch the exemplar videos and read the detailed rubrics on engagement for each level
2. watch each of the eight un-assessed videos, give their rating for each participant according to the rubric and grade them on their engagement for this activity
3. compare the grades they assigned with those of the other rater and then discuss any differences in opinions you may have had.

The raters' discussion was video-recorded and later qualitatively analyzed for content (Cho & Lee, 2014; Elo *et al.*, 2014).

3.2 Analyzing the raters' discussion

The video recording of the raters' discussion was subsequently transcribed and later inductively coded into six themes that were drawn out of the data. The focus was on extracting categories that explained and operationalized the notion of engagement, according to the two teacher-raters' understanding of the descriptive rubric document. Table 1 lists the themes that emerged from this coding. These themes were later coded independently by a second researcher in order to verify the coding. Where there was any discrepancy, the researchers discussed the difference until agreement was reached.

Table 1

An overview of the thematic categories that emerged from the raters' discussion

Thematic category	Gloss
Sequence expansion	The speaker precedes, intervenes and extends on the base turn sequence
Enthusiasm	The speakers appear interested in the talk
Parallelism	The speakers are talking about different topics and there is limited coherence between turns
Embodiment	The interactants use gestures, gaze and so on to display their engagement
Timing	Over-attention to the timer distracts from the interactants' engagement
Authenticity	The interaction appears "genuine"

Two secondary and unexpected points also arose from these discussions: (1) moments when the raters appeared to misuse the rubric, accounting for their grade in ways that was not intended by the rubric, and (2) instances in which the teacher-raters questioned the validity of the rubric. These will be dealt with in detail in the following section.

4. Findings

4.1 Introduction

This section will present the study's findings. Via qualitative content analysis of the teacher-raters' comments, it will expand on the six recurrent themes that emerged during their hermeneutic dialogue, namely sequence expansion, enthusiasm, parallelism, embodiment, timing and authenticity (Section 4.2). On the whole, the raters' comments demonstrate that they were able to capably interpret the notion of engagement in line with the detailed descriptions given in the exemplar-based rubric, although there were also times when they added their own interpretations that went beyond the rubric and instances in which they questioned the test format and its effect on the interaction. Those issues will be taken up in Sections 4.3 and 4.4 respectively.

4.2 Thematic categories

As outlined above, the themes to be discussed emerged from the raters' discussion, and we will examine each in turn in this section, particularly with a view to how they demonstrate the teacher-raters' real-time operationalization of the notion of engagement in regard to the video-recorded discussion tests they watched.

4.2.1 Sequence expansion

One of CA's core concerns is the way in which speakers use sequences of turns to precede, mediate or extend a base sequence of talk, a notion that Schegloff (2007) calls sequence expansion. Most EFL teachers might be tempted to gloss this as "follow-up talk", and that representation would not be entirely inaccurate, although CA's approach would also consider pre-expansions (sequences that lead up to the base pair) and insert expansions (such as repair sequences that appear after the initiation of an action and before the recipient has responded to it) along with post-expansion sequences (Stivers, 2013). The rubric's exemplar descriptions refer specifically to post-expansion as evidence of a high level of engagement. For example, at the A level, the rubric states "their uptake ... is often followed by post-expansions that provide evidence to demonstrate how they understood their partner's contribution" (see Appendix) and cites the following extract from the exemplar video:

A I- (0.3) often went to sannomi↓ya kobe
B **O::::h** san[nomiya]
A [to: play]
B **very fashionable**
A Ye::s
B fa:[shionable]
A [fashionable]
B **ni:ce place**

At the time of writing, terms like "post expansion" are probably not familiar to many Japan-based EFL teachers, but that is not to say they are not able to recognize these practices when they see them. The raters instead talked about expansion in terms of "building on each other's responses" (Bob) or "reacting to the partner" (Jim). Conversely, Bob also rated one speaker as relatively less engaged in the talk by noting, "B asked A 'how many people would you like to share with?' (and) he says, 'Two". That's the end of it, y'know? They could have gone somewhere with that." In other words, Bob noticed that B had initiated a new sub-topic that was hearable as a kind of pre-sequence, or at least made relevant the broader project of post-expanding any response with an account for that answer. Instead, in the video they were rating, the selected next-speaker gives a short (numerical) response but the account is missing. Whether or not this was because A was unable to produce it due linguistic proficiency or the inability to come up with a timely reason for his answer, the raters noticed the absence of the account and it left them with the impression that A was not highly engaged in the conversation. On the other hand, when the speaker was able to extend on the topic in a natural fashion, the teacher-raters assessed them as more competent in terms of engagement.

The raters also occasionally glossed expansion in terms of control and passivity. For example, Bob contrasted the performance of two of the test-takers in the following way: "he's kind of taking control and she's a little more passive but she's trying to open up the conversation in other ways and he's sort of missing all these hints". Bob went on to specify A had "taken control from the start" by grabbing the card and asking the first question, but he nonetheless felt that B was more engaged in that she was "offering more information and trying to expand on the conversation and he's just missing it all". Although control is not mentioned directly in the rubric, the expression "listens passively" does appear once in the A-level exemplar and it is seems that the teacher-raters sometimes interpreted sequence expansion in these terms.

4.2.2 Enthusiasm

The teacher-raters also saw engagement in terms of the test-takers' perceived interest in the topic, as demonstrated through a variety of interactional stances and paralinguistic features of the talk. At one point, for example, Bob noted, "his gestures are kind of strange, but it does show you he's enthusiastic, which I think is engagement". This suggests that the raters viewed enthusiasm not only in terms of spoken contributions to the topic, but also spacio-visual messages deliver via the body (see 4.2.4). Bob was also aware of changes in the speakers' enthusiasm level over the course of the test, and rated them higher as a consequence: "It started out as an exercise and then, I mean, it is all the way through, but then they slowly become more enthusiastic about it"

The raters also noted the connection between the topic and engagement. Where test-takers were unfamiliar or uninterested in a topic, they came across as less engaged, but when they were able to shift the topic to something they liked better their enthusiasm for it rose. As Bob put it, "Now they're actually engaged cause they're interested". Although enthusiasm is obviously a somewhat subjective notion and it is not something that has been dealt with at length in the CA literature, it does appear incidentally at one point in the B-level rubric (see Appendix), which says "the speakers receipt each other's turns more *enthusiastically* and build on the topic with assessments and reciprocal questions in which the turn changes back and forth more frequently" (italics added). In this case, it seems that enthusiasm is interactionally constitutive of immediate uptake, collaborative turn-taking and relevant post-expansion, and is therefore closely linked to other themes that emerged from the raters' discussion. Although it may be difficult for teachers to explicate enthusiasm in interactional terms in this way, it does provide them with a loose starting point that can lead them to search for interactional evidence in the video that gives them that impression.

4.2.3 Parallelism

Analyzing topic management among pairs of L2 speakers of English during a peer-peer discussion test, Galaczi (2008) notes three typical interactional patterns which she terms "collaborative", "parallel" and "asymmetric". The collaborative pattern is perhaps the ideal form of engagement, in which the speakers listen carefully to what their partner is saying and design their next turn to align with what has just been said. In collaborative talk, the turns may be shorter and speaker transition becomes more frequent. On the other hand, asymmetric talk is unbalanced, with one speaker taking a long turn and the other contributing relatively less to the conversation: according to Galaczi, such speakers exhibit moderate mutuality and low equality. Meanwhile, parallel interaction has low mutuality and high equality, since both speakers say quite a lot, but there is little connection between what they say. A typical example is provided in the D-level of the rubric (see Appendix), in which the recipient produces a turn that does not specifically address what the prior speaker has just said, but instead initiates a long turn on a different topic. One regular impression that a listener might be left with is that B has been preparing what she is going to say while A is talking.

It is likely the teacher-raters were unfamiliar with Galaczi's terms for these aspects of topic management, but certainly after seeing them in the exemplar videos they did seem familiar with the practices themselves. In this respect, the rubric itself appears to have had an instructive element for teachers, providing them with terminology to discuss interactional phenomena that they come across regularly in their classes. Unsurprisingly, the raters both viewed parallel topic management as undesirable. For instance, while discussing Student A in one of the ratable videos, Bob said, "I gave him D because he talks about and offers his own information and expands on what he's saying himself, but he never asks the other guy 'what do you think about it?' or, yeah, he doesn't really consider B's position, the other guy's position at all y'know like, whether he's part of the conversation or not y'know, he just wants to say like if he spat out a bunch of English in there he can get graded on how much he said". In another video, Bob formulates this lack of discursive cohesion in terms of low mutuality: "He's not really responding to her: he's just waiting for his chance to speak".

4.2.4 Embodiment

Since the embodied turn in conversation analysis (Nevile, 2015), there has been increasing attention paid to the way the entire body is used in interaction and how any given face-to-face sequence of talk is made up of co-operative, multimodal laminations (Goodwin, 2018). The interactants' gesture and gaze are therefore an integral part of how they display their engagement. As they were based on conversation analytic observations, the rubric descriptions also picked up on elements of embodiment in the exemplar videos, leading the raters to incorporate this into their discussions of the additional videos.

At one point, for example, Bob said, "I thought the gestures are great in that it shows how they're really into the discussion um, y'know they're just, they're trying to explain what type of girls he likes, just does the short part to help the other guy understand and then they respond really quickly to each other". Here Bob recognizes the role of gestures in covering for limitations in the test-takers' spoken English and suggests that the embodiment within their talk enables the recipient to respond in a timely manner and to focus on the content of the discussion, rather than its form. Bob rates this as a desirable feature of learner interaction, and this is in line with the ethos behind the rubric as well. At another point, Bob and Jim made the following observations:

BOB again these guys have got great gestures
JIM yeah, I suspect you can just look at that and it would correlate
really well with just the- the- the outcome of the engagement
BOB that's true yeah

While obviously speculative in nature, Jim's remark here offers some insight, not only on test pairs who actively use gestures, but also on those who do not. Even when watching the videos without the sound, one gets a sense of those test-takers who are actively pursuing interaction and those who are reticent to talk. In fact, some of the exemplar pairs who were rated highly in terms of engagement were not necessarily high in terms of accuracy. This highlights the need for an engagement rating, because it is often highly engaged students who are willing to use their English in real situations beyond the classroom.

The raters also drew connections between sustained mutual gaze and engagement. Jim's initial comment after viewing a pair that he rated poorly was, "Yeah, they're not looking at each other so much, are they?" Likewise, Bob viewed gaze direction as indicative of the recipient's engagement with the talk, suggesting that a speaker who was not maintaining gaze with her interlocutor was being inattentive: "See even like now, just holding the card and looking at it. You know, it's not really listening: she's thinking about "what's the next question?"

4.2.5 Timing

Timing was another means the raters used to specify engagement in the video data, and again this aligns with the descriptive rubric. Their argument here boils down to the perception that over-attention to the timer distracts from the test-takers' engagement in each others' talk. In fact, elsewhere this dataset has been used to show that student gaze shifts toward the timer, particularly in the closing moments of the test, systematically impacts on their interactional engagement (Greer, 2019). In terms of the rubric, this was only mentioned once (within the D-level descriptor), but the raters were able to extrapolate this to other videos and agreed that it provided evidence about the test-takers level of interest in the content of their talk (or their attention to the test setting itself).

At one point, Bob noted "He knew exactly when the time was up" as the test-taker adjusted his talk to the time remaining rather than just forgetting about the timer and focusing instead on what his partner was saying. As Bob summed it up, "I guess the idea is that if you're engaged in what you're talking about time flies". Again, this indicates that the raters were able to incorporate this element of the rubric into their assessment of the students' engagement levels, and perhaps it is a particularly pertinent one to be aware of, since it is easy to identify once the rater is made aware of it.

4.2.6 Authenticity

The final category to surface from the raters' deliberations was authenticity, and this perhaps rests on many of the other categories. In the end, when test-takers are engaged in their conversation, the interaction somehow appears "genuine". They have not over-rehearsed the questions they ask, they do not already know the answers their recipient will provide and they react in a way that treats the conversation as real. An engaged speaker is not just talking for the sake of the test, but in order to impart and receive information from their interlocutor. The rubric points to this facet of engagement in such expressions as "They ... seem to be *genuinely* reacting to the content of what the other person has just said" (italics added), and the raters picked up on this too, as shown in the following excerpt.

BOB I ask you a question you give me a response and all that sort of thing
but um, later they get on the topic of their sports?
I mean they're talking about their training,
JIM yeah
BOB then it's kinda like a hot topic for both of them
JIM yeah
BOB he's like oh **it's starting to become like a genuine conversation**

Although it is not as directly observable as some of the other categories (such as embodiment or attention to the timer), a sense of whether or not the talk is authentic can be an impressionistic starting point from which raters can then ask themselves, "What is it in the talk that gives me that impression?" In addition, if a teacher were to grade a considerable number of this sort of peer-peer discussion test, they would inevitably develop an awareness of the authenticity within the test-takers' interaction, and be attentive to it in their grading, whether at a specific or general level. By the same token, this is something that teachers could then inform students of, encouraging them to focus more on the content of their partner's talk and advising them to display a genuine stance toward it in their reaction.

4.3 Misinterpreting the rubric

As shown in 4.2, the teacher-raters' hermeneutic dialogues provided plenty of evidence to suggest that they were able to understand the notion of engagement from the descriptive exemplar-based rubrics and could identify instances of it within other video samples of EFL discussion tests. In this sense, the rubrics seem to have been effective.

However, at one point it appears that one of the raters also misinterpreted the notion of engagement, at least as it was intended by the researcher and formulated within the rubric. This misunderstanding was in relation to topic shift within the talk. Even though this was not something that the rubric descriptors touched on, on several occasions Jim felt that test-takers who moved the topic away from what was originally written on the topic card should be penalized in terms of engagement, as shown in some of his comments below:

JIM yeah, I wrote many things- I wrote a lot for these two
but the negative thing that I wrote was that they like start
talking about something different

JIM and then I wrote the question does it matter if they're engaged
about something that isn't the topic?

JIM and then also, is it still engagement if they're engaged in a topic
that they weren't randomly given by a piece of paper at the start?

JIM well I did write a big question mark next to it where like, y'know,
how much is that still the topic of myself? Could it be sports?

As far as the developer of the rubric was concerned, topic shift within the conversation did not pose any problem, so long as it occurred in a step-wise fashion (Jefferson, 1993). In fact, moving away from the initial topic might even be considered as evidence of deeper engagement, so long as it is done in a genuine way, since it is likely to have evolved naturally from the speakers' reactions and post-expansions of earlier talk. As Bob told Jim in response to his final point above, "Yeah I didn't worry about them going off topic. I was more worried that now they're actually engaged because they're interested."

The intent of the test is therefore more in line with Bob's position than Jim's on this matter. At one point the rubric does specifically mention topic shift: at the D level it states, "A does not take this opportunity to shift the talk from prepared monologues to a more natural back-and-forth conversation". However, this would imply that topic shift was a normal and expected part of interaction and therefore acceptable within this test. It may have been that the reason for Jim's misunderstanding of the rubric here were due to his lack of familiarity with the test parameters, but it is curious that he related it to engagement. Another explanation might be that Jim interpreted engagement in term of engaging with the assigned topic, rather than engaging with the interlocutor and the emergent interaction. This could be one aspect of the descriptor that requires further clarification in the future.

4.4 Questioning the rubric in relation to the test format

In a similar vein, Bob brought up a point in relation to co-construction that was perhaps more critical of the test setting than of the rubric itself. He said, "Again, I think it's to do with the combination of, the pairing as well. You'd get different responses if they were with different people, for sure." In fact, this is an argument that has received some attention in the literature on peer-peer testing as well (Brooks, 2009; Galaczi, 2008; Philp *et al.*, 2013), and it is difficult to argue that the outcomes of the test are completely unaffected by the ability of a test-taker's interlocutor. With regard to engagement, for example, a partner who did not actively pursue uptake or on-topic talk may cause the interaction to become asymmetrical when it might have been more collaborative with a different person. However, again, this is not so much a criticism of the rubric so much as it is an uncertainty regarding the test format itself.

How best to incorporate this co-construction into the rubric remains a topic for further research. However, Brooks (2009) has suggested that peer pairs produced more interactionally demanding and complex than when the same students were paired with an L1 speaker, and there is no denying that the paired test format has other practical advantages in low-stakes contexts, such as processing the participants in half the time that an interview test would take.

5. Concluding Discussion

This study has explored the potential of a descriptive rubric for a paired EFL discussion test, particularly as a tool for enabling other raters to grade the test in regard to the notion of interactional engagement. Rather than just the single-sentence descriptor that is given to the students in the simplified rubric, the version of the rubric for the teacher-raters was developed using conversational analysis of the selected video-recordings of actual student interaction. These interactions in turn became exemplars for the raters, enabling them to compare the exemplars with other video recordings from the test corpus.

Overall, the inter-rater hermeneutic dialogues demonstrated that the raters were able to operationalize the notion of engagement via the test rubric along the lines anticipated by the test developer, despite the fact that neither of the raters had any particular background in CA. The descriptors were designed to include a balance of CA terminology and non-technical expressions, in order to be both manageable and instructive, and in this respect they appear to have been successful. Although engagement itself is not a term that either of the teacher-raters used to any great extent in their own classroom assessment, they were able to make sense of it by reading the deeply descriptive rubrics in conjunction with the video clips of the test-taker interaction. This suggests that engagement is indeed testable as an element of interactional competence in the Japanese EFL context.

However, one issue that remains is the question of how engagement differs from interactional competence (IC) (Hall *et al.*, 2011) or intersubjectivity. Without a doubt, these concepts do possess a great deal of overlap. Within the peer-peer test context, Galaczi (2014, p. 559) operationalizes IC in terms of topic development organization (degree, extension), listener support moves (backchanneling, confirming comprehension), turn-taking management and embodied features of talk. Meanwhile, within CA, intersubjectivity is more commonly discussed, since the focus there is more firmly grounded in how people understand each other than on how language learners demonstrate their proficiency. Goodwin and Duranti (1992) note that "(i)n order for separate individuals to engage in coordinated social action they must recognize in common what activities are in progress and what those present must do to perform the activity. The central question of intersubjectivity (how separate individuals are able to know or act within a common world) is thus raised as a constitutive feature of social action" (p. 27).

As a practical matter, however, language teaching professionals who are not specialized in Conversation Analysis require an accessible repertoire of language and concepts to use in testing their students' ability, beyond the holy trinity of accuracy, fluency and complexity. Certainly there is a place for debate among CA-SLA researchers about the nature of what comprises interactional competence, but there is also a need for non-specialist teachers to use such notions for practical pedagogical purposes. It is hoped this study has helped bridge the gap between theory and practice, at least within the context of one low-stakes language test. One logical future direction for the study would be to see how the exemplar-based rubric might be used as a form of learner-oriented assessment (Joo, 2016) and even become the basis for pedagogical activities prior to the assessment.

As a small-scale initial exploration into this topic, however, it holds limited generalizability. The raters were both L1 speakers of English with considerable experience in teaching EFL at the tertiary level in Japan, and undoubtedly this familiarity with the sort of test-takers featured in the videos informed their comments. While such teachers represented the rubric's target users, it would also be worthwhile to gather comparative data from other potential users, such as Japanese teachers of English or EFL teachers at other institutions. Likewise, the test-takers themselves were a particular sort of English user: first-year non-English majors at a national Japanese university. Further research with raters who teach in other contexts may prove fruitful.

Acknowledgement

This research was conducted via funding from JSPS Grant-in-Aid 17K03011.

References

Antaki, C. (Ed.). (2011). *Applied conversation analysis: Intervention and change in institutional talk*. Springer.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26*. 341–66.

Cho, J. Y., & Lee, E. H. (2014). Reducing confusion about grounded theory and qualitative content analysis: Similarities and differences. *The Qualitative Report*, *19*(32), 1-20.

Elo, S., Kääriäinen, M., Kanste, O., Pölkki, T., Utriainen, K., & Kyngäs, H. (2014). Qualitative content analysis: A focus on trustworthiness. *SAGE Open*, *4*(1), 2158244014522633.

Firth, A., & Wagner, J. (1997). On discourse, communication, and (some) fundamental concepts in SLA research. *The Modern Language Journal, 81,* 285-300.

Galaczi, E. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, *5*(2) 89-119.

Galaczi, E. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics*, *35*(5), 553-574.

Gan, Z., Davison, C., & Hamp-Lyons, L. (2008). Topic negotiation in peer group oral assessment situations: A conversation analytic approach. *Applied Linguistics, 30*(3), 315-334.

Gardner, R. (1994). Conversation analysis: Some thoughts on its applicability to applied linguistics. *Australian Review of Applied Linguistics, 11,* 97-118.

Goodwin, C. (2018). *Co-operative action*. Cambridge University Press.

Goodwin, C., & Duranti, A. (1992). Rethinking context: An introduction. In A. Duranti & C. Goodwin. *Rethinking context: Language as an interactive phenomenon*, (pp.

1-42). Cambridge University Press.

Greer, T. & Potter, H. (2008). Turn-taking practices in multi-party EFL oral proficiency tests. *Journal of Applied Linguistics*, *5*(3), 297-320.

Greer, T. (2019). Closing up testing: Interactional orientation to a timer during a paired EFL proficiency test. In H. T. Nguyen & T. Malabarba (Eds.) *Conversation analytic perspectives on English language learning, teaching and testing in global contexts* (pp. 159-190). Multilingual Matters.

Hall, J. K., Hellermann, J., & Pekarek Doehler, S. (Eds.). (2011). *L2 interactional competence and development*. Multilingual Matters.

Jefferson, G. (1993). Caveat speaker: Preliminary notes on recipient topic-shift implicature. *Research on Language and Social Interaction*, *26*(1), 1-30.

Joo, S. H. (2016). Self-and peer-assessment of speaking. *Studies in Applied Linguistics and TESOL*, *16*(2), 68-83.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, *70*, 366-372.

Leyland, C., Greer, T., & Rettig-Miki, E. (2016). Dropping the devil's advocate: One novice language tester's shifting interactional practices across a series of speaking tests. *Classroom Discourse*, *7*(1), 85-107.

Philp, J., Adams, R., & Iwashita, N. (2013). *Peer interaction and second language learning*. Routledge.

Pekarek Doehler, S. (2010). Conceptual changes and methodological challenges: On language and learning from a conversation analytic perspective on SLA. In P. Seedhouse, S. Walsh, & C. Jenks (Eds.), *Conceptualising 'learning' in applied linguistics* (pp. 105-126). Palgrave Macmillan.

Markee, N. (1994). Toward an ethnomethodological respecification of second language acquisition studies. In E. Tarone, S. Gass & A. Cohen (Eds.), *Research Methodology in Second Language Acquisition*, (pp. 89-116). Routledge.

Markee, N., & Kasper, G. (2004). Classroom talks: An introduction. *The Modern Language Journal*, *88*(4), 491-500.

Nevile, M. (2015). The embodied turn in research on language and social interaction. *Research on Language and Social Interaction, 48*, 121–151

Okada, Y., & Greer, T. (2013). Pursuing a relevant response in oral proficiency interview role plays. In S. Ross & G. Kasper (Eds.). *Assessing second language pragmatics* (pp. 288-310). Palgrave Macmillan.

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis*. Cambridge University Press.

Sandlund, E., Sundqvist, P., & Nyroos, L. (2016). Testing L2 talk: A review of empirical studies on second-language oral proficiency testing. *Language and Linguistics Compass*, *10*(1), 14-29.

Seedhouse, P. (2013). Oral proficiency interviews as varieties of interaction. In S. Ross & G. Kasper (Eds.). *Assessing second language pragmatics* (pp. 199-219). Palgrave Macmillan.

Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher*, *27*(2), 4-13.

Sidnell, J., & Stivers, T. (Eds.). (2013). *The handbook of conversation analysis*. John Wiley & Sons.

Stivers, T. (2013). Sequence organization. In J. Sidnell, & T. Stivers. (Eds.). *The handbook of conversation analysis* (pp. 191-209). John Wiley & Sons.

Talandis, J. (2017). *How to test speaking skills in Japan*. Alma Publishing.

Walters, F. S. (2007). A conversation-analytic hermeneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing*, *24*(2), 155-183.

Yashima, T. (2002). Willingness to communicate in a second language: The Japanese EFL context. *The Modern Language Journal*, *86*(1), 54-66.

Young, R. F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in language learning and teaching* (pp. 426–443). Routledge.

Appendix

The conversation analytic rubric descriptors at each level, as given to the test raters

Grade	A (Exemplary)
Score	8/10 to 10/10
VideoSample 1 (KTOP 33)	

Descriptive analysis of the students' interactional engagement

Although they are by no means perfect at English, both of these students are highly engaged in the conversation. The turn-taking frequently transitions back and forth between the speakers and they base their contributions on what the other person has just said. Neither takes an extended turn at talk in which the other just passively listens. Instead they jointly take responsibility for the topic and help each other to negotiate meaning through short repairs and clarifications and by using gestures. They proffer follow-up questions on what their partner has just said and only change the topic when the conversation comes to a natural close. On occasions their uptake is only simple, consisting of non-lexical perturbations like "Orgh!" or "Huun", but even so it usually comes swiftly (and sometimes is comfortably overlapped) and is often followed by post-expansions that provide evidence to demonstrate how they understood their partner's contribution.

```
A      I- (0.3) often went to sannomi↓ya kobe
B      O::::h san[nomiya    ]
A                [to: play]
B      very fashionable
A      Ye::s
B      fa:[shionable    ]
A         [fashionable]
B      ni:ce place
```

They mostly focus their gaze on each other and their laughter suggests that they are enjoying the topic. They are so engaged in the talk that they do not monitor the timer, and even continue to bring the talk to a natural close in English after the timer chimes.

Grade B (Proficient)
Score 7/10
VideoSample 2 (KTOP 28)

Descriptive analysis of interactional engagement

Towards the start of the video the turns resemble a series of monologues with each person taking an extended turn to discuss their position, although the other person often provides brief uptake tokens and nods throughout the turn. On the whole, they do not develop the topic based on what they have just learned from the previous speaker, and instead either change the topic or throw it back at their partner by saying "And you?" This is the sort of engagement that would normally get them a C or even a D.

However, toward the second half of the test, we see this pattern start to change, and the speakers receipt each other's turns more enthusiastically and build on the topic with assessments and reciprocal questions in which the turn changes back and forth more frequently. They maintain eye contact well and seem to be genuinely reacting to the content of what the other person has just said. If they had interacted this way more consistently throughout the test, they may have achieved an A for engagement.

A how many children eh: do you want to have.
B uh:::h (x.x) I wan- |two:: two children?
|((two fingers up))
A >**oh two children**.<
B |girls and- a [girl and-]
|((one finger up))
A **[oh!]** |(**good-)**
|((pointing to B))
B hh ye(h)ah [hah hah hah]
A [(good!)] girls (and) [↑Boy]
B [>**yeah**] **yeah yeah**<
A yes.

Grade C (Needs improvement)
Score 6/10
VideoSample 3 (KTOP 38)

Descriptive analysis of interactional engagement

The turns in this test are not so long, the participants do not build on each other's topic in much detail, and there is often extended silence both within and between turns. For example, in the extract below we see that A asks a question. It takes B a while to formulate his response, but when he does finally get it out, A simply receipts it through repetition ("thirty years old") and the change-of-state token "oh". After that there is another extended silence in which A could have developed the topic further, but didn't, so B simply redirects the same question back to A, meaning they have missed an opportunity to engage more deeply.

A .hh m:::m whe:n do you want to (.) marriage.
B uh:: .(s)hh I want to ↑ma:rriage
(4.2)
A >(want to)< [marriage]
B [whe:::n] when I:’m: (3.3) thirty
years old
A $**thirty years old**$=
B =thir(h)ty years [old .hh]
A **[o : : h]**
(3.0)
B when do you (.) marriage?

At other times, A gives minimal receipts before changing the topic to something completely different without marking the transition in any way. Perhaps because B is struggling to formulate his sentences grammatically (or to reach an answer), he often looks away from A and scratches his eyes as he is thinking; this slows the pace of the talk and A's minimal responses appear disinterested in that they do not build significantly on what B has said in the prior turn. At 4:15, they glance at the clock and wind down their talk before it chimes.

Grade D (Unsatisfactory)
Score 0/10 to 5/10
VideoSample 4 (KTOP 35)

Descriptive analysis of interactional engagement

The turn-taking in this test could best be described as a series of parallel monologues. Each person takes a relatively long turn at talking while the other generally listens without much nodding or providing uptake tokens. After the "monologue" the recipient B gives a brief acknowledgement like laughter and "I think so (too)" and then after a long silence he shifts the topic to something unrelated via another extended monologue.

((at the close of an extended turn from A))
A ↓mm (0.5) but (0.6) I m:m: (0.7) my private is
(.) more important [zHAh hah]
B [hh hh] °(hah hah)°
I think (.) [so]
A [(I-heh heh)]
(0.8)
A mm-
(1.6)
B uh:::m (1.8) ma- I think (1.0) uh::: the:::
advantage of share housing is (1.6) m:::m (0.4)
ma-i- (0.6) a::h it's al-always (1.0)
((B goes on to give an extended turn on another topic))

These extended monologue turns and the fact that they do not build on what has just been said sometimes give the impression that they have been preparing their turn while the other person was speaking. They do not use a lot of gestures or facial expressions and the overall pace of the talk is slow. There is far less overlap and far more silence between turns in this clip. A does speak quicker than B, but the points he raises seem to be prepared prior to the test and therefore are not always related to what B has just said. When he does get some new information from B, such as at 4:18 when B tells him he has three friends who share a house, A does little more than say "oh" and gives a nod, leaving a gap of silence that forces B to continue speaking. In other words, A does not take this opportunity to shift the talk from prepared monologues to a more natural back-and-forth conversation.