How do raters understand rubrics for assessing L2 interactional engagement? A comparative study of CA- and non-CA-formulated performance descriptors

Erica Sandlund
Karlstad University, Sweden
Tim Greer
Kobe University, Japan

While paired student discussion tests in EFL contexts are often graded using rubrics with broad descriptors, an alternative approach constructs the rubric via extensive written descriptions of video-recorded exemplary cases at each performance level. With its long history of deeply descriptive observation of interaction, Conversation Analysis (CA) is one apt tool for constructing such exemplar-based rubrics; but to what extent are non-CA specialist teacher-raters able to interpret a CA analysis in order to assess the test?

This study explores this issue by comparing two paired EFL discussion tests that use exemplar-based rubrics, one written by a CA specialist and the other by EFL test constructors not specialized in CA. The complete dataset consists of test recordings (university-level Japanese learners of English, and secondary-level Swedish learners of English) and recordings of teacher-raters’ interaction. Our analysis focuses on ways experienced language educators perceive engagement while discussing their ratings of the video-recorded test talk in relation to the exemplars and descriptive rubrics. The study highlights differences in the way teacher-raters display their understanding of the notion of engagement within the tests, and demonstrates how CA rubrics can facilitate a more emically grounded assessment.

Key words: engagement; conversation analysis; paired discussion tests; interactional competence; English as a foreign language (EFL)

Introduction

Paired student discussion tests in English as a foreign language (EFL) contexts are commonly graded according to some form of rubric, sometimes prepared by the classroom teacher, sometimes by an external organization or educational authority.
Typically, such rubrics involve broad descriptors that have been simplified to fit idealized cases. An alternative approach, however, is to construct the rubric based on an extensive written description of video-recorded exemplary cases at each performance level. In this study, we are interested in one aspect of assessing interactional competence (IC), namely engagement in L2 paired tests, operationalized as an external, publicly available indication of a willingness to communicate (MacIntyre et al., 1998). While some high-stakes tests like the Cambridge English and Trinity tests explicitly measure IC, many interactional EFL tests are still assessed based on variations of a three-part rubric that includes accuracy, fluency and complexity. These categories refer to the test-takers’ linguistic proficiency but do not take into account the importance of collaboration, intersubjectivity and pragmatics in interaction (Weir, 2005).

Conversely, the test constructs in our study possess a strong interactional component, including features uncovered within Conversation Analysis (CA; Sidnell & Stivers, 2013), such as turn-taking, sequence and repair. Our data are taken from two distinct contexts (a low-stakes test at a Japanese university and a high-stakes test in Swedish middle schools), but our primary concern is not to compare the test-takers or their achievement. Instead we are interested in how the raters interpret the test-taker behavior as evidence of engagement based on the descriptive exemplar-based rubrics they have been given – one of which was written in a simplified CA-style. As such, our study is comparative primarily in terms of its focus on the raters' co-accomplished sense-making. Following a review of relevant literature and an outline of the test contexts, we will provide a detailed analysis of three of the categories identified in the rater discussions – post-expansion, parallelism, and naturalness – and demonstrate how the raters use these to assess the candidates' engagement. We conclude with a discussion of the efficacy of incorporating a CA-style analysis into a descriptive oral proficiency rubric.

**Literature review**

**Assessing L2 interactional competence: rater perspectives**

Data-driven approaches to the development of constructs, rating scales, and performance descriptors for interaction are strongly recommended in many interaction-oriented studies of language testing (Kley, 2019; Ross, 2018). However, in order for any rating scale to function consistently, we must also understand how raters interpret and apply them to actual test interactions, since “it is their view of interaction which finds its reflection in the test scores” (Ducasse & Brown, 2009, p. 425). In Ducasse and Brown’s study, raters watched video-recordings of tests and recorded how they perceived the interaction in a two-step process: first, recording their general

https://www.trinitycollege.com/qualifications/english-language/GESE
impression of the interaction and then their comments about specific features (i.e., a think-aloud protocol design). The authors coded the raters’ responses, yielding three categories salient to the raters: *non-verbal interpersonal communication* (gaze and body language), *interactive listening* (supportive and comprehension-oriented), and *interactional management* (topics and turn-taking). This data-driven approach to construct development therefore incorporates elements that raters identify as interactionally competent conduct.

A recent investigation by May et al. (2019) likewise relied on rater feedback to develop a learning-oriented tool for assessing IC in a high-stakes paired test. The researchers created a checklist and materials based on stimulated recall comments from examiners who viewed video-recordings of the test. These were trialed by four teachers and after analysis of further focus group discussion with them, the researchers came up with two IC checklists; a detailed list aimed at testers and a concise one with four categories and brief descriptors for classroom teachers, such as (1) initiates new ideas, (2) keeps the discussion going over several turns, (3) negotiates toward an outcome, and (4) uses body language appropriately. May (2011) likewise recognizes the relevance of non-vocal action: in her study, four raters trained to assess paired interaction in an EAP speaking test viewed embodied interaction as an essential component of the test-takers’ ability to respond appropriately, listing features such as “demonstrates effective body language, including eye contact, facial expressions and gestures that indicate a genuine desire to communicate with partner” (p. 136). The raters also noted “authentic interaction” in their commentary, where features characterizing authenticity included “a flowing discussion and inclusive and cooperative partners” and using descriptive terms such as “natural”, “genuine”, and “authentic” for high-scoring learners (p. 137).

In her study of raters assessing paired tests against a national standard at Swedish upper secondary school level, Borger (2019) examined raters’ written notes, focusing on features of IC that stood out to the raters. Three main interactional resources used by test-takers were seen as indicative of IC: topic development moves, turn-taking management, and interactive listening strategies. In addition, raters also considered the test-takers’ respective interactional roles in assessment decisions. Borger concluded that raters attended to features both within and beyond the assessment guidelines, which implies that guidelines for teachers-as-raters “have to be elaborated, including conceptually grounded reasoning as well as commented examples” (p. 167).

Other studies deal with the development of CA-informed performance descriptors for interaction. Example include Youn’s (2015) analysis of raters’ application of data-driven rating criteria on role-play assessments and Walters’ (2007) study of CA-trained raters’ assessment of L2 pragmatic competence, which included post-rating “hermeneutic dialogues” (p. 169) between raters; i.e., “a series of dialogues regarding differentially rated performances” (p. 172). These dialogues aimed primarily to
uncover their grounds for assessment and to resolve rating differences. In contrast to Walters’ raters, who had some training in CA transcription and analysis, the present study focuses on raters with no CA training.

**Learner engagement**

Over the past decade or so, interactional competence has been an area of increasing interest for both testers (Galaczi & Taylor, 2018) and CA-SLA researchers (Hall et al., 2011). Our study views engagement as a subset of IC that focuses particularly on the speakers’ interest in the topic as demonstrated via the interaction itself. Our analysis restricts itself to the behavioral and social dimensions of engagement, touching on cognition and emotion only to the extent that they are made accessible via the micro-details of social interaction by interactants themselves. In the following review of engagement in classroom discourse, we restrict our discussion to studies adopting such an emic stance to engagement. This view, in turn, also entails the understanding that engagement is observable for raters in their practical work of assessing L2 tests.

In classroom contexts, CA researchers have explored engagement in terms of keeping students “on task”. Waring and Hruska (2011), for example, document the interactional practices a teacher uses to manage student resistance and keep the lesson moving forward, while Leyland et al. (2016) analyze the way a teacher assistant abandons one interactional strategy (“playing the devil’s advocate”) in favor of another (agreeing with the group) in response to students’ disengagement from the discussion. Stokoe et al. (2013) see engagement in student discussions in terms of how the participants align with each other’s comments and make relevant their displays of affiliation. Finally, Greer (2016) considers engagement in terms of student *initiative*, examining how one student goes beyond the teacher-assigned interactional task by deploying self-selected post-expansion sequences that allow him to practice and perfect language formulations peripheral to the task-as-designed. Engagement, therefore, is not only an emotional or reflective state; it is also something that can be instantiated in classroom interaction – in real time by the participants themselves, and retrospectively via video-recordings of the talk.

**Engagement in L2 testing**

Perhaps due to the necessity to evaluate talk and conduct, it is in research on speaking assessment that the notion of interactional engagement has been most closely specified. Galaczi (2008) identifies interactional engagement in such learner actions as “rapid speaker changes and supportive overlaps and latches” (p. 567). She notes that more engaged listeners are able to respond to the speaker’s just-prior turn rather than preparing a longer monologue-like response to be delivered at a sequentially disjunctive point, a practice that she labels parallelism. Similarly Lam (2018) stresses the importance of “producing responses contingent on previous speakers’
contributions” (p. 377). A test-taker who is preparing a turn while their partner is talking is unlikely to be paying complete attention to the turn-in-progress. Gan (2010) found that while high-scoring learners demonstrated a range of social actions for engaging with each other’s ideas, lower-scoring learners structured their interaction mainly according to the discussion prompts rather than on immediately prior talk.

Interactive listening, such as when speakers provide uptake tokens and listener responses at appropriate points during their partner’s turn, is also cited by test-raters as evidence of active engagement in the discussion (e.g., Ducasse & Brown, 2009; May, 2011). Ross (2018) views such listener responses as an interactional means of maintaining engagement with the speaker, and therefore suggests they could provide evidence of the test-taker’s interactional competence.

Unlike linguistically based categories, students’ engagement level therefore ascertains how active learners are during the conversation. Such interactional resources include turn-taking, sequence, and repair, as well as the sorts of constructs that Young (2009, 2011) sees as integral to interactional competence: participation frameworks, register, modes of meaning and boundaries. Given various possible displays of interactional engagement that raters may perceive in their assessment of learner test talk, research operationalizing engagement for construct development is timely and necessary. Furthermore, the success of attempts at operationalizing engagement for assessment are, naturally, dependent on how well they work in terms of guiding raters toward particular understandings and applications. The present study addresses how teachers-as-raters perceive aspects of engagement in paired EFL speaking tests, using scoring rubrics assigned for the particular test they are rating. Following Ducasse and Brown’s (2009) call for further research on rater understandings of interaction, and with the additional aim of contributing to the growing body of work addressing the assessment of interactional competence in L2 language testing, our study explores ways that raters perceive engagement in recordings of L2 test talk in two distinct contexts. The contexts are distinct mainly in the sense that the raters work from different types of scoring rubrics, but also distinct because the tests assessed are drawn from different contexts.

**Data and methods**

In the present study, we focus on raters’ perceptions of engagement when applying scoring rubrics in L2 paired discussion tests, with a focal interest on whether raters presented with CA-grounded exemplars are able to identify features of engagement differently to raters presented with more traditional assessment criteria. In order to compare and contrast raters’ understandings of interactional engagement, we draw on two different EFL speaking tests and their accompanying assessment instructions – a university EFL course test for freshmen in Japan (KTOP), and the National English
Speaking Test (NEST) in Sweden. Authentic test recordings from the two tests were used in creating our second empirical dataset – video recordings of rater discussions about specific test recordings, from each test context. We begin with a description of the two testing contexts, continue with the test recording datasets, and finally, account for the collection of the rater discussion dialogues.

The tests

While they undoubtedly have their differences, the tests the raters are discussing do also have much in common, in that they are both relatively open-ended, peer-to-peer discussion tests of novice language users in EFL contexts.

The Kobe Test of Oral Proficiency (KTOP) is a simple discussion test used as part of a 15-week EFL course among first-year students at a national Japanese university. Students randomly chose one out of six topic cards (yourself, extended family, marriage, travel, share-housing and jobs) and then talked freely in English about that topic for four minutes. As such, each pair structured their talk around one of the topics. The complete dataset consists of 93 pairs, but for the current study, four tests were selected as exemplars at each of the assignable grades (A to D) and the video recordings of a further eight pairs were given to the raters to evaluate and discuss. Following Walters' (2007) work on inter-rater hermeneutic dialogues, the second author produced a detailed rubric to outline what engagement involves at each level by using videos from the KTOP dataset as exemplars. These exemplars (Appendix 1a-d) consisted of a link to the video recording, a graded score and an extended descriptive analysis of each test recording, specifically in relation to interactional engagement. The descriptions were written by an EFL teacher who is also a CA researcher (the second author) based on selected focal segments of the recordings that illustrated significant features of engagement. Although they were informed by CA, they intentionally incorporated less specialized terminology.

The second dataset comes from the spoken section of the Swedish National Agency for Education’s mandatory 9th grade National English Standardized Test (NEST). This is a criterion-referenced, standardized test; however, unlike many international standardized tests, national tests in Sweden are usually graded by the students’ own teacher (see Borger, 2019; Sundqvist et al., 2018). Learners sit in pairs or small groups, and the teacher is present. The 15-to-25-minute test consists of an initial warm-up task (a picture prompt or short narratives about familiar topics) and a discussion section based on a number of topic cards containing statements and questions. Students take turns pulling a new topic card, and continue until the teacher/examiner rounds off the test talk, usually after 10-15 minutes. Statements such as “Everybody should be a vegetarian” are followed by prompts such as “Agree? Disagree? Explain why and give examples. What do your friends think?” The grading criteria are largely aligned with the descriptors of the communicative abilities in the Common European Framework
of Reference for Languages (CEFR) (Council of Europe, 2001). The grading scale is from A to E, where E is the lowest passing grade, and F is a failing grade.

The NEST rater discussions were based on one selected test recording, featuring two 9th graders (Fred and Henrik, pseudonyms). The test recording was collected as part of a project on interaction in L2 tests, Testing Talk, funded by The Swedish Research Council [id 2012-4129]. Data were collected at four schools, encompassing 161 learners in a total of 71 audio-recorded tests. All learners had been assessed by their own teacher (present during test) and by two external raters (using audio-recordings, for research purposes only). The learner test underlying the rater data presented in this paper was selected from the corpus because of its relative typicality (e.g., in terms of duration, performance quality, and lack of any explicit test disruptions), and because there was a grading discrepancy between at least two of the three assessors used in the original study (for a detailed description of the procedure for selecting the test recording, see Sandlund & Sundqvist, 2019). Data for the present study was collected as part of a spin-off project to Testing Talk, focusing on in-service teachers’ assessment of the NEST\(^3\). A recording from the test corpus was used in a training workshop where 11 English teachers (henceforth, the NEST teachers) were divided into four groups in which they rated and discussed their ratings of two learners. All four groups were video-recorded as they discussed their individual assessments, and agreed on a grade for each test-taker. The NEST teachers were provided with the assessment materials, which accompanies the test from the Swedish National Agency for Education, consisting of guidelines, national standards, and audio-recorded benchmarked exemplars, along with descriptive texts for the benchmarks at performance levels from A to E. An overview of the full National Test of English, where the NEST is one part (the remaining two focus on receptive skills and writing, respectively), describes the NEST as a test of “free oral interaction and production” where assessment has a focus on “the willingness and ability to participate in the conversation and to formulate content in a comprehensible manner” (Swedish National Agency for Education, 2014, Teacher information, title page, translated). The same page lists “assessment factors” (Sw. bedömningsfaktorer) under three headings: Willingness and ability to interact and speak, Content, and Language, with descriptive bullet points under each (see Appendix 2). In the booklet accompanying the test materials, the assessment factors (or “generic analytic assessment factors”, see Borger, 2019, p. 156) are also specified in relation to the knowledge requirements in the syllabus for English in Swedish compulsory school, but on this page, the assessment factors are listed under two headings instead of three: Content and Language and expression with bullet point descriptors under each (see Appendix 3). Here, the Content category also lists “complexity and variation – different examples and perspectives”, whereas the Language category contains “communicative strategies”, such as “to develop and carry the conversation forward” and “to solve linguistic problems through reformulations, explanations, and

\(^3\) Funded by the Center for Language and Literature in Education, Karlstad University, Sweden.
clarifications”. These assessment factors are further exemplified in the descriptive texts on the three benchmark sample tests. By way of example, at the benchmark for a B grade, the student is described as showing initiative by asking questions such as “what do you think”, and interacting by responding to her co-interactant’s utterances with turns like “I also think that...” (Swedish National Agency for Education, 2014, p. 21).

Although the test-takers, contexts and conditions of the two tests do have substantial differences, the focus of our study is not on the tests as much as it is on the raters and how they interpret the test-takers' interaction in terms of the rubric they have been given. We are concerned with how experienced teacher-raters conceptualize interactional conduct in ways that relate to the notion of engagement when assessing interaction using the NEST assessment factors. Since the NEST describes interactional conduct in rather general terms, the Swedish dataset constitutes a suitable contrastive sample to the CA-grounded rubrics used in the KTOP tests.

Rater discussions

As a low-stakes in-class assessment tool, the KTOP test is normally graded by the classroom teacher; however, as part of an investigation into the development of a rubric for assessing engagement [JSPS Grant-in-aid 17K03011], two additional teachers were asked to rate video samples from the KTOP test according to the pilot exemplars. The raters (Jim and Bob, pseudonyms) were both L1 speakers of English and taught in the same program. They were veteran EFL teachers with extensive experience within the Japanese university system and had each assessed Japanese EFL students in oral communication tests of their own, although not using the current rubric. The raters had not specifically used engagement as a criterion in their own tests, but they seemed to understand it upon reading the descriptive rubric and watching each of the exemplar videos. Neither rater was familiar with CA transcription or analysis. They were then given eight additional videos from the collection to grade, and rated each test-taker individually by making notes as they watched. They were then asked to compare the grades they had assigned with those of the other rater and to discuss any reasons or differences of opinion. This inter-rater discussion was video-recorded for the present study.

The NEST rater data was collected as part of a research project on collaborative assessment of oral proficiency and interaction, funded by the Center for Language and Literature in Education at Karlstad University. Data consists of four video-recordings of L2 English teachers discussing and grading the selected test recording (Sandlund & Sundqvist, 2019). Thirteen English teachers, all with at least four years of teaching experience, participated in the assessment workshop. The teachers had not received any formal training on interaction or IC prior to the workshop event. After an introduction where all teachers listened to the selected recording and made individual assessment notes, they were divided into groups and instructed to discuss and grade
the two learners together. They were also provided with the NEST assessment materials, including tasks and assessment instructions. All four groups were video-recorded, and rater meetings lasted around one hour. The groups were instructed to assess the test as they would normally do, and did not receive instructions to pay particular attention to specific interactional features.

**Analytic approach**

In order to examine the raters’ perceptions of interactional engagement, we approached the rater data through a process of qualitative content analysis (Cho & Lee et al., 2014; Elo et al., 2014; Silverman, 2000). Starting with the KTOP rater dataset, we searched for recurrent themes related to engagement, particularly with regard to how the raters used the rubric descriptors to identify engagement in the video recordings they rated. We later compared this to the NEST rater discussions, which used a more traditional rubric with no explicit mention of engagement.

In this paper, we will focus on three of the themes; post-expansion, parallelism and naturalness, which were salient in the data. These themes were identified through an extended review of transcripts of the discussion between the two KTOP raters. After the raters’ discussions were transcribed, they were inductively coded for repeated ideas and concepts, namely the ways the raters referred to and operationalized the notion of engagement within the video and audio data they were reviewing. Segments of their conversations were then sorted and grouped according to the emergent sub-themes shown in Figure 1 within each dataset, which were later grouped according to the three over-arching categories of post-expansion, parallelism and naturalness. In some cases, certain sub-themes were applicable to several categories, such as turn-taking. This qualitative analysis later facilitated the cross-contextual comparative analysis that is presented in the next section. As such, the themes in the KTOP rater discussions guided the empirical searches of the NEST rater discussions.

---

4 For further detailed analysis of the KTOP rater discussions, see Greer (2020).

5 The KTOP rater dataset was transcribed for content only, but since the NEST data had already been transcribed according to Jeffersonian conventions (Jefferson, 2004), it was decided to leave them in that form.
As the NEST rater groups featured 3-4 raters each, the content analysis often included sequences where observations of rater impressions were coded across several turns. We have chosen to include responses from co-raters, without aiming for a sequential analysis of the rater conversation (for such an approach, see Sandlund & Sundqvist, 2019).

One difference between the rater datasets is that the NEST raters only had access to the audio-recording, whereas the KTOP raters were supplied with videos. When teachers assess the NEST, they have usually also been present during the test, which offers situated access to the non-vocal conduct of test participants. Given our interest in engagement, which could encompass vocal, interactional, and embodied conduct, this is a limitation for the comparative analysis.

Another difference between the test datasets is the length of the test recordings. While the KTOP tests had a set stop time of four minutes, indicated by the use of a timer, there is no fixed duration of the NEST – instead, the teacher rounds off the test interaction when sufficient output has been generated, or when a particular number of topic cards have been used. The test selected for rater discussions was about 21 minutes, which was average in the corpus of 71 tests. While the KTOP raters occasionally oriented to timing in their discussions (Greer, 2019), this was a non-issue for the NEST raters, and this aspect was therefore excluded from our comparison.

Finally, the KTOP raters discussed eight tests, whereas the four NEST groups all discussed one selected test. Any comparisons between conduct displayed across
different tests in determining grades on performance was therefore not possible for the NEST raters.

**Analysis**

Below, we present three of the most prominent categories identified within the rater discussions: *post-expansion, parallelism, and naturalness*. The analytic section of the paper is structured as follows: The KTOP raters’ identification of interactional conduct under each category, following the CA-grounded rubric, is presented first. After that, their constructions of engagement are contrasted with constructions by raters using performance descriptors not grounded in a CA analysis of interaction data.

**Post expansion**

One aspect of interactional competence that the raters recognized as related to engagement was *post-expansion*, or the ability to extend on a just-prior turn through relevant follow-up talk. In conversation analytic research, post-expansion refers to a turn or an adjacency pair that comes after, but is still connected to, a base adjacency pair (cf. Schegloff, 2007, pp. 115–168). Minimal post-expansions, also known as sequence-closing thirds, often consist of utterances such as *oh, I see,* or *okay* and occur as single turns after the base second pair part but do not project any further talk beyond it, while non-minimal post-expansions are designed to go beyond one turn by constituting a new first pair part, projecting a responsive second pair part (Schegloff, 2007, p. 149). Such expansions may consist of, for example, disagreement-implicative other-initiated repair, topicalization through repeats and partial repeats, and overt disagreement turns (pp. 151–162). Greer (2016) has shown how post-expansions play a role in learner initiative, demonstrating a learner’s willingness to engage with the interlocutor and extend the conversation beyond the bounds of the assigned task. In the context of L2 testing, Lam (2018) argues that “producing responses contingent on previous speaker contribution” (p. 381) in speaking assessments could be included in the construct for IC, and Galaczi (2008) also notes how the most successful candidates managed to relate their contributions to what their co-participant had said and also add something new. In the case of the raters’ perspective in the present study, we were interested in seeing whether the CA-based exemplars would help the KTOP raters see post-engagement-type conduct differently than the NEST raters.

*Post-expansion: KTOP raters*

In our first rater discussion segment, KTOP rater Bob negatively assesses one of the test-takers for not providing a fuller response:
Extract 1. “They could have gone somewhere with that” [KTOP 5]

01 BOB but little things like
02 B asked A how many people
03 would you like to share with
04 he says, “two”
05 (0.9)
06 that’s the end of it, y’know?
07 JIM pfff
08 BOB they coulda gone somewhere with that.
09 JIM yeah yeah mm

In lines 2-4, Bob provides a summary of one segment of the video-recording, using reported speech to illustrate how Speaker B asks Speaker A a question, which yields only a single-word answer from Speaker A. Bob appeals to Jim by saying “that’s the end of it” and then going on to suggest the learner “could have gone somewhere with that”. This interactional assessment of A’s response demonstrates that Bob sees this as a slot in which A could have provided a fuller answer, such as an account for why he would like to share a house with two people. Such a response would constitute a self-selected post-expansion, and on noticing its absence, B could have post-expanded by requesting such an account. Because the test-takers did not extend the talk in this way, the rater views them as less engaged in the conversation. The implication is that the test-takers are not offering sufficient follow-up detail.

In contrast, in Extract 2, Bob rates another pair favorably for "building on each other’s responses" (lines 2-3).

Extract 2. “Building on each other’s responses” [KTOP13]

01 VID "yes, I like udon"
02 BOB they’re definitely building on
03 each other’s responses
04 JIM yeah they are

Throughout this video, the test-takers are actively initiating follow-up questions and post-expanding their responses with additional information. They are listening to each other’s utterances and attempting to provide on-topic contributions. In line 4 Jim voices his agreement with Bob’s assertion, suggesting that they both view post-expansion as evidence of interactional engagement.

In this way, the KTOP raters referenced the descriptors, both tacitly and explicitly. Formulations like "build on" (Ex 2, line 2) also appear in the rubric (C-level), and there were other occasions when the raters mentioned the rubrics specifically.
In Extract 3, for example, Jim refers to the A-level rubric in line 4, displaying that he had not previously considered the idea that the participants’ orientation toward the timer could provide evidence of their engagement level. Bob agrees, reformulating it in the form of a proverb (“time flies”, lines 11-12), to which Jim agrees and assesses the point positively (lines 13 and 15). Clearly, the raters are both referencing the rubrics in their talk and also refining their observational skills along the way.

Post-expansion: NEST raters

In the NEST assessment instructions (see Appendix 2), the descriptors closest to encompassing conduct of the post-expansion type is “communicative strategies for developing and advancing the conversation”, although some of the others, such as “adaptation to the purpose, recipient, and situation” could also cover post-expansive actions. In Extract 4, the raters have been collaboratively critiquing the two test-takers’ interaction (rather than their linguistic proficiency). Marcus (line 135) exemplifies the substandard interaction by stating that the learners do not “question” each other – if one student has expressed a stance on something, the other seems expected to challenge this perspective – something that could possibly come in the form of displaying disagreement through non-minimal post-expansion. In the trajectory that follows, Benitha sums up the gist of the deficiency identified: there is not much interaction between Fred and Henrik:
As such, Marcus’ formulation of how the learners do not ‘question’ each other is paired with a general observation about the test-takers’ quality of interaction, and treated as one piece of evidence of an interactional skill that neither test-taker is displaying. Worth noting, also, is that ‘questioning’ each other’s talk (the Swedish verb *ifrågasätta* is semantically tilted towards ‘call into question’, or ‘express doubt about’ rather than ‘asking questions’) implicitly indicates that a good test interaction implies providing different perspectives on current topical talk, even displaying disagreement (cf. Hüttner, 2014), which in turn may relate to one of the other assessment factors, listed under the *Content* heading: *Complexity and variation – different examples and perspectives* (see Appendix 2). As NEST raters are instructed to assess holistically, several of the assessment factors may be considered simultaneously. In this specific case, it appears as if engaging with an interlocutor’s prior talk is combined with a linguistically and topic management-oriented assessment factor – the display of different perspectives, although the descriptors are not cited explicitly.

In Extract 5 below, another group is discussing an instance where one of the test-takers, Henrik, did expand on his co-participant’s talk. May begins by identifying the instance (lines 10-11, 13) where the topic was what they would do if a friend were using drugs. May then reports on Henrik’s turn at talk in line 16, after which she provides an

---

account for why Henrik’s contribution was exemplified: it is evidence that Henrik “did try” (lines 10, 20), “at least once” (line 23):

Extract 5. “you can at least try” [CASS1: 10-24]

10 MAY Henrik försökte ju där e:h någon gång e:h
Henrik did try there e:h at some point e:h
11 när Fred fick frågan om han skulle hjälpa
when Fred was asked if he would help
12 SAR a:
Yes:
13 MAY e:h säga prata med kompisar [om] droger
e:h say talk to friends [about] drugs
14 BRI [precis]
[exactly]
15 TER m:
16 MAY you can at least try
17 SAR m:
18 MAY tyckte Henrik då
said Henrik then
19 TER ja ja
Yeah yeah
20 MAY e::h så han försökte ju
E::h so he did try
21 SAR [m::]
22 TER [m::]
23 MAY åtminstone en gång
at least once
24 SAR| m:

In contrast to Extract 4, where raters focused on the lack of topical elaboration, raters in Extract 5 bring up an instance where a test-taker offered (unprompted) follow-up talk to something the other test-taker had said. In this case, Fred had mentioned that it would be difficult to talk to a friend about their drug use, and Henrik added that you can at least try to talk to the said friend. What May is pinpointing is an instance of IC in the form of an action that displayed understanding of prior talk (Sert, 2019; Burch & Kley, this issue) and an action that showed some engagement with Fred’s topical talk.

In sum, the NEST raters do pick up on displays of IC that relate to post-expansion; however, in all instances located in the rater data, it appears as if post-expansion type conduct is salient to raters when specifically related to something like opinion discourse (Shi-xu, 2000); that is, when test-takers expand each other’s talk on assigned topics. This may partly be because it is easier to recall examples of conduct when linked to specific topics, but it may also be linked to the NEST descriptors, which emphasize developing topical content –Similarly, Galaczi (2014, p. 565) observed that
“topic extension moves of other-initiated topics” were found in the interactions of high-scoring test-takers, which indicates that the ability to extend other-initiated topics “signal a speaker who is orienting to an active role in the development of topics and provide evidence of linguistic proficiency.” This view was seen in the NEST rater data as well. Also, in the written comments to the benchmarked test recordings from the test developers, the expression of opinions with arguments supporting those opinions is emphasized for the higher grades (C, B, and A). As such, when applying the NEST rubrics, post-expansion conduct tends to become a matter of post-expansion through (partial) disagreement or the proposing of a different topical angle than the prior speaker.

**Parallelism**

As outlined above, Galaczi (2008) noted that engaged listeners are able to respond to the speaker’s current turn rather than preparing a longer monologue-like response to be delivered at a point that may be sequentially disjunctive. This practice is referred to as parallelism. Raters in the present study also rated parallelism as evidence of a lack of interactional engagement in similar ways to Galaczi’s description of parallel interaction: a form of disaligned “dual solo” talk in which “both speakers initiated and developed topics...but engaged little with each others’ ideas” (p. 102). As described below in relation to each dataset, the raters in our study see parallel interaction as evidence of the test-takers’ lack of engagement with their interlocutor.

**Parallelism – KTOP raters**

At the lower levels of the KTOP dataset, this pattern is recognizable when one test-taker takes a long turn and the next speaker’s turn is not directly related to what the previous speaker has just said. This was the case in the video Bob and Jim are watching at the point we join them in Extract 6.
Bob places part of the blame for B’s lack of engagement on the prior speaker, because A produces an extended turn at talk but “he never asks the other guy what do you think about it” (lines 1-5), adding, ”he doesn’t really consider B’s position” (lines 6-7). These observations constitute an accurate summary of parallelism within this section of the data; A’s contributions are not directly related to B’s prior talk and seem as though they may have been prepared prior to the test, or while B was talking (See Appendix 4, Test Fragment 1). They are long turns, giving a sample of the student’s ability level that could be useful in assessing some of the other criteria, or as Bob puts it in lines 11-13, ”he just spat out a bunch of English in there (so that) he can get graded on how much he said”. His turns do not provide evidence of active listening or attention to what his partner has been saying. Bob notes that toward the end of the test when each speaker has completed their extended turns, they appear at a loss for something further to discuss, and only then does A ask some questions that make the discussion more interactive.

In this way, it seems that the KTOP raters have identified episodes of parallelism within the dataset and appropriately linked them to the criteria of engagement, marking them down to a lack of engagement. Likewise, in Extract 7, Bob detected another example of parallel talk, even though the recipient was more actively providing uptake through repetition.
Extract 7. “He’s not really responding to her”

<table>
<thead>
<tr>
<th>Line</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>BOB  yeah, he’s kinda,</td>
</tr>
<tr>
<td>02</td>
<td>he’s not really responding to her</td>
</tr>
<tr>
<td>03</td>
<td>he’s just waiting for his chance to speak</td>
</tr>
<tr>
<td>04</td>
<td>JIM  right</td>
</tr>
<tr>
<td>05</td>
<td>BOB  so like she was saying,</td>
</tr>
<tr>
<td>06</td>
<td>her father works in a bank</td>
</tr>
<tr>
<td>07</td>
<td>the guy went &quot;bank, very busy, very busy&quot;</td>
</tr>
<tr>
<td>08</td>
<td>&quot;so hard so hard&quot;.</td>
</tr>
<tr>
<td>09</td>
<td>He’s just repeating what she says</td>
</tr>
<tr>
<td>10</td>
<td>and then waiting for her to finish</td>
</tr>
<tr>
<td>11</td>
<td>so he can talk about his mum.</td>
</tr>
</tbody>
</table>

The test talk the raters are discussing here is shown in Test Fragment 2 (Appendix 4).

Although the secondary speaker has been providing uptake tokens such as incremental acknowledgements (Iwasaki, 2009) and receipt through repetition (Greer et al, 2009), the raters do not consider this to be genuine interactional engagement in this instance, as evidenced by Bob’s assertion in lines 2-3 that "he is not really responding to her, he’s just waiting for his chance to speak". In line 11, he goes on to suggest that the reason for this assertion is that the speaker changes the topic to one more related to himself as soon as he obtains a longer turn at talk. In other words, Bob views this as another case of parallel talk, in which the students’ extended turns lack discursive cohesion, and the raters agree that this demonstrates less engagement.

Parallelism – NEST raters

The NEST raters occasionally comment on the learners’ lack of engagement with the other test-taker’s topical talk. In Extract 8, Annie uses a similar formulation as Bob did in Extract 6 above: There is a lack of what do you think type questions as topic elaborations:
The raters are pointing to the fact that each new topic card should initiate a new topical discussion, but that these two learners display a parallel understanding of topic management, and “say a few words” on each card but do not interact on each other’s contributions. Specifically, in line 41, Ann enacts what is missing in their interaction: questions like “what do you think”, inviting a co-participant to engage with a particular topic. This indicates that the learners had a tendency to only extend on topics they had initiated themselves (cf. Galaczi, 2014). \textit{How about you}-type questions are sequentially located “after one or more speakers have already made public their stance” and “implying that the respondent’s next turn will be somehow related to the current topic of conversation” (Greer & Potter, 2008, p. 303). The \textit{what do you think} question type, while also making relevant a topically related response, is formatted more narrowly than \textit{how about you}, and more specifically projects a response with an opinion or belief. In Extract 8, Ann uses the question type as an example of the conduct that was missing, which functions as an account for her earlier assessment-relevant description (lines 35-36) of how the test-takers basically read a topic card and provided a few brief comments before moving on.

In similar ways as with post-expansion, the NEST raters seem concerned with topical engagement, where both test-takers, ideally, engage with the talk expressed in a prior turn. In relation to the NEST performance descriptors (Appendix 2), raters here appear to orient to the \textit{complexity and variation} – \textit{different examples and perspectives} descriptor, and to the \textit{communicative strategies} descriptor regarding developing and carrying the conversation forward.

\textbf{Naturalness}

Finally, a third theme we identified in the rater discussions was that being engaged led to a more ‘natural interaction’; that is, more spontaneous and reactive to
immediately prior talk. While test talk can be considered as ‘naturally occurring test interactions, they are not ordinary conversations where participants have to take the achievement to intersubjectivity at every point very seriously (cf. Seedhouse, 2013); neither are everyday conversations structured by pre-set topics. However, engagement as a sub-construct of IC aims to capture aspects of interactional conduct that resembles conversations in non-testing contexts, and our raters indeed identified aspects of the students’ interaction they judged to be ‘genuine’.

Naturalness: KTOP raters

In Extract 9 we can see that the KTOP raters orient to this ‘naturalness’ as reflective of the test-takers’ level of engagement. Here Bob says, for example, that such participant talk is “starting to become like a genuine conversation”.

**Extract 9.** “It’s starting to become like a genuine conversation”

<table>
<thead>
<tr>
<th>Line</th>
<th>Bob</th>
<th>JIM</th>
<th>Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>yeah- I- I changed them from B to A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>because it started out like they’re doing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>what they’re supposed to be doing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>JIM</td>
<td>yeah</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>BOB</td>
<td>I ask you a question you give me a response</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>and all that sort of thing but</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>um, later they get on the topic of their</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>sports?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>JIM</td>
<td>or maybe that was it, sports</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>BOB</td>
<td>I mean they’re talking about their training,</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>JIM</td>
<td>yeah</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>BOB</td>
<td>then it’s kinda like a hot topic for both of</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>them</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>JIM</td>
<td>yeah</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>BOB</td>
<td>he’s like oh it’s starting to become like</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>a genuine conversation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>JIM</td>
<td>right</td>
<td></td>
</tr>
</tbody>
</table>

Bob reported that he revised his assessment of the test-takers (line 1) when the topic changed in a step-wise manner to one that was only tangentially related to the test topic, sport (line 8). Bob typified it as a "hot topic for both of them", suggesting that the participants seemed more genuinely interested in what they were saying and responding to it as actually newsworthy, rather than as just part of the test alone. Bob's observation here is set up as contrastive to the earlier part of the test in which they "started out ... doing what they're supposed to be doing" (lines 2-3), such as initiating and responding, which Bob links to a B grade. It was only when the participants appear to become observably engaged with the topic that Bob classifies it as "a genuine conversation" and upgraded his assessment of their engagement level to an A.
Naturalness: NEST raters

The NEST raters never specifically mention naturalness by using descriptive formulations such as natural, authentic, genuine or real. Instead, they pinpoint sequences in which the topic appeared to interest both learners genuinely, which provided an opportunity for interacting as if it were not a test. In Extract 10, Brianna identifies “some of the questions” (i.e., topics) where the test-takers “had more to say”, and accounts for her claim by offering “this thing with cheating” as an example of such a sequence (line 182). This does not specifically say anything about engagement or naturalness, but rather identifies a topic in which the learners perhaps had more knowledge and interest and were able to make more substantial topical contributions:

**Extract 10. “Able to express their own views” [CASS1: 179-191]**

179 BRI  ja det var ju ändå några av frågorna där
   yeah well there were some of the questions where
180   det blev mer (.) där dom hade mer att säga också
   there was more (.) where they had more to say too
181 SAR  m:
182 BRI  om det här med [med] fusk och
   about this thing [with] cheating and
183 MAY  [m:.]
184 SAR  [m:.]
185 SAR  mm:
186 TER  ja: m:
   ye:s m:
187 BRI  om eih data[spel]ändet och som säkert hade
   about e:h gami[ng] and that they could have
188 TER  [m:]
189 BRI  kunnat (.) där dom säkert hade
   could (.) where they surely could
190   kunnat uttrycka [egna egna] ställningstaganden
   have been able to express their [own own] views
191 MAY  [ja:h m:]
   [ye:ah m:]

The co-raters agree with Brianna, who then offers a second example – the topic of gaming (line 187). However, she also adds that for this topic, she would have expected that the learners had offered even more in terms of “their own views” (line 190). Her analysis is thus clearly linked to the assessment factor of offering different perspectives and examples, and engaging with their interlocutor’s viewpoints. As such, the evident topical suitability for particular learner pairs is something the NEST raters perceive, but with a clear orientation to the assessment factors concerning opinions and personal viewpoints rather than to engagement as an aspect of IC.

---

7 For a detailed interaction analysis of the gaming sequence referred to here, and raters’ discussion of it, see Sandlund and Sundqvist (2019).
Discussion

This paper has investigated learner engagement in EFL speaking tests as a subconstruct of interactional competence, particularly in relation to raters’ identification of its features in actual instances of talk. As such, the conceptualization of engagement in the analyses presented may not bring us closer to an interactional specification, but rather, sheds light on how the raters in our study define it based on the rubric documents they have in front of them. In other words, categories like post-expansion and naturalness are aspects of engagement that originate in the rubrics and that the raters view as worth noting in the test-takers’ interaction. That being the case, we see engagement as a sub-construct of interactional competence which can be made accessible to teacher raters in rubric formulations.

As multiple conversation analytic studies of L2 test interactions have cautioned, test interactions constitute “particular types of speech situations” (He & Young 1998, p. 8) and operate within the frame of testing and the specific constraints on the competencies that learners can display (cf. Sandlund et al., 2016a; van Compernolle, 2011). As such, test interactions are indeed ‘natural’ as test interactions, even though they are not comparable to everyday conversations. Particularly with oral proficiency interviews (OPIs, see Young & He, 1998), a recurrent criticism of the testing format with an interviewer asking a candidate questions is that such testing formats do not allow for inferences to be made about a learner’s IC outside the testing situation.

As Seedhouse (2013) has argued, there is “no requirement to achieve intersubjectivity” (p. 211) in most oral proficiency tests, unlike in naturally occurring conversations, where shared understanding is key. The specifics of the testing context may indeed compromise the ‘naturalness’ of the interactions; however, engagement (as a sub-construct of IC) aims to capture aspects of interactional conduct that resemble non-testing interactions where there are social motivations for engaging with an interlocutor’s contributions. From our observations, it is evident that the CA-informed rubrics helped raters conceive of engagement in a slightly different way than the non-CA descriptors. The Swedish raters, who continuously orient to the formulations in their rubrics, seem to focus more on content than on IC in the sense that we see in the Japanese data. That is, the non-CA rubrics oriented raters toward how the learners developed a topic and expressed opinions through post-expansion, displayed knowledge of or interest in a topic. Alternatively, when orienting to parallelism in the interaction, the raters’ evaluations were tied to lack of engagement with the topical talk of their partner. All four Swedish rater groups rated the peer interaction between the two learners negatively, and also seem to ground their evaluation mainly on the test-takers’ lack of independent/unprompted initiatives in the conversation, on a topical level, although they rarely comment on the interaction in ways that we, as conversation analysts, think of interaction.
Despite similar test formats, the rubrics do indeed focus the raters' attention on different aspects of the interaction, so ultimately that is a basis for our major premise that interventionist CA; that is, conversation analytic work explicitly aiming to influence practice (Antaki, 2011), can play a role in directing raters' attention toward interactional elements of the test. Moreover, raters do not necessarily have to be trained in CA as the CA-informed rubric along with the video samples educates them to see the practices in other tests. The extensive descriptive detail available within the exemplar style rubrics also provides the opportunity to explain multiple aspects of the notion of engagement and the video offers a standard by which other test recordings can be judged. Although the KTOP test was originally designed as an in-class test to be administered by and for a single teacher, the current investigation suggests that the exemplar-based rubric would be an efficient means of training other EFL teachers to grade engagement within their own classes, especially in contexts in which cross-group consistency was deemed necessary.

A different, and perhaps more fundamental question, however, involves whether or not it is valid to assess engagement in the first place. It is important to note that both tests did also devote a significant portion of their grades to purely linguistic output, via such test constructs as accuracy and complexity, and we are therefore not denying the importance of assessing language. However, by focusing additionally on engagement we hope to send a washback-like message to students, encouraging them to be aware of IC and the need to actively participate in the conversation by initiating post-expansive sequences, designing their next turns on what was said immediately prior, and striving for natural communication in which they focus on topics that are genuinely newsworthy to their interlocutors.

**Conclusion**

In this paper, we have presented findings from an exploratory study on how two different rubrics for assessing engagement (as an aspect of IC) impact on raters' perceptions of test-taker conduct in EFL paired discussion tests. Using a qualitative content analysis of rater discussions, we have compared rater considerations of engagement in test-takers' performance. Using data from two complimentary contexts – a researcher-developed test assessment rubric for interaction based on conversation analytic findings from Japan and a criterion-referenced national test of English from Sweden – we examined how raters assessed aspects of interaction we consider related to engagement. We were interested in knowing if the Japan-based raters would be able to apply the CA-grounded rubrics and exemplar recordings in identifying interactional engagement, and retrospectively, how both groups of raters formulated their own understandings of this concept during their assessment discussions. A number of themes were identified in the qualitative analysis, of which three are detailed here. We then searched the non-CA rater dataset for constructions of
engagement, and compared the ways in which raters in the two contexts specified interactional conduct connected to post-expansion, parallelism, and naturalness.

The KTOP rubric descriptors used some terminology that originated from CA, such as change-of-state token, repair, and minimal receipt, as well as CA-style transcripts of the learner interactions. In contrast, the Swedish raters had access to the test constructors’ written commentary to benchmark tests at different grade levels. Analysis of the rater discussions indicates that the combination of video segments and everyday descriptions of the interaction for the KTOP exemplar tests made the CA transcripts and terminology unproblematic for the language teachers who served as raters. As such, our study supports previous efforts to draw on the strengths of CA in developing assessment materials for interaction (Huth & Betz, 2019; Walters, 2007), as well as the growing body of work on interventionist conversation analysis (Antaki, 2011; Sandlund et al., 2016b), which aims to use CA to deal with interactional problems collaboratively with people in specific real-world contexts.

These initial attempts to develop an exemplar-based EFL rubric via a CA-style analysis of prior test samples can therefore be seen as interventionist in the sense that they help language teachers address the real-world challenge of understanding a particular aspect of IC and rating it consistently across multiple cases. By reflecting on how the raters applied the notion of engagement to their discussions, we can consider ways the rubrics can be refined and revised for future tests. In addition, it is hoped that this project will help language teachers consider how engagement (including notions like parallel talk, post-expansive sequences and naturalness) is manifested during oral tests they use in their own classes, and we would encourage others to apply this approach to language assessment within those contexts.

However, as an initial explorative investigation, the study is not without its limitations. The coding scheme was developed only from the KTOP data and it is possible that the NEST raters orient to engagement in ways not captured by the sub-themes identified in the KTOP data. Although the findings may have been different if themes had been identified through iterative review of both discussion datasets, one of the aims in this study was to compare ways that the NEST teacher-raters oriented to the categories identified in the KTOP data. Future research should also strive for greater comparability within the research design, such as by comparing similar raters, rubrics and performances rather than two fundamentally different contexts. That said, the participant-centred stance that underpins the CA approach has informed and shaped our study, and that led us to ask ourselves how teacher-raters operationalise engagement in contexts where it was not the major focus of the rubric.

One question that requires further research is whether engagement is an essential aspect of IC to be included in scoring rubrics in ways similar to the one examined here. We have examined only some of the features of the many possible interactional
practices and activities that test-takers – and raters – may orient to as engagement. In our study, raters drew on a wider array of indicators of engagement than we included in the comparative analysis, such as tone of voice, gestures, perceptions of enthusiasm, and an unspecified ‘sense’ among raters that a particular learner “did not want to be there”. Given that raters occasionally orient to embodied features (cf. May, 2011), the use of video-recordings of paired tests as exemplars, and for assessment, would be essential for further development of CA-grounded rating scales.

The study also offers promise in terms of pedagogy. Raising learners’ (and teachers’) awareness of the specificities of conversation may improve their understanding of what is expected when in learner assessment conversations (cf. Waring, 2018; Salaberry & Kunitz, 2019). Since traditional oral proficiency constructs require learners to show what they know in terms of lexical and grammatical knowledge of the target language, there is a risk that learners may prioritize extended individual turns to demonstrate their linguistic competence instead of engaging with their interlocutor’s turns (cf. Sandlund & Sundqvist, 2011). Findings from the present study contribute to our understanding of how learner interaction is perceived by raters, which in turn could inspire work on awareness-raising classroom activities.

References


Appendix 1a. The KTOP exemplar-based rubric (A grade)

Grade A (Exemplary)
Score 8/10 to 10/10
Video Sample 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 “Ourn” or “Huum”, 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.

\[
\begin{array}{ll}
\text{A} & \text{I- (0.3) often went to sannomiya kobe} \\
\text{B} & \text{O:::h san[nomiya] } \\
\text{A} & \text{[to: play]} \\
\text{B} & \text{very fashionable} \\
\text{A} & \text{yes} \\
\text{B} & \text{fa:[ashionable] } \\
\text{A} & \text{[fashionable]} \\
\text{B} & \text{nice place}
\end{array}
\]

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.
Appendix 1b. The KTOP exemplar-based rubric (B grade)

<table>
<thead>
<tr>
<th>Grade</th>
<th>B (Proficient)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>7/10</td>
</tr>
<tr>
<td>Video</td>
<td>Sample 2 (KTOP 28)</td>
</tr>
</tbody>
</table>

**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 (0.5) 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.
Appendix 1c. The KTOP exemplar-based rubric (C grade)

Grade  C (Needs improvement)
Score   6/10
Video  Sample 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 when do you want to (.) marriage.
B  uh:::.sh I want to ↑marriage
   (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.
Appendix 1d. The KTOP exemplar-based rubric (D grade)

Grade D (Unsatisfactory)
Score 0/10 to 5/10
Video Sample 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:\,: (0.7) my private is (.): more important [zh\Ah hah] B [hh hh ] "(hah hah)"

I think (.): [so ]
A [(I-hoh hoh)]

((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.
Appendix 2. Introductory page to the NEST (translated by the first author)

Part A – focus: free oral interaction and production

[Introductory page with test materials, translated from Swedish, Swedish National Agency for Education, 2014]

Assessment factors

In assessing, the following factors can be analyzed:

Willingness and ability to interact and speak

- Interacting – taking the initiative, understanding what others are saying, and carrying the conversation forward
- Narrating, describing, and making a case for something
- Adapting what is said to the situation, topic, and recipient

Content

- Richness and complexity
- Topic treatment (focused/in-depth – brief/superficial)

Language

- Comprehensibility – the ability to express a message clearly
- Ease, variation, and security – fluency
- Strategies for solving language problems
- Vocabulary and idiomaticity (range, variation, accuracy)
- Articulation, pronunciation, and intonation
- Grammar (range, variation, accuracy)
Appendix 3. Assessment factors NEST in teacher booklet (translated by Sandlund & Sundqvist, 2019)

Content

- Comprehensibility and clarity
- Richness and variation (different examples and perspectives)
- Context and structure
- Adaptation to purpose, recipient, and situation

Language and expression

- Communicative strategies
  - To develop and carry the conversation forward
  - To solve language problems by e.g. reformulations, explanations, and clarifications
- Fluency and ease
- Breadth, variation, clarity and confidence
  - Vocabulary, phraseology, idiomaticity
  - Pronunciation and intonation
  - Grammatical structures
- Adaptation to purpose, recipient, and situation
Appendix 4. Fragments from the KTOP data the raters are discussing

Test fragment 1. KTOP6

01 A  next topic. u time
02   (1.0)
03   uh::
04   (0.8)
05   sh- (1.3) share house
06 B  "share house"
07   |(0.9)
       b-bh |brace table
08 A  uh::
09   (1.0)
10 B  ((clears throat))
11   (0.6)
12 A  [uh ]
13 B  [if you] if you share house (0.7) mm:
14   do you want (0.7) how- how many people
15   do you want to: share [with?]
16 A  [ah: ]
17   (0.4)
18 A  |uh |two- uh two people
     a-hd |tilts left
     a-rh |holds up two fingers
19 B  tw(h)o people.
20 A  two or three (girl or guy).
21 B  .hhh
22 A  a few people.
23 B  a little (.) little people.
24 A  [heh
25 B  [hahaha .hhh
26   |(0.5)
     b-lh |rubs nose
27 A  uh uh many (0.4) uh: (0.3)
28   share house with nn many people is uh:
29   |(2.0)
     a-bh |waves in small circles
30   uh:: that I: don't have |private life
     a-hd |nods
     b-hd |nods
31   |(5.1)
     a-gz |stares at card
     b-lh |covers mouth
     b-gz |stares at card
32 A  "share house"
33 B  mm:::
34   |(0.3)
     a-lh |covers mouth then looks at C
Test Fragment 2. KTOP31

01 A m::m (1.1) my father is wo- working in a bank
02 B bank.
03 A my hometown
04 B yes.
05 A n::n the (0.3) his (0.3) his work is (0.8)
06 so:: hard (0.6) |so hard |
   a-rh |rolls hand
07 B so hard.
08 A and very busy.
09 B "very busy" |
   b-hd |nods
10 A so::: (0.9) I: |don’t like (.) I don’t=
   a-hd |shakes head
11 =want to: work in a bank.
12 B in a bank.
   a-hd |nods
   b-hd |nods
13 (2.1)
14 A "and" (0.7) "hh hh"
15 B my- and my mother work (0.6) my mother
16 work |teacher
17 A teacher?
18 B in-in |school |work in school
   b-gz |away | back to A