Equity in L2 English oral assessment: Criterion-based facts or works of fiction?

Erica Sandlund, Karlstad University

Pia Sundqvist, Karlstad University

Abstract

For assessment to be equitable, it is central that teachers/raters perceive and apply grade criteria similarly. However, in assessing L2 oral proficiency in paired tests, raters must grade test-takers individually on a joint interaction performance. With a conversation analytic approach, we examine closely one recording from a 9th-grade national test of L2 English with an aim to uncover some aspects that underpin vastly divergent assessments (as assigned by three raters) of one test-taker. Findings pointing to issues such as moral stance, rater experiences, and the interlocutor effect are discussed in light of equity in L2 oral proficiency testing and assessment.

Keywords: L2 oral proficiency, interactional competence, morality, turn-taking, language assessment, equity

1. Introduction

Whether being involved in English language teaching at the university level or in compulsory school, testing and assessing students' oral and written proficiency is central. In dedicating this paper to Solveig Granath's career, we know that even though her own research interests lie outside the scope of second/foreign (L2) language testing and assessment, Solveig is a passionate teacher, loved by her students (evidenced not least in her repeated nominations to the Student Association's Best Teacher Award), and partly, we believe, because of her dedication to making each examination task an opportunity for learning. Feedback from Solveig on examination tasks is always rich and detailed, so issues on assessment and equity seem close to Solveig's heart in her professional practice.

Having said this, however, for many teachers formal assessment is a time-consuming task and assigning a grade to a student's test performance in alignment with externally set criteria is a challenge, not least with regard to mandatory national tests in school. With the national

Sandlund, Erica and Pia Sundqvist. 2016. "Equity in L2 English oral assessment: Criterion-based facts or works of fiction?" *Nordic Journal of English Studies* 15(2):113–131.

tests in English and other core subjects in focus, reports from the Swedish Schools Inspectorate have indicated that equity in assessment is a problem (e.g., Skolinspektionen 2012). Re-assessments of student performances in English (including tests of reading/listening comprehension and writing) have revealed diverging views between external raters and the students' own teacher. While the teacher obviously has more knowledge about each student's abilities as compared to an external rater, the reports show that criteria for assessment may not be interpreted in the same way by different raters. In this paper, we wish to bring the issue of equity in assessing English oral proficiency out of the woodwork, and discuss some of the issues that may come into play in diverging assessments. We examine one recording from the speaking part of the mandatory high-stakes, summative, 9th-grade national English test in this case study, and set out to locate some reasons for *why* particular students' oral proficiency may be difficult to assess, resulting in different grades assigned by different raters. We wish to contribute to an ongoing debate on L2 assessment equity (see, e.g., Moss, Pullin, Gee, Haertel, & Young 2008) with this study and to problematize the paired oral proficiency test format in relation to assessment criteria.

2. Literature review

In honor of Solveig and her scholarly interests, this literature review begins with a brief etymological account of the key terms from the title, before relevant academic work related to the topic of this paper are discussed.

According to the *Oxford English Dictionary* (2016), the general meaning of the noun *equity* is '[t]he quality of being equal or fair; fairness; impartiality; even-handed dealing'. The first record of *equity* (from Latin *æquitas*) dates back to the early 14th century (Shoreham: "Thet hys hys pryvete Of hys domes in equyte"). The first record of the noun *assessment* (probably of Anglo-Norman origin) appears about a century later in reference to the determination or adjustment of taxation. In its educational use, however, there is no record of *assessment* until in 1956 (attributed to H. Loukes). The dictionary defines *assessment* as '[t]he process or means of evaluating academic work; an examination or test'. In this paper, the evaluative aspect of assessment is focused.

In applied linguistics and second language acquisition studies, research on L2 oral language is a fairly recent phenomenon, gaining scholarly interest from the 1950s and onwards (Fulcher 2003). L2 oral language has been defined as learners' ability to converse with one or several interlocutors (cf. 'interactional competence' in an L2, Kasper & Ross 2013: 9). Whereas external examiners are frequently used internationally in speaking tests, in Sweden – the setting for our study – test-takers' own English teacher acts as the examiner. From an international perspective it is also more common to adopt a test format in which there is a native-speaker examiner together with only one test-taker, the oral proficiency interview, OPI (see, e.g., Fulcher 2003). The test examined here differs in that it involves two or sometimes more test-takers. It can be noted that paired L2 OP tests have grown increasingly popular and it has been argued that such tests reflect natural conversation better than OPIs (Ducasse & Brown 2009). When Brooks (2009: 341) compared the two test formats (OPI versus dyadic), she found that dyadic tests resulted not only in higher scores for test-takers, but also in "more interaction, negotiation of meaning, consideration of the interlocutor and more complex output." From an assessment perspective, however, using two (or more) test-takers may be problematic due to the fact that the spoken output is a joint product (He & Young 1998) by individuals who later are assessed individually (May 2011; Sandlund & Sundqvist 2011). A general finding appears to be that paired tests allow for more flexible test-taker contributions and a wide range of complex actions. Not surprisingly, then, the role of the interlocutor is highly important because s/he is likely to influence both scores and interaction – sometimes positively, sometimes negatively (Davis 2009; Iwashita, Brown, McNamara, & O'Hagan 2008; Lazaraton & Davis 2008). Galaczi (2008) suggests that test-takers with limited L2 skills are not very involved in interactions with their interlocutors, a topic she further explores in Galaczi (2014), where a broader view on interactional competence is recommended. Regardless of test format and group sizes, research has shown that preparation affects test results (Farnsworth 2013), and it appears that gender may also play a role in L2 speaking tests (Amjadian & Ebadi 2011).

In a study of our own, we focused on how test-takers managed interactional trouble connected to the test tasks (Sandlund & Sundqvist 2011). The findings indicated that some task management strategies

appeared to be rated less favorably than others, even though some (for instance, negotiation of understanding of the test task) were perfectly productive for the students themselves in terms of test-wiseness. Being test-wise has to do with, among other things, the willingness to play along in a test situation (Bachman 1990). In tests where stakes are high for test-takers, having such an ability may prove rewarding in terms of assessment and, needless to add, in such tests, the issue of equity is essential. Equity is particularly important when the construct of interest is complex and comprises a number of different variables, as is the case with a construct such as OP. It can be noted that Sandlund and Sundqvist (2011) demonstrate that the validity of the English speaking test may be threatened by demands of topical knowledge irrelevant to the intended construct, and that the teachers'/examiners' objective to elicit enough assessable talk result in differing and unwanted patterns of interaction (Sandlund & Sundqvist 2013). Finally, Ducasse (2010) argues that there is a need for more research on what takes place in the interaction between test-takers in dyadic L2 speaking tests; for example, no rating scales have been developed based directly on empirical data from observed performances of such interactions. Although we do not center on developing rating scales here, this study makes a sought-for contribution in that it is based on exactly the type of interaction test data Ducasse (2010) is referring to.

3. Research question

We are interested in exploring reasons for diverging assessments of L2 English oral proficiency and, therefore, ask: What possible reasons for raters' diverging assessments of oral proficiency can be found in the interaction between students in a paired oral proficiency test? A few possible answers will be provided through detailed examination of one test.

4. Method and materials

4.1 Data and participants

Recordings and assessment data were collected as part of a research project (*Testing Talk*, the Swedish Research Council, Reg. no. 2012-4129) on the speaking part of the 9th-grade national test of English in

Sweden. In *Testing Talk*, 71 recordings of paired/small group tests were collected at four schools. In total, 161 students (aged 15–16) from ten classes participated. For the assessment part of the project, three separate assessments of student performances were collected: one from the students' own English teacher (in total, six certified teachers with more than ten years work experience were involved in these assessments) and one each from two external raters (equally qualified). Provided with audio recordings of all speaking tests, the external raters assessed and scored student performances independently. They were instructed to follow standard procedure, that is, the instructions included in the materials provided by the Swedish National Agency for Education. Raters' commenting in writing on test performances on the score sheets was optional but encouraged. In addition to the speaking test assessment data, students' scores on the other parts of the national English test (see below), the global national test grade, and the final grades in all school subjects were also collected.

Some additional data could also be collected. For instance, External Rater 2 handed in copies of her original scribblings on which she based her "official" comments. As it happened, during the course of the project, it also came to our knowledge that some teachers employed self- or other teacher-created matrices for assessment of OP, and these documents were shared with the project team.

For two participating classes, the teachers first assessed test performances independently for students in their own class and provided us with documents of student grades; these teachers then collaborated in co-assessing some performances with the help of the recordings. In only a few cases, the initially assigned grade was changed; the grade *after* co-assessment became individual students' official test grade. It ought to be mentioned that these two classes differed from the others in that they were Content and Language Integrated Learning (CLIL) classes (see, e.g., Dalton-Puffer 2007), which means that the participating students had been exposed to English as the medium of instruction in other subjects than English throughout grades 7–9 and, in addition, they had had English lessons on a daily basis for three years.

For the purpose of this paper, a targeted search was conducted for students where there was considerable divergence between the grades assigned. The criterion for *diverging assessments* was that at least two out of the three assessments differed with at least two grades (for

example, from A to C or from B to D). This search helped identify test recordings deemed as relevant for the scope of the study of which one, for several reasons, was selected for detailed scrutiny. For instance, in the selected test recording, the teacher and External Rater 2 had assessed one of the students, Leo, with a C-, whereas External Rater 1 had evaluated Leo's performance as an A. Moreover, the selected test happened to be from one of the CLIL classes and Leo's originally assigned grade by his own teacher was, in fact, a D. Moreover, all three assessors were experienced teachers of English but only one of them (Leo's teacher) knew the student and had performed additional classroom-based assessments that could possibly color her assessment, as indicated in tendencies observed by the Schools Inspectorate (2012) for writing (which, as speaking, is a multifaceted language ability leaving room for interpretation on the part of the assessor).

The analyzed test recording was 23 minutes and 6 seconds. The involved students are Leo and Magnus (pseudonyms). Magnus' performance was awarded a B by the teacher and both raters. External Rater 2 had commented on this specific test (see Table 1).

Table 1. External Rater 2's comments on Leo and Magnus (translated from Swedish).

Student	Grade	Comment
Leo	C-	*Good fluency and relatively good vocabulary + idiomaticity* *Somewhat poor in terms of content, briefly explains what he means, does not deepen his contribution, does not interact particularly well*
Magnus	B	*A lot of production* *Fluency disrupted at times* *Reformulates, clarifies* *Adapts somewhat to his interlocutor, communicative strategies to lead the conversation forward can be developed more* *Good vocabulary and rather sure of accurate grammar use*

4.2 Test instructions and text details

In Sweden, national tests are summative and share a twofold purpose: (i) to contribute to equity in assessment and grading and (ii) to yield data for evaluation of goal-attainment (www.skolverket.se). The present paper is particularly relevant in light of the former purpose. The national test in

English is a typical proficiency test which does not assume prior knowledge of a specific content. It consists of three parts: Part A focuses on oral interaction and production ("speaking"), Part B on receptive abilities (reading/listening comprehension), and Part C on written production. In this study, data are drawn from the 2014 speaking test.

In preparing for test administration and examination, teachers receive a booklet with detailed instructions and a CD with sample test recordings for different grade levels commented on in the booklet with references made to relevant grade criteria. During the test, teachers may prompt students if they run into difficulties but, generally, the present teacher/examiner should remain in the background (Swedish National Agency for Education 2013).

Over the years, instructions regarding the number of students per speaking test has varied and, at times, instructions have been unclear in this regard (Sundqvist, Sandlund, & Nyroos 2014). In any case, regardless of whether there are two or more test-takers involved, the speaking test adopts a three-step format, beginning with a rather monologic warm-up section before moving on to step two, which involves more of dialogic speech between test-takers, and step three, where even more interaction between test-takers is the aim. This type of design in which the level of difficulty increases gradually is typical for speaking tests, as shown in a recent research overview of L2 oral testing (Sandlund, Sundqvist, & Nyroos 2016). In the test examined here, pictures were used for warm-up. In the next phase, one at a time, test-takers drew so-called topic cards (blue) from a stack of cards. On each card, there is a statement and some questions to be read aloud and discussed. After that, yellow topic cards were used in a similar fashion, again with statements serving the purpose of triggering further test-taker interaction. It needs to be mentioned that because of a 6-year secrecy put on the 2014 national test, the first sentence on the topic cards has been concealed in the transcripts below, and we are only at liberty to discuss the topic card formulations in general terms.

Since teachers act as examiners for the speaking test, specific information regarding how to assess students' output in relation to relevant grade criteria are given in the booklet. More specifically, teachers/examiners are instructed to focus on *content* and *language and ability to express oneself* in the assessment. As regards content, assessment should focus on the following variables: (i) intelligibility and

clarity, (ii) rich and varied content (providing different examples and perspectives), (iii) context and structure, and (iv) ability to adapt to the purpose, the interlocutor(s), and the situation. For language and ability to express oneself, assessment should focus on: (i) the use of communicative strategies (developing the conversation and leading the conversation forward as well as solving linguistic problems with the help of reformulations, explanations, and clarifications), (ii) fluency and ease of speaking, (iii) variation, clarity, and accuracy (vocabulary, phraseology, and idiomaticity; pronunciation and intonation; grammatical structures), and (iv) ability to adapt to the purpose, the interlocutor(s), and the situation (Skolverket, 2013: p. 28).[1] After the completion of a test, the teacher assigns a grade to each student test performance by checking the appropriate box (see Figure 1), and as mentioned, scores on the test are interpreted in relation to criteria.

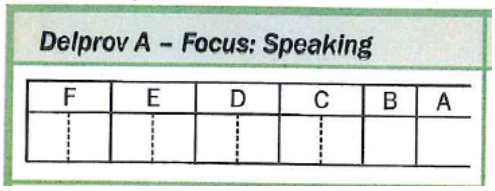

Figure 1. Grade grid for Part A, Speaking (Skolverket, 2013).

There are six grades, A–F, where F is assigned when criteria are not met. To a large extent, the criteria for English are aligned with the communicative abilities described in the *Common European Framework of Reference for Languages* (Council of Europe 2001), and a passing grade (E) corresponds to level B1.1 (Council of Europe 2001).

4.3 Analytic procedure

Adopting a conversation analytic (CA) approach to the test interaction data (e.g., Sidnell & Stivers 2013), we are particularly interested in how the interaction unfolds in situ, and how participants format their

[1] Before the curriculum implemented in 2011, another variable (*willingness and ability to interact and talk*) was also included; this variable still remained on one sheet in the 2014 test materials. Thus, it is possible that teachers/raters included it in their assessments.

contributions in relation to immediately prior talk. For this paper, we look specifically at the organization of turns and turn shifts, and the ways in which students orient to each other's prior turns. Transcriptions presented below use standard CA conventions for depicting turn features and sequential organization (Sidnell & Stivers 2013).

Due to space limitations, the full transcript cannot be included here. Instead, having analyzed the entire recording, we selected three sequences, presented in chronological order. Extract 1 occurs 11 minutes into the test and Extracts 2a and 2b at the end. All sequences are from the third "interactive" part of the test.

5. Findings

5.1 Extract 1: Turn-taking and moral dilemmas

In the first fragment, it is Leo's turn to draw a topic card. It presents a moral dilemma dealing with how to act when an elderly person boards a full bus:

(1) [Rec 11011181, 11.21–13.19]

```
51    LEO   should I?
52    TEA   (   ) (.)
53    LEO   eh:m (0.9) ((reads test card aloud)) (2.0)
54          hh we:ll (0.3) I::: (.) wouldn't really do:
55          (1.0) s (0.4) s- so much for that person (.) if
56          that person didn't r- really: (0.2) >come up to<
57          me and ask for the seat (0.6) b'cause then I
58          could stand up a:nd (.) let her sit down? (0.6)
59          because the elderly (0.7) n- needs it (1.2)
60          a::n (1.1) o:r if it's like (1.5) a djounger
61          person (0.2) siddin' next to me (1.0) I would
62          tell (.) that person maybe yump u:h (0.4) stand
63          up? (0.6) and give that (person) to theu:h (0.2)
64          elderly¿ (1.6)
65    MAG   *myeah* (1.6) I I I w- I have actually (0.4)
66          been in kindof that situation (0.3) .hhh but I'm
67          not sure IF ifsh if (.) the person
68          didn't ↑Ask for the seat (0.7) I- I-
69          don't think I really would dare I- I would
70          feel embarrassed by standing up an'like
71          (0.2) but I- I know it's the right thing to do¿
72          .hh (0.4) and of ↑course if they asked I would
```

```
73          (.) e:h (.) of course give them my seat
74          anduh (1.1) but I'm not sure e:hm (0.7) if
75          (.)  if there-or if ↑there is really a full
76          bus I would s- probably stand up (0.8) or I
77          would a:sk the person would you like to
78          sit here (1.2) an'uh (.) if they say (.) yes
79          please I would of course stand up (0.3) a:n hh
80          (1.1) but e:h .hh (0.8) I'm not really that
81          s:ocially (0.4).hh  I'm a liddl: (0.7)uhd h(0.3)
82          shy? so I (.) it would beu:h (.) >yeah (0.5) I
83          wouldn't be very (0.5) brave (0.2) and (0.2) but
84          I I would try to ask them at least try to
85          have the courage to ask them (1.3) andu::h (.)
86          ↓yeah
87          (4.9)
88          this (side) yeah?
89    LEO   yeah.
90    MAG   (.) should I take a third one?
```

A first observation is that the turn-taking pattern observed is rather common in the paired tests in our corpus. It generally consists of relatively long turns prior to turn transition, few (if any) overlaps between the speakers, and test-taker orientations to the activity as a conversation where one party 'exhausts' his viewpoints on the topic before a co-participant offers his (or hers). As compared to everyday conversation, this pattern is strikingly different.

As Leo embarks on his first response, the turn-initial, drawn-out "w:ell" in 54 can be heard as projecting an upcoming disagreement or disaffiliation (Heritage, 2015; Pomerantz 1984). As a "departure-indicating" particle (Heritage 2015: 89), *well* in turn-initial position has also been shown to project upcoming "'my side' responses to descriptions and evaluations in which the speaker's perspective becomes a new point of departure for subsequent talk" (2015: 101). In the remainder of the first part of Leo's turn in 54–64, he does indeed account for his own perspective, and one that could be seen as socially problematic – i.e., that he would not offer his seat to an elderly person unless directly prompted to do so. Even though openly formulated as a 'what would you do' question, the test topic bears with it, like all morally charged issues, a preference toward responding in a particular, morally appropriate way, and it is to be expected that diverging from this underlying preference will require additional interactional work on part

of the speaker. Turn-initial *well* may be one such indicator, and indeed, Leo also states that he would ask a younger person to stand up for the elderly, thereby positioning himself as not first in line to give up his seat.

Magnus, however, takes a different stance in his account (65–86). While he also indicates that he would not necessarily offer his seat to an elderly person, Magnus gives a different account for why – he puts forth his own shyness and social ineptness and that he "wouldn't be very brave" (81–83) as reasons for not verbally offering his seat. His highly personalized account, which he states relates to first-hand experience, contains multiple angles on the topic, such as knowing "the right thing to do", a detailing of his personal reasons for being unable to abide by the moral code, and a desire to "have the courage to ask". In direct adjacency with Leo's turn, Magnus' account stands out as more exhaustive but also, possibly, as morally superior. In terms of grammar, lexis, and tempo, both students produce their turns with few production problems, and both contributions are clearly 'on task'. In sum, this fragment shows lengthy 'monological' accounts from both speakers, but they take different moral stances on the topic.

5.2 Extract 2a: Delimiting a topic

Extract 2a is the last topic card of this test, and the recording stops when the full sequence ends after Extract 2b (below). This fragment also shows Leo as the topical talk initiator, albeit prompted by Magnus' directive in line 111. The topic card is formatted as a statement proposing that people do "too much" of something (in this case, care too much about clothes and fashion). This means that an agreeing (*yes*-type) response would align with the formulation, whereas disagreeing with the statement (in part or fully) would require a first turn displaying some version of a negative response. Leo's first turn unit is *well*-prefaced, indicating at least partial disagreement (see Extract 1). He also makes a selective characterization of "some people" and "mostly teenagers" as indeed representing a citizen category that "cares too much" (115–116), which supports partial disagreement:

(2a) [Rec 11011181, 20.40–21.56]

```
91    MAG   °>pick a c(h)ard any c(h)ard<° ((whisper voice))
92    LEO   e:hm (.) ((reads topic card formulation))
93          (1.0)
```

```
94    LEO   well (.) some people la- mostly
95          ↑teenagers does? (0.8) e::hm (0.5) if
96          somebody like (.) doesn't have afford
97          for (0.9) some kind of clothes (h) then they
98          shouldn't buy it (0.3) they should just go
99          with their (.) clothes that they have (0.9)
100         afford for (0.6) or >yeah< (ff) ande:h (.)
101         some people care just too much like (1.5) pushing
102         some people out (0.2) b'cuz (0.7) they don't
103         havu:h (0.7) fashion clothes (.)they
104         but >whaddif they can't< afford it (0.4)
105         an' it's not (.) their fault¿
106         (1.7)
107   MAG   yeah. (.) I agree (1.1) hh (.) u::h (0.3) I like
108         nice clothes though bud I don't really care if
109         people >wearitornot< (0.9) I don't alwe-
110         always wear fashionable clothes bud (1.3) I
111         like it bu::t SOme people care (0.8) too
112         much (0.9) you shouldn't be pushing anyone
113         out or something like that .hhh because
114         they're not wearing (0.8) nice clothes (0.3)
115         cuz as you said peop- all >not everyone can
116         afford (1.1) .hh really nice clothes so:, (1.7)
117   TEA   °good?°=
```

Leo's first turn (114–125) brings up two related aspects of the topic of fashion in the form of an unspoken contrast to the topic card, in which he lets on that 'not caring' may not necessarily be the case; instead, that someone who may come across as not caring about fashion may care, but be limited by economic means. He further links limited financial resources to the possibility of social exclusion (122–124), and condemns people who 'care too much' and resort to "pushing some people out" for something children from low-income homes cannot change. His turn shows a few grammatical errors, such as "most teenagers does" (concord) and "if somebody like (.) doesn't have afford for". It is likely that the turn design shows an L1 transfer: in Swedish, it is common to say *om någon inte har råd med*, and a possible direct translation would be "if somebody doesn't have afford with" (Leo's selected option). The construction "afford for" re-occurs in 120, produced in an unmarked manner, indicating that Leo does not spot a potential problem. However, a correct expression using the same verb surfaces in the hypothetical question in 124: "but what if they can't afford it". Thus, although Leo

does not display obvious repair strategies, the construction eventually becomes grammatically accurate.

In line 127, Magnus aligns with Leo's claims ("yeah I agree") and recycles Leo's formulation about not "pushing" individuals out of social groups and "not everyone can afford" (135–136). Such recycling could be viewed differently in terms of assessment: on the one hand, Magnus does not go far beyond what Leo has already said but merely restates issues in Leo's contribution; on the other, such recycling can also be heard as displays of interactional competence – Magnus displays attentiveness to Leo's turn and is able to build upon prior talk. This is also visible in his direct reference to Leo's preceding turn, where Magnus indicates reporting prior talk with "as you said".

Another feature of Magnus' turn, which echoes his treatment of the bus dilemma, is that he begins on a personal note. Instead of treating the topic in general terms, like Leo does, Magnus begins his turn with a "my side" telling (cf. Heritage 2015) where he admits to liking nice clothes. He immediately adds, however, that his preference does not entail a comparably strong interest in what other people wear. He elaborates on this point in the admission that he does not always wear fashionable clothes himself, but that even so, it would not be an appropriate reason for excluding others. It is possible that Leo's rather forceful statements about social class and fashion occasion Magnus' extended accounts of agreeing with the gist of Leo's contributions, while also mitigating his own standpoint, but that remains speculative. Regardless, we would like to emphasize that Magnus, again, treats a topic in relation to his personal experiences, which is one notable difference between his and Leo's contributions. As we move through the second part in Extract 2b, we begin with line 137 (also above) which is a positive assessment from the teacher after Magnus' turn.

5.3 Extract 2b: Competing angles

(2b) [Rec 11011181, 21.58–23.04]

```
118   TEA   °good?°=
119   MAG   =A:E:h the latest fashion that's (1.0) it
120         differs a ↓lot (0.8) as the most questions
121         (1.5)
122   LEO   well I think it's pretty (.) pathetic that
123         some people (0.3) care about (.) fashion
```

```
124         (1.0)
125   MAG                  [ y e  a  h  ]
126   LEO   if like (0.2) >[a person on ] school<
127         (0.6) doesn't (.) have uh (0.4)
128   MAG   °oj°
129   LEO   the latest fashion (.) then you shouldn't
130         be mean to that person (0.8)
131   MAG   no of course not nej no
132   LEO   mm
133   MAG   (1.9) but fashion could ↓mean (0.3) a lot
134         of things=
135   LEO   =yeah
136   MAG   it could mean like catwalks with
137         ridiculous clothes (0.8)
138   LEO   HHHhhh
139   MAG   like no one would wear
140         (0.3) in public (1.3) and it could be
141         just (.) the brands (0.5) and the brands
142         doesn't really matter (0.9) but it does
143         madder (0.3) for a loddof people (0.7) but
144         it shouldn't madder. (0.9) although it's (0.5)
145         a lotta times there it's (0.5) better quality
146         with brands but hh (1.4) .h they are also a
147         lot (1.2) u::h (.) expens- (0.4) more expensive
148         so (2.1) (hhh) yeah, (0.6)
149   TEA   o↑kay
150   MAG   is that it? (0.5)
151   TEA   you're done?
152   LEO   yeah.
```

The assessment *good* does not reveal the exact assessable object. Nevertheless, Magnus does not treat the teacher's *good* as an indication that they are finished and, instead, he latches on with a more general statement regarding what 'counts' as fashion. His turn (138–139) opens up for a broader set of possible responses ("it differs a lot (.) as the most questions"), but formatted as a claim, it is still tilted towards a structural preference for an agreeing response. Again, Leo's response is *well*-prefaced, which may be his diversion from this new possible trajectory. Leo disattends to the new angle and instead reconnects to his own earlier talk about social exclusion and bullying (148–149). Leo's claim, including the rather strong assessment "pretty pathetic", links caring about fashion to the school context only, and takes a stand against 'fashion-based meanness'. Magnus emphatically agrees (150), but like

Leo, does not elaborate on Leo's topical trajectory. The turn-initial *but* in 152 displays that he will not completely align with Leo's claim, and Magnus' turn reaches backwards to his own prior contribution about fashion being a matter of definition (138–139). Leo provides minimal acknowledgement in anticipation of more talk from Magnus, who in turn elaborates on different ways of defining fashion (high fashion, brands, quality, and cost). The teacher offers an *okay*, which Magnus treats as a pre-closing signal, and the test interaction is brought to a close.

There are a few things to pay particular attention to in the examination of Extracts 2a–b. First, (2b) shows a different sequential organization as compared to (1) and (2a), with more turn shifts, shorter contributions, and also attempts by both students to set the topical agenda and sticking to it. Both students display fluency and use some relatively advanced idiomatic expressions (*pretty pathetic, catwalks, ridiculous clothes*), and both show slight production troubles on certain words. One of the differences observed is that Magnus, yet again, begins his first topical contribution on a personal level, whereas Leo maintains a more abstract, non-personalized approach, but also makes strong personal value claims on his selected topical perspective. Notably also, the teacher's positive assessment comes in adjacency to Magnus' turn, which occasions further elaboration. In sum, what makes (2b) stand out is that it shows heightened involvement or increased participation (Sandlund 2004), but also, a competition for a topical angle.

6. Discussion and conclusion

Having examined the test recording in its entirety, did we find any explanations for why Leo received diverging assessments while Magnus was assigned the same grade by the teacher and the raters? In terms of the criterion *language and ability to express oneself*, both test-takers display similar competence, for instance, in terms of fluency and accuracy. In terms of the interaction as such, and assessment criteria like *content* and *communicative strategies,* a few observations deserve further discussion, and so do some issues related to the test format, the criteria, and the raters in question.

Even intuitively, readers will be able to spot differences between this test interaction and naturally occurring conversation. First, to a great extent the topic cards determine topical content, and it is rather obvious

that card topics make possible different degrees of involvement from the students. Having personal experience of something, as opposed to discussing something on an abstract and speculative level, is bound to have an effect on the type of interaction that unfolds. Magnus' ability to draw on personal experiences and claim *epistemic access* to the topic (cf. Stivers, Mondada, & Steensig 2011) (whether evidence of true experience or plain test-wiseness, cf. Bachman 1990) may make his performance come across as more varied and rich. Also, the test format, where students take turns reading and commenting on each new card, establishes a specific turn-taking pattern where test-takers tend to exhaust their commentary on a topic before a turn shift is made relevant. This leads to extended, less co-participant-oriented turns with few overlaps and interjections, making each test-taker's contribution stand out almost as a miniature monologue. This sequential organization highlights differences between a first and second turn on a given topic, for example, in terms of length and topical richness. Possibly, then, 'going first' could be an advantage, and Magnus' extensive turns when 'going first' in other parts of the test (not shown here) may restrict the kinds of contributions Leo can make without coming across as repetitive. As such, being paired with a slightly more proficient *or* more talkative interlocutor may be a challenge, as a degree of competitiveness over the floor may come into play, and perhaps especially so for CLIL students. Also, it is possible that students of slightly different proficiency levels, accomplishing test interaction jointly, are compared *against each other in situ* (rather than against assessment criteria) and differences in conduct and proficiency may stand out more.

Furthermore, discussing moral issues, especially in the presence of one's teacher, can be a risky affair. As members of a community, norms are shared and people generally know what the morally appropriate stance would be when presented with a moral dilemma. When diverging from such expected moral stance-taking, the speaker may be held accountable. When comparing our focal students, both are open about dilemmas in the hypothetical situations, but Leo's approach does on occasion deviate more from the expected norm. We cannot claim with certainty that moral stance influences assessment, but note the fact that this is one of the ways in which the two test-takers differ.

In the assessment instructions, two typical features of interactional competence are included: the *ability to adapt to the interlocutor* and the

ability to use communicative strategies for leading the conversation forward. In an ongoing study on collaborative assessment of the speaking test (Sandlund & Sundqvist, forthcoming), preliminary findings suggest that the use of communicative strategies is a criterion that teachers seem to find very important, but also that being 'too talkative' and 'taking over' is viewed as undesired. As assessment criteria are rather abstractly phrased, teachers will inevitably have personal interpretations of them, and perhaps unintentionally value some more highly than others. It is possible that Leo's slightly shorter contributions and infrequent orientations to Magnus' prior turns are heard as less communicatively oriented in contrast to Magnus' use of phrases like "as you said", and that some raters over-value such conduct. However, Magnus' relative talkativeness may explain some of these differences in situ.

In terms of factors possibly impacting assessment equity, we also need to comment on the different raters in relation to their perceptions of Leo's performance. As mentioned, re-assessments of, for instance, writing have revealed that students' own teachers tend to assign higher grades than external raters do. This study has shown the opposite, as the teacher originally assigned a far lower grade than both external raters. The Swedish system with teachers assessing their own students is unusual internationally, but regardless of whether one believes this system has consequences for assessment equity or not, it is obvious that a teacher has knowledge of each student's general performance, personality, and communicative style. Such knowledge probably generally benefits students, as it relaxes the pressure to perform perfectly on one test occasion. A teacher has several opportunities to assess a student's oral proficiency in school and can therefore overlook minor problems that can be related to, for example, unfortunate pairings or test anxiety. However, it is just as plausible that Leo actually performed 'better than usual' together with a more talkative and proficient student like Magnus, and that this improvement goes unseen by the teacher, who is so familiar with Leo's performance on other occasions. Consequently, in terms of equity, teachers' knowledge of individual students' abilities may be a double-edged sword.

Finally, as the particular teacher here teaches CLIL classes, she can be expected to be used to high-performing students. A rater's experiences of homogenous and heterogeneous student groups may be

one aspect that comes into play in assessing the national test; the external rater who assigned an A to Leo had no experience of teaching CLIL classes, which most likely made both Magnus and Leo stand out as highly proficient students.

Equity in assessment, well, is it a criterion-based fact? Or is a belief in assessment equity for the national English speaking test a fiction-like utopia, like our somewhat provocative title indicates? Clearly, assessing oral tests of the kind examined here is a challenging and complex task where many factors come into play. While the optimal conditions for assessment equity for speaking tests remain to be demonstrated, we hope this case study raises some issues worthy of further discussion.

References

Amjadian, M., & Ebadi, S. (2011). Variationist perspective on the role of social variables of gender and familiarity in L2 learners' oral interviews. *Theory and Practice in Language Studies, 1*(6), 722-728.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26*(3), 341-366. doi: 10.1177/0265532209104666

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Dalton-Puffer, C. (2007). *Discourse in content and language integrated learning (CLIL) classrooms*. Amsterdam: John Benjamins.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367-396.

Ducasse, A. M. (2010). *Interaction in paired oral proficiency assessment in Spanish*. Frankfurt am Main: Peter Lang.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26*(3), 423-443.

Farnsworth, T. (2013). Effects of targeted test preparation on scores of two tests of oral English as a second language. *TESOL Quarterly, 47*(1), 148-155. doi: 10.1002/tesq.75

Fulcher, G. (2003). *Testing second language speaking*. Harlow: Pearson Education.

Galaczi, E. D. (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. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics, 35*(5), 553-574. doi: 10.1093/applin/amt017

He, A. W., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessent of oral proficiency* (pp. 1-24). Amsterdam: John Benjamins.

Heritage, J. (2015). *Well*-prefaced turns in English conversation: A conversation analytic perspective. *Journal of Pragmatics, 88*, 104.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24-49.

Kasper, G., & Ross, S. J. (2013). Assessing second language pragmatics: An overview and introductions. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 1-40). Bristol: Palgrave Macmillan.

Lazaraton, A., & Davis, L. (2008). A microanalytic perspective on discourse, proficiency, and identity in paired oral assessment. *Language Assessment Quarterly, 4*(4), 313-335.

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly, 8*(2), 127-145. doi: 10.1080/154303.2011.565845

Oxford English Dictionary. (2016). Oxford: Oxford University Press.

Pomerantz, A. (1984). Agreeing and disagreeing with assessments: some features of preferred/dispreferred turn shapes. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action* (pp. 57-101). Cambridge: Cambridge University Press.

Sandlund, E. (2004). *Feeling by doing: The social organization of everyday emotions in academic talk-in-interaction.* (Diss.), Karlstad University, Karlstad.

Sandlund, E., & Sundqvist, P. (2011). Managing task-related trouble in L2 oral proficiency tests: Contrasting interaction data and rater assessment. *Novitas-ROYAL, 5*(1), 91-120.

Sandlund, E., & Sundqvist, P. (2013). Diverging task orientations in L2 oral proficiency tests – a conversation analytic approach to participant understandings of pre-set discussion tasks. *Nordic Journal of Modern Language Methodology, 2*(1), 1-21.

Sandlund, E., & Sundqvist, P. (forthcoming). Teachers' collaborative assessment of an L2 English speaking test.

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. doi: 10.1111/lnc3.12174/epdf

Sidnell, J., & Stivers, T. (Eds.). (2013). *The handbook of conversation analysis*. Chichester: Wiley-Blackwell.

Skolinspektionen. (2012). *Lika för alla? Omrättning av nationella prov i grundskolan och gymnasieskolan under tre år*. Stockholm: Skolinspektionen.

Skolverket. (2013). *Lärarinformation inklusive bedömningsanvisningar till Delprov A. Ämnesprov i engelska Årskurs 9*. Stockholm: Skolverket.

Stivers, T., Mondada, L., & Steensig, J. (Eds.). (2011). *Knowledge and morality in conversation. Rights, responsibilities and accountability*. Cambridge: Cambridge University Press.

Sundqvist, P., Sandlund, E., & Nyroos, L. (2014). National speaking tests in English – does group size matter? *LMS Lingua*(3), 29-31.

Swedish National Agency for Education. (2013). *English. Ämnesprov, låsår 2012/2013. Lärarinformation inklusive bedömningsanvisningar till Delprov A. Årskurs 9*. Stockholm: Swedish National Agency for Education.