Teaching and Testing L2 Interactional Competence

Bridging Theory and Practice

Edited by M. Rafael Salaberry and Silvia Kunitz

13 Doing Versus Assessing Interactional Competence

Erica Sandlund and Pia Sundqvist

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

13 Doing Versus Assessing Interactional Competence

Erica Sandlund and Pia Sundqvist

Introduction

The question of what counts as competent participation in interaction in a second language (L2) presents challenges for language assessment research and practice. Specifying features of talk and conduct that constitute such 'interactional competence' (IC) (Kasper & Ross, 2013; Kramsch, 1986; Pekarek Doehler & Petitjean, 2017; Young, 2011) involves determining *which* "methods of interaction" (Waring, 2018: 58) are indices of IC in an L2. Additionally, can these methods be specified to the extent that they can be taught and tested? Theoretically, as Hall (2018) argues, the interdisciplinary roots of the IC concept have led to a conflation of ideas from two approaches—conversation analytic (CA) investigations into the interactional organization of human sociality, and studies relating to the objects of L2 learning.

While L2 scholars struggle with the specification and operationalization of IC, practitioners (teachers and raters) also struggle with the assessment of L2 interaction. In a study of how trained raters interpreted interaction in a paired speaking test, May (2009) noted that they incorporated everything from body language to assertiveness and cooperativeness in their IC ratings, which "attests to the complexity of the decisions they have to make, and the lack of meaningful descriptors guiding them to this decision" (p. 416). May's observation brings an essential aspect of testing and assessing IC to the table—namely, raters' understandings of rating criteria for interaction. While studies examining learner discourse in search of patterns indicative of changes in participation in L2 interactions (cf. Pekarek Doehler, 2018) have been crucial to our understanding of practices displayed at various levels of L2 proficiency, the local, situated context of assessment can also provide insights into how IC is conceptualized in relation to specific conduct. While L2 learners in interaction draw on L2 resources acquired over time, they also draw on a range of abilities "developed since infancy to deal with generic features of social interaction" (Pekarek Doehler, 2018: 6). Accordingly, the conflation of ideas Hall (2018) observed in scholarly conceptualizations of IC

may well have its counterpart in raters' practices: raters (and test developers) are faced with the challenge of determining whether an oral assessment task captures what has been taught/learned in an L2, or whether the assessment also pertains to a "universal interactional competence that L2 users and learners . . . orient to as human beings to do the cooperative work of human sociality" (Hall, 2018: 31).

In this chapter, we target raters' perceptions of learners' IC, but in contrast to much previous work on rater interpretations of learner performance and scoring rubrics (e.g., Ducasse & Brown, 2009), we approach raters' understandings in situated assessment talk. With a CA approach, we examine how groups of L2 English teachers participating in a training program for the assessment of L2 oral proficiency and interaction collaboratively assess learners' IC in a paired L2 speaking test, the high-stakes national test of English for ninth-graders in Sweden. In addition, we examine how teachers (as raters) discuss particular learner (test-taker) contributions in presenting arguments for their scoring in relation to rubrics. Such reported speech (RS) turns can give an insight into the kinds of interactional conduct that raters perceive as salient to IC. Furthermore, we examine the sequences in which the reported turns were originally produced in the paired test interaction, and analyze the interactional work accomplished in situ. The opportunity to see how particular actions in test interaction are re-situated for assessment purposes can inform test constructs and assessment criteria and, additionally, yield knowledge on the interactional basis for RS (Clift & Holt, 2007). An overreaching aim of the study is to illuminate IC assessment as a social practice and, through this lens, obtain insights into the kinds of conduct that raters treat as competent participation in L2 interaction. Furthermore, by working backwards from raters' identification of IC-relevant conduct and locating the turns made relevant in the learner interaction assessed by raters, the study also sheds light on divergences between learners' orientation to situated test interaction, on the one hand, and raters' subsequent recontextualization of these actions in assessment talk, on the other.

The analyses of rater interaction and learner interaction respectively serve two purposes: (1) to uncover whether raters' reconstruction and interpretation of particular turns play a role in positive or negative assessment of IC, and (2) to uncover what interactional work the actions reported on served in situ for the learners. Findings have relevance not only to research on IC for L2 teaching and testing, but also for understanding the complexity of assessing social interaction, which in turn can yield important knowledge for developing rater training for the assessment of L2 interaction.

In what follows, we situate our study in conceptualizations of and empirical work on IC in first language (L1) and L2 contexts. We then review relevant work on the testing and assessment of L2 interaction.

Next, we turn to assessment and examine studies on rater perceptions of L2 IC, assessment moderation and rater agreement in assessing speaking. We subsequently present the study design, the data and our analytic approach. The analysis is presented in two parts. First, we present findings regarding raters' collaborative construction of learners' IC, including a focus on their reports of a particular action in the learners' interaction. Then, we turn to the learner test interaction and examine interactional work accomplished in the sequence that raters report on in their assessment discussions. Finally, we contrast and discuss the findings from the analyses of rater and learner interaction, respectively. We conclude by discussing some implications of contrasting assessment talk with learner interaction.

Assessing L2 IC

Competent Participation in Social Interaction

Current applications of the concept of L2 IC have grown out of influences from several sources, such as Hymes's ethnography of communication (Hymes, 1972; see also a recent review by Hall, 2018), Canale and Swain's (1980) model of communicative competence (see also Bachman, 1990) and second language acquisition (SLA) studies of language socialization (e.g., Ochs & Schieffelin, 1984), but also from a large body of CA work on social organization (Schegloff, Jefferson, & Sacks, 1977; Sidnell & Stivers, 2013) and L2 interactional development (Waring, 2018; Young, 2011). Hall (2018: 25) notes that the term *IC* is frequently used to refer to two rather different things, namely, "the underlying competence of learners on which they draw to participate in their interactions, their interactional competence," as well as "the diverse language-specific linguistic forms and other resources that learners develop from their participation in diverse social contexts over time." In teaching and assessing L2 IC, this argument is particularly important, as test constructors and raters face the task of deciding whether a particular assessment format captures a kind of general human knowledge of social organization and interaction along with L2-specific practices that learners have acquired.

In what Kasper and Ross (2013) refer to as a "rather obvious paraphrase," IC is, simply, the "competence to participate in interaction" (p. 9). When participating in interactional events, people "coordinate their actions, make these recognizable to each other and continuously adapt them according to the needs of the moment" (Pekarek Doehler & Petitjean, 2017: 6). Since adaptation to the setting, the current activities and the local context are central to 'competent' participation, interactional practices are both "context-independent . . . and context-sensitive" (Pekarek Doehler & Petitjean, 2017: 7). If IC is viewed from a CA perspective on the orderliness and normativity of interactional organization,

IC represents the "the ability to use the context-free interactional organizations . . . in a context-sensitive manner to participate in social activities" (Kasper & Ross, 2013: 24). Consequently, IC development, whether in an L1 or L2, means the application of a universal competence, and the context-sensitive deployment of practices in particular settings. For professionals in institutional interactions, becoming an expert means learning to manage and adapt to the specifics of a particular discursive practice (Ngyuen, 2012: 4–5), and for young people, the path between school and work requires the acquisition of "a growing degree of interactional flexibility" (Pekarek Doehler & Petitjean, 2017: 14) and an ability to revise their conduct in relation to particular institutional practices. This also applies to the transfer of knowledge of interaction into L2 contexts, or as Pekarek-Doehler (2018: 6) puts it, L2 learners "draw on interactional abilities . . . developed since infancy to deal with generic features of social interaction, yet they also recalibrate these in the course of L2 development."

The IC concept thus has its challenges for teaching and assessment, such as (1) how IC can be defined scientifically[1] (see Hall, 2018), (2) what L2 IC entails in practice in terms of its teachability (Waring, 2018) and (3) what L2 IC entails in practice for the testing of L2 interaction, as it is difficult to separate L2 IC from a general (L1 or 'social') competence in creating assessment tasks and scoring rubrics. Many promising attempts to define and specify IC-relevant conduct for testing purposes have been made, some of which we examine next.

Testing L2 IC

L2 speaking is often referred to as one of the 'complex' language abilities because of the "individualized uniqueness and complexity" (Wang, 2010: 108) of the tasks and performances to be assessed. While there are a number of empirical studies on L2 IC in non-testing contexts, such as classrooms, institutions or study-abroad contexts (Gramkow, 2000; Hall, 1995; Masuda, 2011; Walsh, 2012), research on L2 IC in testing contexts has grown steadily since Young and He's (1998) seminal collection on interaction in language proficiency interviews. Expanding on Canale and Swain's (1980) concept of communicative competence, He and Young (1998) emphasized the joint management of communication as central to IC, underlining that IC is *not* "a trait that is independent of the interactive practice in which it is (or is not) constituted" (p. 7), and called for an approach to oral testing based on how test participants display their understanding of the situated practices in which they participate.

In the context of formal educational assessment of oral proficiency, interactional abilities usually constitute part of the assessment construct (see, e.g., Kasper & Ross, 2013; Kramsch, 1986; Young & He, 1998). IC forms "an integral aspect of test-taking, especially with regard to

oral tests of language abilities" (van Compernolle, 2013: 327). This is also evident in the large bulk of studies on L2 speaking tests informed by CA, which, to different degrees, deal with how test takers display L2 interactional skills in assessment contexts (see, e.g., Greer & Potter, 2008; Kasper & Ross, 2013). CA-based research on L2 tests of interaction covers different assessment formats, from well-known L2 tests like the International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL) to unique, national or course-specific test formats (for a recent review, see Sandlund, Sundqvist, & Nyroos, 2016). Not all test constructs emphasize the same aspects; for instance, some focus on L2 pragmatics (Kasper & Ross, 2013), others on oral proficiency (Seedhouse, 2013) and oral production and interaction (Sandlund & Sundqvist, 2017).

As all common L2 oral test formats—oral proficiency interviews, paired tests and small-group tests—involve at least two interlocutors, the issue of co-constructed interaction and individual assessment has long been a concern for research on L2 IC testing (Chalhoub-Deville & Deville, 2005; Young, 2011). It has even been argued that since IC is not "an attribute of an individual participant . . . we cannot say that an individual is interactionally competent; rather we talk of interactional competence as something that is jointly constructed by all participants" (He & Young, 1998: 7). In this view, IC is how linguistic and interactional resources are "employed mutually and reciprocally by all participants in a particular discursive practice" (Young, 2011: 428). Along these lines, it has also been suggested that "the awarding of shared scores for interactional competence is one way of acknowledging the inherently co-constructed nature of interaction in a paired speaking test" (May, 2009: 397).

Tests of L2 IC generally aim to provide a measure of competence that can be transferred to other, non-testing contexts, along the lines of what Waring (2018: 58) refers to as "capturing and foregrounding what enables successful interaction in the real-world." However, studies of test interaction have shown how the testing context entails domain-specific demands on talk, and, consequently, on the IC that test takers *can* display in particular test formats (cf. van Compernolle, 2011). Depending on the similarities between non-testing (target competence) contexts and the "discursive architecture of language testing practices" (Young, 2011: 440), the IC tested may be considered as transferable to, or distinctly different from, other contexts. It has also been argued that while L2 oral tests are designed for obtaining assessable output, there is no requirement to accomplish intersubjectivity in, for example, oral proficiency interviews (Seedhouse, 2013), whereas participants in everyday conversation are persistent in maintaining understanding through practices like repair (Hayashi, Raymond, & Sidnell, 2013).

Studies on L2 test interaction have examined the kinds of conduct that learners display in tests that can be said to constitute evidence of L2 IC

or IC development (Butler & Zeng, 2014; Gan, 2010; van Compernolle, 2011). Aside from examining how learners orient to the interlocutor, questions, tasks, turn-taking and mutuality (e.g., Kasper & Ross, 2007; Nakatsuhara, 2008; Okada & Greer, 2013; Sandlund & Sundqvist, 2013), a common approach is to compare scores with interaction and look for interactional features surfacing at different proficiency levels (Davis, 2009; Ducasse & Brown, 2009; Galaczi, 2014; May, 2009; Sandlund & Sundqvist, 2011).

Last, it is worth noting that while constructs and rubrics may be based on specific definitions of L2 IC, participants in situated test interaction deal with contingencies arising within particular moments. As such, studies on authentic L2 test interactions, together with the bulk of knowledge on interactional organization (e.g., Sidnell & Stivers, 2013), are well suited for contributing to test construction and the development of IC assessment criteria (cf. Schegloff, Koshik, Jacoby, & Olsher, 2002: 15). However, the real-life consequences for test takers—their scores—are highly dependent on how raters construe IC from scoring rubrics and how the rubrics are applied to actual learner performances. Next, we address some rater aspects of L2 IC assessment.

Assessing L2 Interaction: Rater Interpretations

Assessing broad language abilities such as oral proficiency and interaction in an L2 is no straightforward task, and raters' (or teachers') professional judgments "will inevitably be complex and involve acts of interpretation on the part of the rater, and thus be subject to disagreement" (McNamara, 1996: 117). Therefore, Ducasse and Brown (2009) argue for studies of raters' perceptions of L2 test interactions:

> While it is important to understand the impact that peer-to-peer tests have on the language performance of the learners, it is also important to investigate what language experts—teachers and assessors—value while rating pairs, because it is their view of interaction which finds its reflection in the test scores.
>
> (Ducasse & Brown, 2009: 425)

In the context of paired tests of Spanish, Ducasse and Brown (2009) asked teachers/raters to view video recordings of tests in full and record their impression of the interaction. In a second viewing, raters were asked to pause the test recording and record their own comments on particular aspects of the interaction (i.e., a think-aloud protocol). The raters' views were then coded, 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). Ducasse and Brown (2009)

recommend that their findings on what raters perceive as proficient conduct in interaction be considered in constructs for paired tests.

Ducasse and Brown's (2009) findings on the saliency of embodied action align with those of May (2009, 2011), who admits that in terms of rating scales, there would have to be agreement on what exactly constitutes effective body language, and that perhaps "this Pandora's box has remained closed for very good reasons" (May, 2011: 140). May's studies (2009, 2011) also aimed to identify features of IC salient to raters and to inform the development of IC rating scales. In May's (2009) study, it was evident that raters struggled with individual assessment of IC in asymmetric interactions, where it became difficult to "unravel the impact of the dominant candidate on her partner's opportunity to display his interactional competence" (p. 404). It was also concluded that as raters perceived many aspects of the interaction as mutual achievements, it seems "counter-intuitive" (p. 417) to rate candidates individually on these. Additionally, May (2011) observed that raters tended to compare candidates relative to each other rather than against rating scales in their assessment (cf. also Sandlund & Sundqvist, 2016).

Other studies dealing with the development of rating scales for interaction include, for example, Youn's (2015) analysis of raters' application of data-driven rating criteria on role-play assessments, and Walters's (2007) pilot study of CA-trained raters' assessment of L2 pragmatic competence, which included post-rating "hermeneutic dialogues" (p. 169) between raters, aiming to uncover their grounds for assessment and to resolve rating differences. These studies are part of construct development efforts, whereas the study presented here aims at uncovering the *interactional* accomplishment of collaborative L2 IC assessment (*moderation*; cf. Sadler, 2013), using fixed rating criteria in a high-stakes national test.

Nevertheless, the studies of raters and the rating of IC reviewed above show the value of examining not only test interaction data but also rater data in understanding IC assessment. To our knowledge, no studies of L2 IC to date have approached raters' situated cognition in rater moderation group discussions—a practice for educational assessment often recommended, but seldom researched (Jönsson & Thornberg, 2014).

Assessment Practices: Moderation and Collaboration

Research on educational assessment has repeatedly emphasized issues of rater variability as a problem for assessment equity (cf. Meadows & Billington, 2005; Wang, 2010). A remedy often brought forth is rater collaboration through collegial discussions, so-called *collaborative assessment* practices (see, e.g., Erickson, 2009; Swedish Schools Inspectorate, 2013) or *consensus moderation* (Sadler, 2013). In a review of collaborative assessment research, Jönsson and Thornberg (2014) account for reasons why teachers/raters should work together on assessment: for

increased rater consistency and equity, for increased agreement on what assessment criteria mean, for professional development purposes and for formative assessment in the classroom (pp. 390–394). Through negotiations about authentic student performances, a space for reaching a shared understanding of assessment criteria opens up (Jönsson & Thornberg, 2014: 392).

As Sadler (2013: 5) notes, for moderation to be productive, "delving below the marks to find the reasons for the differences" is key to understanding and leveling rater inconsistencies. As such, rater discussions on selected student performances may function as teacher learning communities (Wiliam, 2007) where extreme positions can be smoothed out, and where raters have to account clearly for 'evidence' underlying their assessment decisions (Adie, Klenowski, & Wyatt-Smith, 2012). Empirical research on rater moderation, or *collaborative assessment* (henceforth CASS) practices as a social, interactional process, is scarce. As with any type of institutional group decision-making process, we can expect that CASS interactions contain agreement and disagreement, negotiations of what to base decisions on, and the supporting of arguments with evidence (cf., e.g., Halvorsen & Sarangi, 2015). The current study is one attempt to access raters' orientations to learner performance, displayed in such negotiations.

Data, Design and Methodology

The current study is informed by research on L2 test interaction and rater perceptions of interaction. We adopt a CA approach to rater interaction (i.e., the CASS moderation) and learners' test interaction. The study treats IC in L2 oral tests as members' accomplishment in rater interactions and the learner test interaction respectively, and specifically aims to uncover (a) differences in how particular social actions in test talk are oriented to by different raters and rater groups and (b) the function of these actions in the test sequence. Research questions guiding the study are:

1. In what ways do teachers participating in CASS meetings collaboratively construct IC when reporting on particular learner turns as arguments in their assessments?
2. What interactional work was accomplished by the turns reported in the original paired test interaction?
3. What can a comparison between the IC constructed by raters and an analysis of the turns in situated talk reveal about raters' perception of competent interactional conduct in an L2 test?

The study draws upon two data sets: video-recorded interactions between groups of teachers as raters (CASS groups) assessing L2 interaction and an audio-recorded L2 speaking test featuring two learners that the teacher groups discuss. Figure 13.1 shows the study design and the data we analyze.
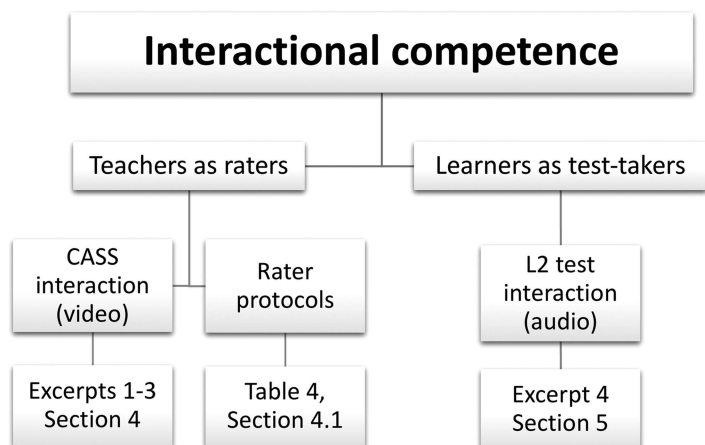
Interactional competence

```
                    ┌─────────────────┐
                    │ Teachers as     │
                    │ raters          │
                    └─────────────────┘
```


Figure 13.1 Study design and data

In what follows, we detail the study participants; the test of L2 oral production and interaction that the teachers are collaboratively assessing, i.e., the National English Speaking Test (NEST) in Sweden (Sundqvist, Wikström, Sandlund, & Nyroos, 2018; Swedish National Agency for Education, 2013); and the analytic procedures. As the analysis will demonstrate, IC as constructed by raters instructed to ground their decisions in given assessment criteria becomes delimited to learners' paired talk. This results in negative evaluations of learner IC with reference to the learners' interaction with the co-present *teacher* (who is an active participant in this particular test). Furthermore, the rater groups' constructions of IC are restricted by their attempts to stay close to the descriptors in the scoring rubrics, which leave room for interpretation.

Data and Participants

The principal data set consists of four video recordings of L2 English teachers doing CASS of a paired speaking test (the NEST). The teachers were attendees at a professional development day for English teachers organized by a university research center. The teachers had volunteered for a workshop on CASS of L2 speaking. All (*N* = 13) had a teacher degree in English and worked at different schools. All had been working for more than four years, except for one, who had only worked as an English teacher for six months (range 6 months–25 years, mean number of working years = 13). All were familiar with the NEST—only one teacher stated that she had *not* graded the NEST previously. Participants had not received any formal training on interaction or IC prior to the workshop, but most were familiar with the test construct. They were also provided with the NEST assessment materials, including tasks and assessment instructions.

Table 13.1 Participants in the CASS groups

CASS group	Participants (pseudonyms)	Recording length
Group 1	Brianna (BRI), May (MAY), Sarah (SAR), Terri (TER)	52 min
Group 2	Kari (KAR), Annie (ANN), Lise (LIS)	51 min
Group 3	Victoria (VIC), Alison (ALI), Katherine (KAT)	42 min
Group 4	Jessica (JES), Max (MAX), Beth (BET)	55 min

Additionally, a second data set is a test recording from the NEST featuring two learners (Fred and Henrik, year 9) and their co-present teacher (Kajsa). The audio-recorded test is part of a corpus of 71 high-stakes NESTs (see below). After presentation of the setup, the teachers (henceforth *raters*) listened to one test recording in full, while taking notes and making initial assessments individually (cf. May, 2011). Raters were then divided into four groups, combined so that they consisted of teachers from different schools, and each group was assigned a separate room. The NEST recording was made available via a web link so that the groups, if they wanted, could choose to re-listen to parts of the recording together on a smartphone. CASS meetings were video-recorded in their entirety. Meetings ranged between 42 and 55 minutes (see Table 13.1), and the researchers were not present.

The NEST in Sweden

The authority responsible for national tests is the Swedish National Agency for Education. The national test of English is a mandatory proficiency test taken in year 9 of compulsory school (ages 15–16). The test consists of three parts: part A tests speaking, part B tests receptive skills (listening and reading comprehension), and part C tests writing. The learners' English teacher is responsible for preparing, administering and assessing performance on all three parts. Using teachers as examiners for high-stakes standardized tests (rather than external examiners) has a long tradition in Sweden (Erickson & Börjesson, 2001). Consequently, the governing authority puts great faith in teachers' professionalism; however, the setup can be a challenge in terms of achieving standardization and equity (Sundqvist et al., 2018; Swedish Schools Inspectorate, 2013).

In preparation for the national test, teachers receive an instruction booklet that includes information about all three test parts. For part A (the NEST), instructions encompass details on grading, and there is an accompanying CD with sample recordings of test-taker performances corresponding to different grades, also described in the booklet with reference to criteria. Thus, the sample performances serve as benchmarks (National Assessment Project, 2015). The instructions stipulate that the examining teacher should remain in the background of the test takers'

conversation, but they are allowed to help out should the learners run into difficulties (Swedish National Agency for Education, 2013).

NEST tasks are designed for interaction in pairs or small groups. The test begins with a warmup task where students take turns producing descriptive talk on a topic (e.g., their home environment or a picture prompt), followed by more advanced tasks aiming to generate peer interaction using *topic cards*. Each topic card lists either a statement with instructions (e.g., *"Money makes people happy. Agree? Disagree? Explain why and give examples."*) or a question (e.g., *"What can you do to protect the environment? Discuss your own opinions about this question and also what other opinions there are."*). Learners then take turns drawing topic cards and initiating talk on each topic. For the test in the present study, there is a secrecy bond until 2020, which means that no exact formulations of tasks in the test recording may be revealed.

The test construct targets learners' "oral production and interaction" (Swedish National Agency for Education, 2013: 28). Teachers are instructed to award each performance a holistic grade from A to F, based on a number of assessment descriptors. In the booklet, there are two main headings, "Content" and "Language and expression" (referred to as "Assessment factors"), and under each, more specific assessment descriptors are listed (see Figure 13.2) (Swedish National Agency for Education, 2013; translated from Swedish).

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

Figure 13.2 NEST assessment descriptors

To assist teachers in NEST assessment, the booklet also contains extracts from the current curriculum for English. The descriptors denote what is required for goal attainment for the grades E, C and A in English at the end of year 9 (encompassing all aspects of English proficiency, not only speaking). Grade D is awarded if all criteria for grade E plus most (but not all) for grade C are attained. Similarly, grade B is awarded if all criteria for grade C plus most (but not all) for grade A are attained. Grade F is assigned when passing criteria are not met. When teachers have decided on a grade, they record the grade by ticking a box (Figure 13.3). As indicated by the dotted lines, grades F through C can be either 'minus' or 'plus,' whereas no such evaluative distinction is made for grades B and A. In other words, although a 6-grade scale for *grades* is employed, the actual scoring of NEST is made on a 10-grade scale.

The grading criteria are largely aligned with the descriptors of the communicative abilities described in the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). A passing grade in English in year 9 corresponds to level B1.1 ('Independent user') in CEFR (Council of Europe, 2001: 34). However, a majority of learners in Sweden receive higher final grades in English. National statistics from 2016 regarding NEST reveal that 22.4% of the whole student population scored an 'A' and very few failed, only 3.6% (www.skolverket.se/statistik-och-utvardering). In the 2011 European Survey on Language Competence, encompassing 53,000 students from 14 countries tested on comprehension, Swedish ninth-grade students scored the highest in English. A great majority were at level B2, and some even at level C1 ('Proficient users' in CEFR) (www.survey-lang.org). In other words, L2 English oral proficiency is generally high among Swedish youth.

Part A- Focus: Speaking					
F	E	D	C	B	A

Figure 13.3 Assessment of NEST grades (teachers tick one box)

CASS Workshop Design: Selecting a Test Recording

The test recording was collected as part of a project on interaction in L2 tests, funded by the Swedish Research Council (reg. no 2012–4129). The corpus was collected in 2014 at four schools, with 161 learners in a total of 71 audio-recorded tests. All test takers had been assessed by their own teacher (also a participant in the recordings) and two external raters (based on recordings). Interrater reliability between the three raters was high (r_s = .787**, .836**, .872**, Cronbach's alpha = .942).

In order to select a test for the CASS workshops, data were scanned for learners who had been given grades that differed by *at least two grade steps* by *at least two of the three raters*. The rating discrepancies indicated to us that these performances were perceived differently by the original raters, and these discrepancies were hypothesized to promote rater moderation (cf. Sadler, 2013) and negotiation (cf. Falk & Ort, 1998). Application of the grade variation criteria to the corpus yielded a subcorpus of 31 test takers in 23 NEST recordings (see Table 13.2).

Most of the diverging grades (42%) were found in the C to A continuum. Diverging assessments were found in equal amounts in clusters D–B and E–C (26%). Findings reported by Millet (2018) suggest that it is easier for raters to assign higher grades than intermediate grades. A hypothesis, then, was that performances at the extremes of the continuum (such as F, E and A) would be easier for teachers to agree upon. Furthermore, since the curriculum descriptors included descriptive texts only for the grades A, C and E, it was hypothesized that CASS discussions would possibly be richer if based on a cluster 'in between' the curriculum descriptors. Together, these assumptions motivated a focus on the D–B cluster.

Having weighed concerns such as recording quality, one recording was selected, featuring two male learners, Henrik and Fred (duration: 21 minutes and 13 seconds). Their teacher, Kajsa, was present, and as it turned

Table 13.2 Grade variation in the subcorpus

Grade difference	Test takers (N)	%
F–D	1	3
E–C	8	26
D–A	1	3
D–B	8	26
C–A	13	42
Total	**31**	**100**

out, she was often an active participant in the test interaction. In order to ensure participants' anonymity, the recording was anonymized using slight pitch editing, and bleeping was added when names and places were mentioned. The CASS raters had no demographic knowledge of the students, and only had access to the anonymized recording. In authentic NEST assessment, the grading teacher also has access to students' visual orientations and embodied action, which can be a limitation in assessing the learners for the CASS raters (cf. May, 2009). However, as CASS is usually done with audio recordings only, we consider this limitation acceptable for the purpose of this study.

The test used had a grade divergence for both learners, where one fit the criteria of a difference of *at least two grade steps by at least two of the three raters*. The grades assigned in the original data set by Kajsa and two external raters are presented in Table 13.3. As shown, the greatest difference lies between Kajsa and rater 2 on Fred's performance.

Table 13.3 Teacher and external rater grades for focal recording

Rater	Henrik's grade	Fred's grade
Teacher	B	B
External rater 1	C+	C+
External rater 2	C–	D+

CASS Group Assessments

The full recording was played, and all raters were instructed to take notes on an assessment form, including assigning a tentative grade to each of the two learners. Subsequently, raters, divided into groups, were instructed to discuss the learner performances and reach consensus on grades. All groups handed in a joint rater protocol with a written motivation for their decisions. As expected, the CASS groups also assigned varying grades (Table 13.4).

Table 13.4 CASS groups and assigned grades

CASS group	Teachers (N)	Teachers' individual grades: Henrik				Teachers' individual grades: Fred				CASS grade: Henrik	CASS grade: Fred
1	4	C+	C-	C	D	D-	D-	D	E	C+	D-
2	3	D+	D	D	n/a	D-	E	E+	n/a	D+	D-
3	3	B	B	C+	n/a	D+	D+	C-	n/a	C+	D+
4	3	B	B	C+	n/a	D+	D+	C-	n/a	C+	C-
Total	13										

It is worth noting how all groups ended up with consensus grades of C or D for both students, while the initial, individually assigned grades varied greatly. Thus, it can be assumed that CASS negotiations of raters' perceptions of the learner performances had taken place.

Analytic Focus

The study aims to uncover how CASS raters collaboratively display their understanding of IC. This is accomplished by (1) examining raters' overall views on interaction in the test recording, and by (2) specifically focusing on sequences where raters report on particular student contributions in qualifying their assessments in relation to scoring rubrics. Thus, Step 2 focuses on RS in rater interaction, such as in Example 1. Here, raters of CASS Group 1 (Brianna, May, Sarah and Terri) are discussing an instance where the learners were discussing the moral dilemma of finding out that a friend was involved with drugs. The learner sequence in question is briefly recapitulated by May (line 333) until shared understanding is displayed by co-participants:

Example 1: Speech report [File_CASS Group 1_333–337]

```
333  MAY    e:h säga prata med kompisar [om] droger
             E:h say talk  to   friends   [about] drugs

334  BRI                               [precis]
                                       [exactly]

335  TER    mm:
336  BRI    mm:

337  MAY  → you can at least try

338  SAR    m:

339  MAY    tyckte Henrik då
            thought Henrik then
```

In line 337, then, May launches a turn ("you can at least try") produced in English, indicating that it is reporting what Henrik had said in the test recording. The turn is also positioned as RS in line 339 with the verb *thought*, which here functions as a verb of speaking (cf. Sandlund, 2014). In Example 1, the RS has an illustrative function, displaying May's perception of something specific in the recording worthy of

bringing into the assessment discussion. Turns like this have guided our data search.

Research on RS in interaction (e.g., Drew, 1998; Goffman, 1974) has shown that people report on talk for a variety of interactional purposes, such as making jokes, assessments and complaints (Clift & Holt, 2007). We find reports of talk to be of particular interest when comparing the rater interactions with the learners' test interaction. In the rater talk, RS represents the gap between situated interaction (i.e., original turns produced as part of 'doing test interaction') and the recontextualization of such turns in an interaction where they may serve as qualifications for assessment. In contrast to raters' general discussion of learners' IC, RS instances do the work of re-opening the test interaction for joint scrutiny. While RS turns are always 'inauthentic' in relation to the original, situated talk reported (cf. Clift & Holt, 2007), the reports in our data become part of negotiation sequences, and thereby built into the local assessment machinery as well as into the wider context of teachers' professional judgment (cf. Allal, 2013). RS of test takers' talk brought forth as evidence in negotiating a particular assessment of learners' displayed IC, and thus raters' treatment of test-taker actions, may be a rich source for understanding IC assessment.

After rater protocols and overall evaluations displayed in rater discussions were examined (Step 1), CASS recordings were marked for all instances in which the raters reported on test takers' turns.[2] The reported-on talk was then located in the learners' test talk, yielding a collection of RS in rater interaction along with the corresponding learner sequences. The sequences were analyzed sequentially in order to establish the work accomplished by RS on learner talk in the rater talk (Step 2). The next step was to examine the students' interaction in the sequences reported on (Step 3). For the purpose of this chapter, one instance in the students' talk that surfaced in all four rater groups was selected for presentation.

Transcription

Audio and video data were transcribed according to CA conventions (Hepburn & Bolden, 2013; see the appendix to this volume). Embodied conduct in video recordings is included in transcripts in the form of still images and on separate lines below the transcription of talk. The plus sign, +, marks the onset and termination of embodied action in relation to vocal/verbal actions. Double parentheses indicate the transcriptionist's scenic descriptions not captured in symbols or images. A parenthesis with three periods (. . .) at the beginning or end of a transcript indicates that immediately prior or subsequent parts of an ongoing (lengthy) turn have been omitted.

IC as Constructed by Raters

Before examining the RS turns, we situate our analysis in the raters' views on the test interaction as a whole. Thus, our first analytic presentation examines data from the written rater protocols and one excerpt from CASS Group 3.

"No Interaction": CASS Reflections on the Test as a Whole

The Rater Group Protocols

First, it should be mentioned that all four rater groups commented on the lack of interaction between Henrik and Fred throughout. In group rater protocols,[3] Group 1 had noted "no interaction" as a minus for both learners. Group 2 wrote "weak interaction" about Henrik's performance, and "no use of strategies, question-answer, used Swedish a few times" about Fred. Group 3 noted that Henrik displayed strategies to overcome interactional problems, but that Fred showed "an unwillingness to take initiative and carry the conversation forward." Group 4 wrote that the teacher should have steered the test interaction so that "the students *interact*—talk to each other." Consequently, the rater groups were in agreement about the insufficient degree of interaction between the learners, something we will return to via Excerpt 1 below.

"It's Not a Conversation Between Them": Interaction in Group 3

The rater groups deal extensively with the lack of interaction indicated in the written protocols, as illustrated in Excerpt 1 (Group 3). Prior to line 582, the raters have been re-listening to parts of the recording and commenting on how the learners seem unwilling to communicate with each other, and how the teacher had to repeatedly prompt them. Here, a formulation from the assessment materials, i.e., the ability to develop and carry the conversation forward, is discussed in relation to what the learners display. As we enter, Victoria has made it known that while she accepts that one test taker is weaker than the other, he also displays "good things" in parts of the test (line 583). Katherine does not affiliate with Victoria, but instead refutes the assessment by claiming that Fred does not "do anything on his own," constructing the displayed ability to take 'independent initiative' as central to a positive assessment of *communicative strategies* (line 584). After Victoria has acknowledged agreement to this claim, Katherine elaborates by referring to the scoring rubrics (lines 589–590) in supporting her claim about initiative.

Excerpt 1: "On his own initiative" [CASS3_00001_07.00 - 07.23]

582	VIC	(...) att han som ä <↑svag> säjer ju[1]
		the one who is <↑weak> do still say
583		ju ändå ↑<u>bra</u> grejer i↑bl<u>a</u>nd
		↑<u>good</u> things on ↑occasion
584	KAT	ja men han (0.3) han gör ju inget på egen hand
		yes but he (0.3) he doesn't do anything on his own
585	VIC	eget bevå:g nej
		on his own initiative no
586	KAT	ne:j.=
		No:.
587		(0.3)
588	VIC	=nej.
		no.
589	KAT	<u>ing</u>et på eget- (0.3) och det är ju <u>ett</u> utav
		nothing on his- (0.3) and that's *ju* <u>one</u> of the
		PRT
590		kriterierna +här att du t ar [<initia↑tiv>
		+*lifts up booklet))*
		criteria here that you take the [<initiative>

591	VIC	[att man interag<u>e</u>rar ja
		[that you interact yeah

```
592    KAT      och att du för sam[talet] <vi:dare↑>
                and that you carry the con[versation] forward↑
593    ALI                          [m:        ]
594    ALI      m:

595             (0.8)

596    KAT      .hh nu tar ju inte den andra heller #e::#
                .hh now the other one doesn't either #e::#

597             han interagerar ju inte heller
                he doesn't interact either

598    VIC      ne:j dom <saknar> de
                No: they <lack> that

599             (0.6)

600    KAT      dom pr- (.)  +dom (0.2) det är inget samtal
                they pr- (.) +they (0.2) it's not a conversation
                             +gaze turns to VIC
601             (0.4)

602             mellan dom?
                between them?

603    VIC      ne:ej+
                No:o +
                     +shakes head twice
604    KAT      utan det är ett samtal mellan,
                instead it's a conversation between,

605             (0.2)

606    VER      det blir som [en inter↑vju::]på nå vis
                it's like [an interview] in a way

607    KAT                   [m e d    läraren]
                             [with the teacher]

608    KAT      de blir som en +intervju:.
                it's  like  an +interview.
                               +lifts up hands
609    VER      ja:
                Ye:ah
```

Interestingly, *taking the initiative* and *carrying the conversation forward* are not listed in the booklet (Figure 13.2). Instead, they are mentioned on a separate sheet of paper accompanying the topic cards for part A. We cannot determine from the video whether Katherine points to the booklet or the task package, but, regardless, she is orienting to the test materials as evidence for why *taking the initiative* is central to the construct, thus a valid claim on her part. Victoria signals agreement with Katherine's account by making public a connection between Katherine's claim of initiative and her own formulation "that you interact yeah" (line 591). By completing Katherine's turn in overlap, Victoria is co-constructing the claim that *initiative* is relevant to the interaction construct, and she also adds the second aspect from the card package: "that you carry the conversation forward"—something clearly treated as not displayed by Fred. Katherine remarks that these features apply to Henrik as well (line 597), to which Victoria makes reference to a lack of something, presumably referring to the learners' displayed IC in the test as a whole (line 598). Katherine produces an upshot of prior talk (600) with the formulation "it's not a *conversation* (0.4) between them" and "instead it's a conversation between." Victoria completes her turn with a formulation: "it's like an interview in a way," after which Katherine continues her turn ("with the teacher") and recycles Victoria's proposed description with falling, 'closing' intonation (line 608).

In Excerpt 1, the raters collaboratively connect the assessment descriptors *take the initiative* and *carry the conversation forward* to the lack of these features in learner interaction. They also display that they partially ground this evaluation in observations of how the teacher had to intervene repeatedly, resulting in an interview format rather than a paired interaction. While raters display dissatisfaction with the students' displayed IC, they are also working in a stepwise fashion towards a shared understanding of features of the interaction and the assessment task, and they display orientation to the assessment descriptors in establishing a joint evaluation. However, what we do *not* see here is the raters linking particular learner actions to IC assessment. In pursuit of such evidence, we next present two selected sequences in which rater groups report on a particular action produced by Fred.

### *"My Mom": Raters' Reports of Talk as Evidence For/Against IC*

First, we note that Fred's action, on which the CASS groups report in Excerpts 2–3 below, was topicalized in all four rater groups. For the sake of presentation, we selected reports produced in two of the rater groups. We first examine how Group 4 (Jessica, Max and Beth) orients to a particular turn in a RS format. Then, we analyze how Group 2 (Kari, Annie and Lise) topicalize the same action in their talk.

"So Says My Mom": RS in CASS Group 4

We begin with a sequence that occurs late in the rater meeting where the raters are to make a joint decision on test grades. The raters have decided to go over the criteria one by one and sum up their impressions about the learners in relation to each criterion. Prior to Excerpt 2, Max has commented on the students' interaction, stating that the teacher in the recording perhaps should have steered the students more towards peer-to-peer interaction. Jessica, subsequently, has pointed to the topic cards, which all include the formulation *Discuss with your friend*, indicating to her that the learners know what is expected. In line 1739, Beth sums up the gist of their perception of the students' task management in the form of a polar question:

Excerpt 2: "So says my mom" [CASS Group 4_File_AA000210_03.40–03.56]

```
1739      BET      så att egentligen [a s s å diskuterar] dom=
                            so really [I mean do they discuss]=

1740      MAX                          [men det glöms bort]
                                       [but it's forgotten]

1741      BET      =↑nå'n (.) uppgift med var↑ann?
                   =any (.) task topic with each other?

1742               (0.6)

1743      MAX      .hhHHH vid nåt tillfälle har jag för mig
                   .hhHHH at some point I recall

1744               att dom    [g i c k    i n? ]
                   that they [w e n t    i n? ]

1745      BET           [>ja dom kommenterar] kanske<=
                        [>yeah they comment] maybe<=

1746      MAX      =ja::?=
                   =Ye:ah?=

1747      BET      =A: så säger mamma [min mamma=
        ➤         =A: so says mom    [my mom =
```


mom   my mom
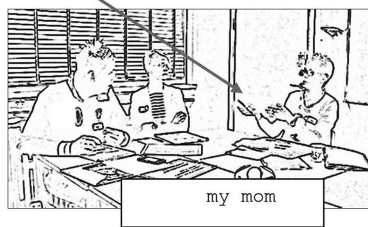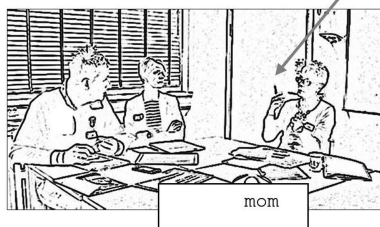
```
1748      MAX                          [ja:?
                                       [ye:s?


1749      BET          =tycker [också så +eller nå'
          ➡️           =thinks [also so  +or somethin'
                                        +lowers hands


1750      MAX              [ja <preci:s>
                           [yeah <exactly>

1751      JES          ja::,((nods three times))
                       Yea::,


1752      BET          [de var]
                       [it was]


1753      MAX          [ m::: ]
                       [ m::: ]

1754      BET          de här me dataspelande
                       this with computer gaming


1755      MAX          ja <preci:s.>
                       yeah <exactly.>

1756      JES          °m.°
1757                   (1.4)
1758      MAX          m.
1759                   (1.9)
```

Beth's turn (lines 1739, 1741) with an interrogative syntax ("do they discuss *any* (.) task topic with each other"; cf. Hayano, 2013: 396) has a built-in preference for an agreeing 'no' response, as it appears to assume that the number of task topics where learners actually talk to each other is very low. This is visible also in Beth's emphasis on "any," and the sequential placement of the turn after preceding talk about the students' insufficient paired interaction (Max's overlapping turn at line 1740 is actually a response to prior talk on whether the learners are following the instructions by discussing with each other). There is a 0.6-second gap, and together with Max's turn-initial in-breath, his response to Beth has the shape of a dispreferred turn. Thus, Max does not affiliate with Beth, but instead accounts for his recollection of one particular instance in the test where the students did, in fact, interact in a way that agrees with the group's understanding of the assessment criteria for *interaction*. He does

not specify the recalled instance, and as he drops out of the overlap with Beth's turn (line 1745), he only partly reveals that whatever the learners did ("that they went in?") counters Beth's extreme case formulation (cf. Edwards, 2000).

Beth's response shows that she understands Max's somewhat vague reference as being when "they comment maybe," and she proceeds to exemplify such conduct (lines 1747, 1749). The enactment of talk begins with "A: so says mom," accompanied by embodied action: Beth moves both hands to her right as she produces "mom." She then immediately produces a second speech report: "my mom thinks also so," which is accompanied by a mirroring of the first gesture, but this time to her left at the production of "my mom." The emphasis on the first "mom" and on the "my" in the second part creates a contrast, and together with her hand movements, Beth displays that she is in fact enacting both Fred and Henrik, but without specifically indicating which speaker is referenced in the separate enactments. Her report-final "or somethin'" indicates the possible inaccuracy of the enactments and makes it hearable as a non-verbatim example of an occasion when the learners did talk to each other. Max emphatically agrees (line 1750), and Jessica produces minimal acknowledgment (line 1751, while nodding), after which Beth offers further claims establishing agreement on the instance in question (line 1754, "computer gaming").

In all four CASS groups, the raters frequently evaluate displayed IC in general terms, without reference to particular contributions. What is particularly interesting in the cases where the raters *do* refer to particular turns is that a comparative analysis of IC as conceptualized by raters and the interactional work accomplished by the same turns in the situated test interaction becomes possible. As in Excerpt 1, we noted that raters in Excerpt 2 collaboratively questioned the degree of paired interaction between Fred and Henrik. However, in Excerpt 2, the RS turn is offered as a contrasting example of when the learners *did display* IC. When the speech report is made available, Max and Jessica can re-inspect it, and allow for it to pass as a representative example of IC. As such, the RS works as an understanding check, but also as a device in the social process of moderation (Jönsson & Thornberg, 2014; Sadler, 2013) where raters can probe their own and others' interpretation of what is to be assessed and reach a shared understanding of scoring rubrics.

"A Little Spontaneously": RS in CASS Group 2

In Excerpt 3, the same focal turn becomes topicalized. Again, the RS turn is located after the raters have discussed the lack of interactional initiative from the learners, and from Fred in particular. In a similar

sequential organization as in Excerpt 2, one participant produces a turn that shifts the focus from negative overall assessments to accounts of competent learner conduct (cf. Excerpt 2, "at some point I recall that they went in"). As we enter, Lise refers how Fred "got going" in a few places in the interaction (line 1139). The RS turn is produced by Kari in line 1148:

Excerpt 3: "Somewhat spontaneously" [CASS_Group_2_File_00007_04:40–05:00]

```
1139 LIS   men a::h? (0.3) lite grann (0.2) kom han igång [va=
           But yeah a little (.) he got going [right=


1140 KAR                                       [ja:?
                                               [Ye:ah?


1141 LIS   =men inte,
           =But not,


1142       (0.2)


1143 KAR   >för de va< de vart ju iallafall om de
           >Cuz it was< it was ju¹ at least if


1144       nu var han som ↑sa så >så där kom det ju
           it was him who ↑said that >so there it came out

1145       lite spont↑a:nt
           somewhat spont↑a:neously


1146 LIS   [m:]
1147 ANN   [m:]

1148 KAR → +not if you ask my: .hh mo[ther
           +  gestures with right hand
```



```
1149 LIS                               [ja
                                       [Yeah
```

```
1150 KAR   eller va han sa
           or what it was he said

1151 LIS   .hhja
           .hhyeh

1152       (0.9)

1153 KAR   ock då vart de lit- då kom <han>
           an' then it became a little- then came <he>

1154       (0.4)

1155       då ly:ste han   lite        [där å]
           then he was shi:ning a little [there an']

1156 ANN                               [ ja  ]
1157 LIS                               [ ja  ]

1158 KAR   sa någonting som inte va:, (0.7) >han han
           said something that wasn't wa:s, (0.7) >he he

1159 KAR   + knocks pen to temple twice

1160 KAR   [kom på det hä:r]
           [came up with thi:s]
```



```
1161 ANN   [han tänkte ↑själv]             [ja:?]
           [he thought for him↑self        [ye:s?]
```

Following Lise's comment, which is mitigated in line 1141 ("but not"), Kari picks up on the part of the test where Fred was more active, and through lines 1143–1145, she proposes a candidate identification of such a sequence, while also displaying some uncertainty as to whether it was actually Fred who "↑*said* that." Prior to line

1139, the group members have been attempting to determine whether it was Henrik who mentioned the singer Marilyn Manson and Fred who mentioned his mother, but only in such unspecific terms, so Kari is probably referring back to the general talk about "mom" or "Marilyn Manson." This is treated as unproblematic by the co-participants, and Kari assesses the hitherto unknown turn as produced "somewhat spont↑a:neously." In line 1148, the turn referred to is enacted: "not if you ask my:. hh mother" with the exactness of the report mitigated in line 1150. Thus, the RS makes the original turn available for recognition by co-participants.

Lise provides acknowledgment, treating Kari's report as still incomplete. As Kari's RS was produced in service of obtaining a shared understanding of which sequence they are currently discussing, an assessment of the report in connection with its identification may be expected. Kari does indeed elaborate on the representativeness of the reported turn with a positive assessment of Fred's local performance ("then he was shi:ning a little there," line 1153). She then accounts for the assessment: "said something that wasn't wa:s, (0.7) >he he," and appears to be searching for an appropriate description of the reported turn. Her search becomes fully available in line 1159 when she pauses and knocks her pen to her temple. Annie, in overlap with Kari's trouble resolution, offers an upshot of Fred's conduct: "he thought for him↑self ye:s?" (line 1161).

In Excerpt 3, the RS turn is first launched as part of a problem-solving sequence aiming to determine with certainty the respective contributions of Henrik and Fred in a part of the test where the raters felt there was more interaction. Kari uses the RS as support for her viewpoint that Fred displayed something distinctive in this sequence, namely, spontaneity. This is collaboratively built into several assessment-relevant turns (*shining*, *thought for himself*, *he came up with*), which all imply that taking the initiative to initiate topic elaborations is indicative of IC. The rater interaction thus reveals a view of the types of contributions that are valued for the construct 'interaction,' but also makes it clear that the occasion was uncharacteristic of Fred's overall performance.

Having examined the CASS sequences where raters draw on RS as part of their assessment negotiations, we now turn to the learner interaction sequence in question and examine the sequential environment for the production of the "mom" turn.

"You Haven't Met My Mom": Interactional Work in the Test Recording

As we enter the test interaction containing the "mom" contribution, Henrik has initiated talk on a new topic card with a statement to be agreed or disagreed with. We are not at liberty to reveal the exact formulation,

but the topic involves video-gaming. The card statement is formulated as an opinion, and test takers are instructed to agree/disagree with whether people worry excessively about teens' gaming habits, and account for why.

As we begin, Henrik introduces an old media debate concerning a murder case in the United States in which blame for the crime had been attributed to the perpetrator's love for video games and for the rock singer Marilyn Manson (lines 123–131). He displays disagreement with such views on video games (line 133). In line 136, Henrik supports his view by referring to how "many people" play video games (line 138), presumably referring to the fact that not all gamers commit violent crimes as a result. Henrik closes his topical talk in line 139 by referring back to the topic card statement, i.e., whether people worry too much about the issue, to which he displays disagreement:

Excerpt 4:1: "You haven't met my mom" [TT2014_File_32111231_08.02–09.55]

```
123  HEN   if its- we jzt had an example when
124        (therewz)a kid that wz playing like (0.5)
125        Marilyn Manson and CO:D
126        (0.2)
127        and he got- (.) he killed somebody and
128        >theywr like< ↑O:H it's only because he
129        listened to Marilyn Manson n: (0.8) .hh
130        ((thudding sound))
131         and played violent ge- >video games °an'°
132        (1.3)
133        don't th[ink THAT's THE ↑CA::SE=
134  TEA   =((coughs))
135        (0.3)
136  HEN   cuz there's there are many people playin'
137        (1.3)
138        >°videogames bud°< I: don't think
139        they caring s: so much about it s: really
140        .hh
141        (2.1)
142  TEA   Do you agr↑ee (.) Fred
143        (0.6)
144        that people care too much? .HHH
```

```
145        (1.0)
146  FRE   ye:h.
147        (0.4)
148  TEA   yeah?
149        (1.2)
150  FRE   If you <do:n't agree> (.) hhh
151  →     you haven't met my mom.
152  TEA   HHHh >heh heh °heh°
153  FRE   She: is
154        (0.7)
155  TEA   ((loud coughing))
156  FRE   #e: a ao:#
157        (0.8)
158  TEA   .pt sh'thinksh you play too ↑much?
159        (1.3)
160  FRE   Ye::h .hhh=
161  TEA   =Mmhmm?
162  TEA   Do you thi[nk (.) you play too much?
163  FRE            [°o(hh):r,°
164        (2.7)
165  FRE   #nn:::# #a::h#
166        (1.4)
167        not nowadays (0.3) but (.) I >↑did
168  TEA   Mmhmm?
```

In line 142, the teacher allocates the next turn to Fred, using a yes/no interrogative pinpointing the task instructions (i.e., agreement). Here, the teacher's intervention tells us that she is expecting Fred to continue, and when no response is forthcoming (line 141), she makes Fred accountable for a next action. In line 144, perhaps prompted by the 0.6-second silence, she clarifies for Fred what he is expected to agree on. In this particular turn, Kajsa narrows down the types of responses expected from Fred by asking for agreement with the *topic card* rather than with Henrik's turns. Consequently, Kajsa's turn here temporarily steers learners away from the paired format. Fred's agreeing response ("ye:h") is not immediately forthcoming but produced after another noticeable silence (line 145). The response, albeit minimal, is structurally aligned with the yes/no question posed, but the teacher repeats the agreement token with a rising intonation, indicating that she treats the response as incomplete. A silence follows, but Fred appears to read Kajsa's turn as a pursuit of an account,

and the focal turn for the present analysis is produced in lines 150–151: "If you <do:n't agree> (.) hhh you haven't met my mom." Fred's turn functions as an account for his agreement but has the design of a negated conditional statement. Consequently, Fred's turn not only addresses the card instructions, but also specifies the conditions for a possible disagreeing response—it would be impossible to disagree with the fact that people worry too much about teens and gaming if you knew Fred's mother.

Our audio-only access restricts a full analysis of Fred's turn (i.e., there may be non-vocal invitations to laughter), but Fred thus makes a rather sophisticated joke on the basis of the topic card instructions, which attends both to the teacher's expectation of an account and to the topic card statement. This indeed occasions laughter from Kajsa, and between lines 153 and 157 Fred appears to be initiating accounts for his mom's role in agreeing with the card, albeit with some trouble. Picking up on the displayed trouble, the teacher offers a candidate understanding (line 158): ". pt sh'thinksh you play too ↑much?." There is some delay (line 159) before Fred's agreeing response, and the teacher produces additional prompts (lines 161–162) as Fred's responses still indicate trouble in producing anything beyond agreement (line 160, silences in lines 159 and 164). Fred cuts off something and drops out of the overlap (line 163). As Kajsa completes her question turn (line 162), there is now a new context for Fred to respond to: whether Fred himself thinks that he plays "too much." Fred's drawn-out disagreeing response in 165 may reflect a space for processing the new question and preparing a response, or just a 'thinking space' for deciding on his opinion. A response is offered (line 167)—he does not play excessively nowadays but used to. Kajsa offers acknowledgment with a rising intonation, treating Fred's account as still unsatisfactory.

After a lengthy silence, Fred continues with an upshot: his mother is still "all over" him about gaming (line 172), as shown in Excerpt 4:2:

Excerpt 4:2: "You haven't met my mom" [TT2014_File_32111231_08.02–09.55]

```
169          (2.1)
170  FRE     But she's ↑STILL hhh
171          (1.8)
172          is: (0.7) all over me
173  TEA     >huh huh< .HHH (.) do you think e:hm
174          (0.6)
175  HEN     >ye:ah if-<
176  TEA     why: do you think she's
177          (2.2)
178  FRE     #°Hhhmm°#
179  TEA     she's talking with you about those things
180          all the time
```

```
181  FRE    She's talking about (0.7)h- how it
182         affects me
183         .hhh all the time.
184         (0.6)
185  TEA    Mmhmm?
186         (2.2)
187  FRE    Sure I:? can't feel it but (1.5) I'm a
188         hundred percent su:re (.) that it affects
189         me n- in ↑some ways,
190  TEA    Mmhmm?
191         (0.6)
192  FRE    buddit haven't- (0.4) hasn't have to be a
193         ↑bad way.
194  TEA    Mmhmm?
```

As we can see, Kajsa produces a new question (initiated in line 173 and restarted in line 176) with a notable silence before its completion in line 179. The *why* question is shaped for an answer in the form of an account for Fred's mother's concern, and Fred aligns with this preference in his response (lines 181–183), where he reveals that his mother talks about how gaming "affects" him "all the time." Kajsa's continued receipt with rising intonation (185) results in an assessment of his mother's stance, where he acknowledges likely effects that are not necessarily negative (lines 187–189, 192–193). The exchange continues for another 16 lines (omitted here) until Henrik, in lines 211–212, joins in with a view revised from what he had expressed at the beginning, namely, that he agrees that mothers worry too much, but not the "rest of the world" (lines 215–216):

Excerpt 4:2: "You haven't met my mom" [TT2014_File_3211 1231_08.02–09.55]

```
            ((16 lines omitted))
211  HEN    yeah if we are going to talk about the
212  →      MO:MS, the Hhh
213  TEA    °Hhhh°
214  HEN     then a:h the- (.) they care too much but
215         (0.4) I don't think (.) the ↑rest of
216         the world,
```

Henrik thus shows that he has been monitoring his co-participants' talk closely. Thereby, he is able to make a contribution that reaches back to his own prior talk and to fit his new turn into the interactional trajectory that has transpired between Fred and Kajsa.

The "mom" turn, then, is produced as an account for *agreeing* with the topic statement for which Henrik had reported a disagreeing stance. It also brings something new to the ongoing interaction in the sense that it is brief enough to warrant more probing or elaboration, but still attends to the request for additional accounting implicit in Kajsa's "yeah?" response (line 148). Furthermore, it also displays disagreement with Henrik's stance on the card statement, facilitated by the fact that Fred does not address Henrik, but rather Kajsa's question, and he also attributes his disagreement to a third party: his mom. The trajectory that follows shows Fred producing additional talk on the topic when prompted by Kajsa, and while he does tailor his responses in structural alignment with the questions, he also displays an understanding of the ongoing interaction as a matter of awaiting prompts and aligning with them, rather than producing more independent topical contributions. Fred thus is orienting to local contingencies (the local establishment of an interview format) in an interactionally competent way, but perhaps not as much to the communicative event as a peer-to-peer interaction where he is assessed on how he orients to his co-test taker (cf. Sandlund & Sundqvist, 2011). Proficiency-wise, Fred's more extended turns show competent use of L2 vocabulary and turn design, and Henrik, who has appeared to treat the latter part of the sequence as being between Fred and Kajsa, displays a close monitoring of their interaction, which results in a well-fitted revision of his earlier claim. Henrik's silence mid-sequence also shows that the test takers treat NEST interaction as a matter of talking one at a time and awaiting completion of their peer's topical talk before turn transitions (cf. Sandlund & Sundqvist, 2013).

Discussion

Participants in the rater and learner interactions, respectively, are faced with different challenges: producing versus identifying ratable samples of IC, where learners can only hope that what they collaboratively construct in the test is also salient to raters as IC. For test developers, equally, tasks and rating scales need to be concise enough to generate the kinds of interaction and rater interpretations that the construct aims to capture. Add to the mix the difficulty in discerning IC as a human, social competence, and as L2-relevant IC (cf. Pekarek Doehler, 2018), and the challenge for learners of sticking to instructions while producing L2 discourse that resembles everyday interaction between peers. The present study has touched on some of these challenges; thus, we now return to review our research questions.

As for research question 1, which aimed at uncovering the ways in which CASS raters collaboratively construct IC when reporting on learner turns in their assessments, our analysis shows raters struggling to apply fixed assessment criteria to learner conduct. We also see how raters work jointly to support their initial, intuitive evaluations of the learners' IC with general as well as specific accounts of test conduct. In the rater sequences examined, we see that an overall negative assessment of the learners' IC is also countered by the joint identification of particular moments that cast learners' IC in a more positive light. The RS turns examined in both rater groups pertain to the same sequence in learners' talk that, given it was topicalized in all four groups, was salient to the raters as a sequence where IC was displayed. In Group 1, whose RS of "mom" was not presented here, one participant assesses the "mom" sequence in the following way: "that was really charming," which speaks to the difficulty in scoring IC strictly based on pre-set scoring definitions of what is successful L2 conduct.

Teachers in the workshop draw on the reported turns in supporting or challenging co-constructed views of the learners' IC. Not surprisingly, the talk reported is presented very differently in the rater groups. We saw that in Group 4 (Excerpt 2), Beth enacts a dialogue between Henrik and Fred that has little resemblance to the original interaction as it was only Fred who talked about his mother. The group's interaction reveals that the paired format is central to their assessment, and that the learners do not display enough paired talk to count as displaying IC with reference to the instructions to test takers. However, the RS turn is used as a representative example of temporary displays of IC in which the learners engaged in topic elaboration with each other. In Group 2 (Excerpt 3), Kari makes evident that the RS is not true to the original, but instead reports the gist of Fred's action and how that action symbolized competence in the form of spontaneity and an unprompted topical contribution. As such, the reports are deployed not in the service of reproducing the original, but for the purpose of either jointly identifying a sequence where IC was (or, in other cases, was not) displayed or offering examples of temporary displays of IC for the other participants to inspect and evaluate in making a joint grading decision. The RS thus serves as a focal example of what raters had wished to see across the interaction in order to award a higher grade. The RS turns thus provide an insight into raters' situated cognition regarding the application of assessment factors for the test.

As for research question 2—the interactional work accomplished by learners in the sequence reported on by raters—we concluded that the teacher in this test was an active participant, and thus implicated in the trajectory that develops. It is safe to assume that this teacher wanted to

help the learners by prompting more topical discussion (cf. Sandlund & Sundqvist, 2013). Unsurprisingly, however, the outcome resembles an interview format rather than a paired interaction, which may lead the learners to anticipate topical prompts from the teacher in ongoing talk. The "mom" turn in the learner talk is produced in a slot where the teacher's continuers had indicated that an elaboration was expected, and Fred indeed produces a turn that is not specifically an agreeing/disagreeing response but an independent angle on the topic, which had the double function of orienting to the card statement and offering an account for his agreement. The teacher probes Fred's new angle with a series of questions, to which Fred responds fluently and competently, but the interview format that develops is treated by the rater groups as negative in terms of IC.

As for research question 3, pertaining to the gains of comparing the RS turn with its original production, we see that for the raters the importance of adhering to the paired format overrides the learners' displays of competence in interacting with the teacher, even though both learners display that they can flexibly adapt to the contingencies of the moment and that they can "use the context-free interactional organizations . . . in a context-sensitive manner to participate in social activities" (Kasper & Ross, 2013: 24). The combined analysis of test interaction and rater interpretations of the same underscores the necessity of either preparing learners extensively for what is expected and assessed (i.e., orienting to a co-present third participant should be avoided) or examining and assessing the competence displayed in the interaction as a whole, including when learners orient to the teacher. The second option is appealing from a CA perspective as it aligns with a view of interaction as both universal and context-specific (cf. Pekarek Doehler & Petitjean, 2017).

Our findings are also relevant in light of the function of reports on talk in making claims, where RS "provides a key resource by which speakers can provide evidence for a position, or attest to the factuality of a claim or version of events" (Wooffitt, 2008: 244). Indeed, raters draw on RS to support their judgments of the interaction, but also in the service of identifying and highlighting counter-examples to be considered in assessment, such as in the case of the "mom" turn. Also, from a CA perspective, the RS turns provide insights into the gap between post-interaction reports and their original location. When "importing speech or thought into conversation from a situation removed from the here-and-now" (Couper-Kuhlen, 2007: 119), the original action takes on a new life as a resource for doing something else, such as for justifying or assessing something. In the case of CASS talk of a past interaction, the reported turns serve to substantiate collaborative views-in-progress. By enacting or reporting

on learner turns, raters also made public a contrast (cf. Sandlund, 2014) between the reported turns as IC and the interaction as a whole as non-IC with reference to the paired format. In sum, we believe the analysis of raters' RS turns opened a window into rater cognition on IC as it allowed for an examination of what was salient to them in the original sequence. While the saliency to raters probably pertained to more than just the one reported turn, the turn itself is launched as representing an exception, and also allowed us to examine why in our analysis of the learner sequence.

In terms of rater moderation, we also observe that CASS as an assessment practice has potential in terms of uncovering rater understandings of what IC in general, and scoring rubrics in particular, actually mean aside from intuitive interpretations. Along these lines, we believe the raters in our CASS groups do qualified, in-depth work in checking their own and others' understanding of what the assessment factors mean in relation to learner conduct in a recording that has been inspected together. For example, raters in Excerpt 1 do the work of establishing a shared interpretation of what *taking the initiative* and *carrying the conversation forward* mean in relation to learner conduct, which is what Jönsson and Thornberg (2014) refer to as efforts to enhance teachers' shared interpretations (Sw. *samsyn*) of steering documents, scoring rubrics and their application to learner performance.

Conclusion and Implications

In a study of language proficiency interviews, van Compernolle (2011: 132) argued that while "providing a response where conditionally relevant" (such as in response to a teacher's question) does present evidence of learners' IC *as test takers*, such routinized testing formats may be unproductive for fairly assessing IC. As our study has shown, what raters treat as insufficient L2 IC is linked to the interview format that developed in the paired test examined here. However, in light of our study, we can also partially challenge the critique of the interview format, as the raters nevertheless identified an interview-like sequence as one where one learner did display IC-relevant conduct. In this sequence, Fred's contribution stands out as unprompted, which is indicative of the kinds of conduct the raters perceived as L2 IC. However, the raters paid little or no attention to how Fred competently aligned with the teacher's questions, which tells us that IC, for these raters, is not assessed mainly in terms of how well learners adapt to the contingencies of the moment, nor how they monitor ongoing talk.

When educators or educational authorities administer tests of L2 'speaking,' they rely on test constructs to capture the abilities desirable for future purposes. In the case of high-stakes national tests, the desired

abilities are grounded in the curriculum in question, and as such, they are focused on what learners as future communicators should be able to do in their L2 at a certain stage of development. Our findings on the precedence given to paired talk over relevant responses to the teacher underscore the relevance of further studies on what can be considered to be useful L2 IC in other contexts (cf. Young, 2011) and, consequently, what should be included in constructs.

Adie et al. (2012) observed that most studies of CASS work have been based on teachers' self-reported experiences of the positive effects of moderation and called for research studies that approach CASS from other angles. The present CA study is one such attempt, and one we believe holds promise for further studies on assessment practices *as* social interactions, where different views are made publicly available for confirmation or negotiation (Jönsson & Thornberg, 2014: 392). The findings from the present study also indicate that examination of raters' interaction in assessment can reveal weaknesses in scoring criteria that leave too much room for interpretation, and as such, rater interactions also serve as important sources of information for the development of tests and assessment criteria on IC. Examining the collaborative interpretative work in assessment talk allows insight into one step in a didactic transposition process (Chevallard, 2007; Achiam & Marandino, 2014), where subject knowledge is transformed from academic disciplines, through steering documents and curricula, to subject teaching and learning in the classroom. From this vantage point, teachers' interpretations and applications of formal assessment criteria will undoubtedly impact their teaching and assessment practices. Further, the examination of such displays can yield knowledge of relevance to the calibration of subject constructions in curricula, syllabi and national testing constructs.

Notes

1 Hall (2018) suggests replacing the concept of IC with *interactional repertoires*, whereas others emphasize the need to reconsider the meaning of *competence* to denote observable interactional *practices* and *methods of interaction* (e.g., Waring, 2018).
2 No a priori division was made between direct and indirect reported speech (cf. Clift & Holt, 2007).
3 Translated from Swedish.

References

Achiam, M., & Marandino, M. (2014). A framework for understanding the conditions of science representation and dissemination in museums. *Museum Management and Curatorship*, 29(1), 66–82. doi:10.1080/09647775.2013.869855

Adie, L. E., Klenowski, V., & Wyatt-Smith, C. (2012). Towards an understanding of teacher judgment in the context of social moderation. *Educational Review*, *64*(2), 223–240. doi:http://dx.doi.org/10.1080/00131911.2011.598919

Allal, L. (2013). Teachers' professional judgement in assessment: A cognitive act and a socially situated practice. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 20–34.

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

Butler, Y. G., & Zeng, W. (2014). Young foreign language learners' interactions during task-based paired assessments. *Language Assessment Quarterly*, *11*(1), 45–75. doi:10.1080/15434303.2013.869814

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47.

Chalhoub-Deville, M., & Deville, C. (2005). A look back at and forward to what language testers measure. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 815–832). London: Routledge.

Chevallard, Y. (2007). Readjusting didactics to a changing epistemology. *European Educational Research Journal*, *6*(2), 131–134. doi: https://doi.org/10.2304/eerj.2007.6.2.131

Clift, R., & Holt, E. (2007). Introduction. In R. Clift & E. Holt (Eds.), *Reporting talk* (pp. 1–15). Cambridge: Cambridge University Press.

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

Couper-Kuhlen, E. (2007). Assessing and accounting. In E. Holt & R. Clift (Eds.), *Reporting talk* (pp. 81–119). Cambridge: Cambridge University Press.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, *26*(3), 367–396. doi:10.1177/0265532209104667

Drew, P. (1998). Complaints about transgressions and misconduct. *Research on Language and Social Interaction*, *31*(3–4), 295–325.

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

Edwards, D. (2000). Extreme case formulations: Softeners, investment, and doing nonliteral. *Research on Language and Social Interaction*, *33*(4), 347–373.

Erickson, G. (2009). Nationella prov i engelska: En studie av bedömarsamstämmighet. Retrieved from www.nafs.gu.se/digitalAssets/1319/1319572_bed.np-eng-g.erickson-2009.pdf

Erickson, G., & Börjesson, L. (2001). Bedömning av språkfärdighet i nationella prov och bedömningsmaterial. In R. Ferm & P. Malmberg (Eds.), *Språkboken* (pp. 255–269). Stockholm: Myndigheten för skolutveckling.

Falk, B., & Ort, S. (1998). Sitting down to score: Teacher learning through assessment. *Phi Delta Kappan*, *80*(1), 59–64.

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

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring students. *Language Testing*, *27*(4), 585–602. doi:10.1177/0265532210364049

Goffman, E. (1974). *Frame analysis*. New York: Harper and Row.

Gramkow, K. (2000). Overlap management and interactional competence. *Odense Working paper sin Language and Communication*, *19*(2), 17–31.

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

Hall, J. K. (1995). "Aw, man, where you goin'?": Classroom interaction and the development of L2 interactional competence. *Applied Linguistics*, *6*(2), 37–62.

Hall, J. K. (2018). From L2 interactional competence to L2 interactional repertoires: reconceptualising the objects of L2 learning. *Classroom Discourse*, *9*(1), 25–39.

Halvorsen, K., & Sarangi, S. (2015). Team decision-making in workplace meetings: The interplay of activity roles and discourse roles. *Journal of Pragmatics*, *76*, 1–14. doi:https://doi.org/10.1016/j.pragma.2014.11.002

Hayano, K. (2013). Question design in conversation. In J. Sidnell & T. Stivers (Eds.), *Handbook of conversation analysis* (pp. 395–414). Oxford: Wiley-Blackwell.

Hayashi, M., Raymond, G., & Sidnell, J. (2013). Conversational repair and human understanding: An introduction. In M. Hayashi, G. Raymond, & J. Sidnell (Eds.), *Conversational repair and human understanding* (pp. 1–40). Cambridge: Cambridge University Press.

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

Hepburn, A., & Bolden, G. B. (2013). The conversation analytic approach to transcription. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 57–76). Malden: Blackwell Publishing Ltd.

Hymes, D. (1972). On communicative competence. In J. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293). Harmondsworth: Penguin.

Jönsson, A., & Thornberg, P. (2014). Samsyn eller samstämmighet? En diskussion om sambedömning som redskap för likvärdig bedömning i skolan. *Pedagogisk forskning i Sverige*, *19*(4–5), 386–402.

Kasper, G., & Ross, S. J. (2007). Multiple questions in oral proficiency interviews. *Journal of Pragmatics*, *39*(11), 2045–2070. doi:10.1016/j.pragma.2007.07.011

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.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, *70*(4), 366–372.

Masuda, K. (2011). Acquiring interactional competence in a study abroad context: Japanese language learners' use of the interactional particle "ne." *The Modern Language Journal*, *95*(4), 519–540.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, *26*(3), 397–421. doi:10.1177/0265532209104668

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

McNamara, T. (1996). *Measuring second language performance*. New York: Longman.

Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability*. London: National Assessment Agency.

Millet, I. (2018). The relationship between grading leniency and grading reliability. *Studies in Higher Education*, *43*(9), 1524–1535. doi:10.1080/03075079.2016.1265497

Nakatsuhara, F. (2008). Inter-interviewer variation in oral interview tests. *ELT Journal*, *62*(3), 266–275.

National Assessment Project. (2015). National Assessment Project. Retrieved from http://nafs.gu.se/english/?languageId=100001&disableRedirect=true&returnUrl=http%3A%2F%2Fnafs.gu.se%2F

Nguyen, H. t. (2012). *Developing interactional competence. A conversation-analytic study of patient consultations in pharmacy*. New York: Palgrave Macmillan.

Ochs, E., & Schieffelin, B. (1984). Language acquisition and socialization: Three developmental stories and their implications. In R. Shweder & R. Levine (Eds.), *Culture theory: Essays on mind, self and emotion* (pp. 276–320). New York: Cambridge University Press.

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

Pekarek Doehler, S. (2018). Elaborations on L2 interactional competence: the development of L2 grammar-for-interaction. *Classroom Discourse*, *9*(1), 3–24.

Pekarek Doehler, S., & Petitjean, C. (2017). Introduction: Interactional competences in institutional settings—Young people between school and the workplace. In S. Pekarek Doehler, A. Bangerter, G. de Weck, L. Filliettaz, E. González-Martínez, & C. Petitjean (Eds.), *Interactional competences in institutional settings. From school to the workplace* (pp. 1–26). Cham: Palgrave Macmillan.

Sadler, D. R. (2013). Assuring academic achievement standards: From moderation to calibration. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 5–19. doi:10.1080/0969594X.2012.714742

Sandlund, E. (2014). Prescribing conduct: Enactments of talk or thought in advice-giving sequences. *Discourse Studies*, *16*(5), 645–666.

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

Sandlund, E., & Sundqvist, P. (2017, January). *Equity in the assessment of "interactional competence" Rater inconsistencies and interactional organization*. Paper presented at the ICOP-L2 conference, Neuchâtel, Switzerland.

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

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*(1–2), 361–382.

Schegloff, E. A., Koshik, I., Jacoby, S., & Olsher, D. (2002). Conversation analysis and applied linguistics. *Annual Review of Applied Linguistics*, *22*, 3–31.

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

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

Sundqvist, P., Wikström, P., Sandlund, E., & Nyroos, L. (2018). The teacher as examiner of L2 oral tests: A challenge to standardization. *Language Testing*, *35*(2), 217–238. doi:10.1177/0265532217690782

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.

Swedish Schools Inspectorate. (2013). *Olikheterna är för stora. Omrättning av nationella prov i grundskolan och gymnasieskolan, 2013 [The differences are too great. Re-assessing national tests in compulsory and upper secondary school, 2013]*. Stockholm: Swedish Schools Inspectorate.

Walsh, S. (2012). Conceptualizing classroom interactional competence. *Novitas ROYAL*, *6*(1), 1–14.

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

van Compernolle, R. A. (2011). Responding to questions and L2 learner interactional competence during language proficiency interviews: a microanalytic study with pedagogical implications. In J. K. Hall, J. Hellerman, & S. Pekarek Doehler (Eds.), *L2 interactional competence and development* (pp. 117–144). Bristol: Multilingual Matters.

van Compernolle, R. A. (2013). Interactional competence and the dynamic assessment of L2 pragmatic abilities. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 327–354). Basingstoke: Palgrave Macmillan.

Wang, B. (2010). On rater agreement and rater training. *English Language Teaching*, *3*(1), 108–112.

Waring, H. Z. (2018). Teaching L2 interactional competence: problems and possibilities. *Classroom Discourse*, *9*(1), 57–67.

Wiliam, D. (2007). Content then process: Teacher learning communities in the service of formative assessment. In D. Reeves (Ed.), *Ahead of the curve: The power of assessment to transform teaching and learning* (pp. 182–204). Bloomington: Solution Tree.

Wooffitt, R. (2008). The dead in service of the living. In R. Clift & E. Holt (Eds.), *Reporting talk* (pp. 244–269). Cambridge: Cambridge University Press.

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing. Advance online publication*, *32*(2), 199–225 doi:10.1177/0265532214557113

Young, R. F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 426–443). London: Routledge.

Young, R. F., & He, A. W. (Eds.). (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency*. Amsterdam: John Benjamins.