6 Rating and Reflecting: Displaying Rater Identities in Collegial L2 English Oral Assessment

Erica Sandlund and Pia Sundqvist

1 Introduction

Assessing complex language abilities such as speaking in interaction presents challenges for the development of constructs, scoring rubrics and the practice of assessment. There is a longstanding conviction that training of raters is crucial for the reduction of variance in test scores due to rater factors (e.g. Wilkinson, 1968). Variation in rater severity across rater groups may be the result of many factors, for example, construct interpretations, rater backgrounds and individual biases (Eckes, 2009; Elder *et al*., 2005). McNamara (1996) discusses four dimensions that may play a role in rater variability: rater consistency, rater leniency or severity, rater's use of the rating scale and rater bias or interaction. Holistic rating scales for speaking and interaction present particular challenges for reaching consensus on performances at different levels, as 'a single score may not do justice to speaking' (Fulcher, 2003: 90) and raters are only required to account for an impression of an overall quality rather than for the presence of a certain number of specified features. Consequently, even when the same score is assigned by different raters, there is no way of ascertaining that raters have based that assessment on the same grounds (Jönsson & Thornberg, 2014), or that raters have understood and used a particular rating scale in the same way. As Fulcher (2003) puts it, there is 'little point in building construct models to support the empirical development of rating scales if raters then pay no attention to it' (Fulcher, 2003: 143), and rater training interventions are generally designed to 'socialize raters into a common understanding of the scale descriptors' (Fulcher, 2003: 145). Such socialization of raters, we argue, could also include opportunities to reflect upon individual rater biases in relation to specific learner performances.

Assessment researchers, as well as policymakers and other stakeholders in education, sometimes promote collaborative assessment (i.e. practices of social moderation or consensus moderation, e.g. Linn, 1993; Sadler, 2013) as a remedy for challenges with assessment equity, especially in the context of large-scale standardized testing, as has been the case with the national tests of core subjects in Sweden (Erickson, 2009; Swedish National Agency for Education, 2009; Swedish Schools Inspectorate, 2013). Moderation is defined 'as a practice of engagement in which teaching team members develop a shared understanding of assessment requirements, standards and the evidence that demonstrates differing qualities of performance' (Grainger *et al*., 2016: 551), which makes moderation an organized practice for the verification of assessment judgments against standards (Bloxham *et al*., 2016).

Whether moderation is applied for the sake of achieving validity and reliability in high-stakes grading or as a professional development practice (cf. Jönsson & Thornberg, 2014, on different goals of collaborative assessment, CASS), a closer look at moderation and training as *interactional events* is warranted, as raters' varying perceptions of assessment criteria are reflected in learner scores (cf. Ducasse & Brown, 2009: 425). Raters from different walks of life, carrying different experiences from their own local contexts, may 'attend more or less closely to different sets of criteria, depending on their professional background (…) and a host of other factors' (Eckes, 2009: 43). How raters perceive their own rater characteristics may therefore provide us with insight into one dimension of the professional practice of doing second/foreign language (L2) speaking assessment: the role of rater identities in assigning and accounting for scores in rater training or moderation activities. Like any other collaborative work practice, assessment discussions require participants to reveal their individual views on grading and have their professional judgments challenged by others. As such, the very act of sharing one's professional judgment also means displaying publicly one's professional competence and/or identity. In this chapter, we approach rater variation specifically from the perspective of the raters' displayed perceptions of their rater 'profiles' in collegial assessment activities, that is, when teachers-as-raters jointly and collaboratively assess learner performances, or discuss individually made assessments (cf. Jönsson & Thornberg, 2014). We adopt a qualitative, interactional approach to raters' discussions on L2 speaking in situated assessment talk, and with a conversation analytic (CA) approach we examine how teachers-as-raters, participating in training interventions for the assessment of L2 oral proficiency and interaction, orient to and position themselves as members of particular rater categories.

In line with an interest in rater training which includes *reflections* on professional practice (cf. Mann & Walsh, 2013), we focus specifically on the interactional management of 'rater identities', displayed as raters' orientations to degrees of *severity* and *leniency* when delivering and accounting for assessments of learner productions. Thus, we examine raters'

reflection-in-action, as different rater identity positionings are claimed, mitigated, negotiated and linked to the current assessment tasks. The study is grounded mainly in two research areas: rater perspectives in the assessment of L2 speaking; and, methodologically, professional identity work in talk and interaction (e.g. Antaki & Widdicombe, 1998; Benwell & Stokoe, 2006; Richards, 2006; Stokoe, 2012).

2 Assessing L2 Speaking: The Rater Perspective

Assessment of language skills means 'the act of collecting information and making judgments about a language learner's knowledge of a language and ability to use it' (Chapelle & Brindley, 2002: 268); however, assessing L2 proficiency has sometimes been described as capturing 'a moving target' (Leclercq & Edmonds, 2014: 5), and thus a challenge for assessment. High-stakes, standardized testing procedures are part of systems of accountability in education (e.g. Lundahl, 2016), and educational authorities, schools and individual teachers are responsible for aligning teaching and assessment with set standards. As such, for the assessment of speaking and interacting in an L2 to function as intended, raters, as well as teachers-as-raters, must develop their *assessment literacy* (Popham, 2009, 2011) in making professional judgments about a learner's L2 proficiency and interactional skills in line with standards. Ideally, assessments should not deviate from those of other raters. In this section, we review work on efforts to develop raters' assessment skills with a particular focus on the assessment of L2 speaking. For the sake of clarity, we use the term *rater* consistently to refer to professionals assessing such tests, whether teachers or trained expert raters.

2.1 Rater dialogues, moderation and rater training

Popham (2009, 2011) identifies a need to develop more extensive teacher education modules and in-service training programs in order to build up teachers' assessment literacy. The refinement of assessment skills can be viewed as a shared knowledge base for professional learning communities and also as benchmarking for the sake of assessment validity and reliability. Many studies of training efforts report positive outcomes in terms of higher post-training inter-rater reliability and agreement (see Davis, 2016), with novice raters, often 'excessively severe or lenient' (Davis, 2016: 118), seemingly most affected by training. Other studies have shown that rater variation in terms of severity was not reduced to acceptable levels after training (Lumley & McNamara, 1995). Weigle (1998) explored differences in rater severity and consistency and found support for the idea that rater training is more successful in assisting raters to give more predictable scores (intra-rater reliability) than in assisting them to assign identical scores (inter-rater reliability). Elder *et al.* (2005) report positive outcomes of rater training where raters received *individual*

feedback on their rating performance, while Knoch (2011) saw no effect of feedback on rater performance over time. However, for the sake of stimulating reflection and awareness, individual feedback can work to prompt self-reflective talk (Sundqvist *et al*., 2020).

Variants of what Sadler (2013; see also Linn, 1993) refers to as *consensus moderation* or *social moderation* is another route towards increasing raters' shared understanding of constructs and criteria for assessment. While studies of the effects of moderation on equity and rater agreement are scarce, there appears to be consensus regarding the positive effects of moderation activities as a form of professional development. In a review of literature on collaborative assessment, Jönsson and Thornberg (2014) emphasize the pedagogical potential inherent in having teachers work together on assessing authentic learner performances. Furthermore, moderation activities focusing on building learning communities (Wiliam, 2007) for teachers or raters contribute to developing their 'assessment literacy as well as knowledge of standards' (Bloxham *et al*., 2016: 649), especially when discussions on specific learner performances contain disagreements, which provides opportunities for professional learning and negotiation (cf. also Adie *et al*., 2012). Central to effective moderation is that the social climate allows for 'the representation and exploration of dissensus', which means that the potential embedded in disagreements and challenges is nurtured as an opportunity for learning (Moss & Schultz, 2001: 65) – even though such disagreements may constitute a threat to members' professional identities (cf. Schnurr & Chan, 2011).

Worth noting is that most studies of moderation work have been based on teachers' self-reported experiences of the effects of moderation (Adie *et al*., 2012) and not on examinations of the interaction in collaborative assessment activities, which is the focus of the present chapter. Among the few studies conducted, Jølle (2014) examined transcripts of audio-recorded paired rater dialogues on the assessment of L1 writing. Data were analyzed using two main categories: the *referents* that raters drew upon in judging student texts; and the *responses* to collegial contributions (i.e. 'the way the responses distribute between rejections, yes-buts, follow-up questions and acceptance is seen as an indicator of the quality of the assessment dialogue', Jølle, 2014: 42). Jølle (2014: 37) concludes that the quality of the rater dialogues did not change substantially over time, which left the author with 'the impression that raters often reached consensus without much discussion'. In his study of the usefulness of training in CA in assessing L2 pragmatic competence, Walters (2007) employed so-called *hermeneutic dialogues* post-assessment between two raters. One of the aims of the post-assessment conversations was to resolve rating differences dialogically, since 'even identical or similar scores between raters do not necessarily imply similar judgments' (Walters, 2007: 169). However, the study focused primarily on the aspects of pragmatic competence that the raters initially disagreed on rather than on the

post-assessment dialogues themselves. May (2011b) was interested in features of paired speaking tests that were salient to raters in relation to interactional competence (IC), and had four raters view a video of a paired speaking test while making notes and recording a stimulated verbal report. Subsequently, raters sat in pairs to discuss their ratings, and their discussions were video-taped. The discussions were then coded with a focus on the features of IC that raters attended to, and reported in the form of sample statements illustrating the different categories. Again, while May's (2011b) study certainly evidences the relevance of examining rater discussions, its main aim was not to examine the rater dialogues as institutional interactions between professionals.

Sandlund and Sundqvist (2019) examined moderation meetings for assessing L2 oral proficiency and interaction, adopting video-recordings of rater discussions and a CA approach. The study aimed to uncover how teachers-as-raters conceptualized IC by examining sequences in which raters reported on specific turns or sequences in the paired test they assessed together. The study concludes that enactments and reports of specific learner contributions served to identify evidence of IC-relevant conduct, to support collaborative views-in-progress and to offer counter-examples to negative assessments in immediately prior talk. The authors conclude that examining rater talk as interaction through a CA lens holds promise for understanding how raters apply scoring rubrics and for developing assessment instructions to raters.

In sum, rater discussions in moderation or training efforts constitute a form of *reflective practice*, where raters do assessment but also reflect upon their own rater performance – something that can then be studied empirically (cf. Eckes, 2009: 44). *Reflection*, then, has been defined as an activity 'in which individuals engage to explore their experiences in order to lead to new understandings and appreciations' (Boud *et al*., 1985: 19). Mann and Walsh (2013) have emphasized the need for a shift from written reflective practice to reflection as a dialogic, collaborative and data-led process, as such a take on reflective practice 'is more likely to elucidate the "real world" of professional practice and help work towards better outcomes in professional development' (Mann & Walsh, 2013: 293). In adopting such an approach to the design of rater training and moderation, new insights into the role of rater characteristics and reflection thereupon may be accessed through a focus on participants' situated orientations and actions. It is in this vein that this study targets rater positionings in L2 speaking assessment.

3 Data and Analytic Considerations

3.1 Participants and test data

Participants were teachers of English in Sweden, recruited for participation in two different research and professional development projects run by university researchers. The first was a rater training program for

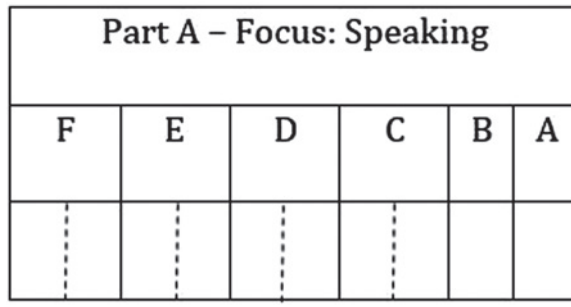


Figure 6.1 Assessment form tick box

assessing the National English Speaking Test (NEST) for Year 6 of compulsory school (NEST 6), and the second was a training workshop for collaborative assessment of the NEST for Year 9 of compulsory school (NEST 9). The NEST is developed by test constructors at the University of Gothenburg on behalf of the Swedish National Agency for Education. As a proficiency test with a traditional design, it includes a productive warm-up task (e.g. picture description or talking about one's family), followed by a peer-peer conversation guided by topic cards. Generally, the test administrator is the students' own English teacher, who thus serves the dual role of administrator and rater (see, for example, Sandlund & Sundqvist, 2019; Sundqvist *et al*., 2018). Topic cards are used for the test conversations, and they carry statements or questions to prompt the learner conversation (e.g. 'There is nothing wrong with junk food'). On average, a test takes around 10 (NEST 6) or 15 (NEST 9) minutes. NEST performance is assessed on a 10-graded scale, from Grade F to Grade A. Since each of the grades F through C is assessed as either 'low' or 'high', the scale is 10-graded (rather than 6-graded), as illustrated in Figure 6.1, which represents the assessment tick box for teachers to mark their test grade.

As such, teachers considering, for example, a C grade for the test, should select one of the two boxes below C to indicate a strong or weak C grade. The test developers also provide information and samples of old national tests and assessment materials for both NEST 6 and NEST 9 on their website (see NAFS Project, 2021a, 2021b; Swedish National Agency for Education, 2014, 2015). Below, we account for the projects in which the data for the present study was collected.

3.1.1 The NEST 6 project

The data collection tied to NEST 6 was collected as part of a research and development project on assessment in two compulsory school subjects in Sweden: Swedish and English. The main objective was to devise and evaluate a training program that could contribute to equity in assessment in English and Swedish, respectively. Here, we target the training track for the assessment of L2 English speaking only.

Participants in the English track were 11 primary school teachers taking part in the training program (all women; mean age: 43; mean years working as teachers: 9.2). They taught English in Grades 4–6 (aged 10–12) at different schools. A background questionnaire revealed that their academic English education varied from nothing to as much as two semesters of English at university level. On average, participants had assessed NEST 6 almost five times, and they had also assigned term grades in English almost five times.

The rater training program was designed to contribute to equity in assessment by developing participating teachers' *awareness* of their own profiles ('identities') as raters of either English oral proficiency and interaction (the English track) or of Swedish writing proficiency (the Swedish track) – that is, complex productive language abilities. All participants had responded to an open call to participate in the combined research and professional development program. Some parts of the training program were jointly conducted with all teachers, but other parts were subject specific.

The training program had three integrated components (for a detailed description of its contents, see Sundqvist *et al*., 2020). The first component was detailed feedback on each participant's individual assessment. The second was theoretical input focusing on language assessment, while the third consisted of repeated moderation sessions in small groups. Altogether, the program offered three full-day meetings on campus. The first day is referred to as the 'pretest day' (June), the second as 'rater training day' (when participants were video-recorded during the actual intervention, August) and the third as the 'posttest day' (September). On these three days, we collected assessment and questionnaire data from the participating teachers and they were also offered various lectures. The lectures were considered particularly relevant to their development as raters of L2 English speaking, and central concepts in the field of assessment (such as *benchmarks*, *construct relevant/irrelevant criteria*, *formative* versus *summative assessment*, *high-stakes* versus *low-stakes testing*, *reliability, standards*, *test construct* and *validity*) were introduced and discussed.

At the pretest day, 10 student performances in five paired NESTs were assessed by the 11 teachers. Each student was scored independently by each rater on the 10-graded scale, yielding 220 assessments of student performances in total. Following the pretest day, each participant was sent an email containing individual feedback. The purpose of this feedback was to raise each teacher's awareness of her own rater profile. Thus, the feedback included information on each participant's assessments, information about assessments in the rater group as a whole, and *benchmarks* (established reference scores for the ten performances – in this case, scores supplied by the Swedish National Agency for Education). Assessment data revealed that the English group assessed fairly close to the benchmarks at the pretest, and this was made explicit in the emails. To be specific, the mean difference for the group from the benchmarks was 0.40; that is, our

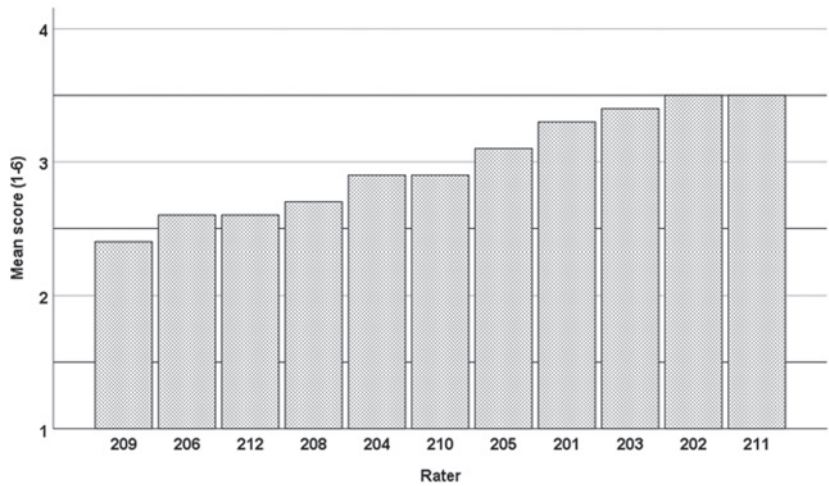


Figure 6.2 Severity continuum: raters' mean scores based on the assessment of 10 student performances on the pretest day, from the strictest rater to the most lenient

11 English raters differed in their assessments on four occasions during the pretest in that they were more strict than the benchmarks (for details, see Sundqvist *et al*., 2020). In order to further prompt rater profile awareness, the feedback included explanations about how each rater had adopted the grading scale compared to the group (see Figure 6.2). There was also an individual table in the email summarizing each rater's own assessments on the 10 student performances. Finally, as preparation for the rater training day, raters were encouraged to reflect on their own profiles as they surfaced in pretest performances, for example, whether they were lenient or strict (or neither) compared to the benchmarks and to the rater group.

At the rater training day, assessment topics were further discussed and explained in lectures by the authors. Particular attention was given to the assessment criteria for the NEST 6 and the test construct *oral production and interaction*, and to being professional as a teacher in terms of aligning with standards. In order to facilitate the raters' discussions, we introduced three bird metaphors, selected to symbolize aspects of rater severity. The first metaphor was *the rater as a hawk* (that is, a severe rater, traditionally rating lower than the benchmark). The second was *the rater as a dove* (that is, a lenient rater, traditionally rating higher than the benchmark), and the third was *the rater as a blackbird* (that is, a 'benchmark' rater, traditionally rating close to or on the benchmark). The metaphors were explained to participants using examples, and the potential consequences for individual learners and their development were discussed.

For the moderation sessions on the rater training day, the participants were in groups of three or four and these were audio- or video-recorded (see Section 2, Table 6.1). In total, they assessed eight student performances

Table 6.1 Overview of data of 10 rater recordings used in the present study

Recording	Data type	Minutes	Raters (*n*)	Tests (*n*)	Students (*n*)
NEST6_2_A	Video	36	4	1	2
NEST6_2_B	Video	31	4	2	4
NEST6_2_C	Video	26	4	1	2
NEST6_3_A	Audio	21	4	1	2
NEST6_3_B	Audio	20	4	2	4
NEST6_3_C	Audio	7	4	1	2
NEST9_1	Video	52	4	1	2
NEST9_2	Video	51	3	1	2
NEST9_3	Video	42	3	1	2
NEST9_4	Video	55	3	1	2
TOTAL		**341**	**37**	**12**	**24**

in four test recordings. They had been given clear oral and written instructions about listening to one recording at a time, starting each test recording with individual listening and independent assessment (for details, see Sundqvist *et al*., 2020). As soon as all members of a group had completed their assessments, group moderation sessions began with the opening of an envelope that contained the official assessment comments and benchmark grades for the learner performances from the National Agency for Education. The teachers had been instructed to then 'out' their own rater profile in the group based on how they had understood their individual feedback. The moderation sessions continued with comparisons of the assessments made with regard to the benchmarks. The various components of the moderation sessions together served the purpose of raising rater awareness. After the posttest, effects of the rater training model were investigated using a many-facet Rasch measurement (MFRM) model (Linacre, 2017; Rasch, 1980), and an intraclass correlation coefficient (ICC) two-way random effects model (McGraw & Wong, 1996), which is reported on in Sundqvist *et al*. (2020).

3.1.2 The NEST 9 project

The data collection tied to NEST 9 was done in 2015 as part of a research project on collaborative assessment (CASS). This dataset is comprised of four video-recordings of L2 English teachers involved in CASS of one paired NEST 9 (see Table 6.1). The teachers had signed up for a professional development day for English teachers organized by a research center at a university, which offered a selection of workshops – one of which was organized by the authors. The workshop was announced as an opportunity to engage in CASS of L2 English oral proficiency and participants consented to filling out a brief background questionnaire and to

video-recording of the CASS discussions for the purpose of research. In total, 13 teachers (12 women; one man) participated. Questionnaire data revealed that all participants had a teacher degree in English. In terms of experience, on average they had worked for 13 years. All knew the NEST 9 well, except for the least experienced participant who was yet to assess her first NEST 'for real'.

The workshop was divided into three parts. In the first part, the authors gave a lecture on research on L2 oral proficiency testing and assessment. In the second part, a selected, authentic paired NEST 9 test recording (with test-takers Fred and Henrik, pseudonyms) was played to the whole group. Raters were instructed to take notes and make initial independent assessments of the two learner performances (cf. May, 2011a). The third part was the actual moderation meetings. Participants were divided into four groups and assigned separate rooms. In each room, the participants had access to the assessment materials for NEST 9 (tasks and assessment instructions), and a web link to the test recording for re-listening on their smartphones/tablets.Their task as raters was to discuss the performances by Fred and Henrik and reach consensus on grades for each test-taker. Afterwards, each of the four groups handed in a joint rater protocol with arguments supporting their grading. The meetings lasted between 42 and 55 minutes (see Table 6.1) and the researchers were not present. As opposed to the procedure in the NEST 6 project, in which the use of benchmarks aimed to achieve calibration against standards, the NEST 9 project centered on description of moderation activities and participants' displayed understandings of the rubrics. As such, participants were only instructed to discuss their judgments and agree on a grade, but were not presented with benchmarks afterwards.

3.2 Recorded data

Based on data collected in the two projects, we have used 10 rater recordings (audio and video) in the present study, amounting to a total of 341 recorded minutes and involving 37 different raters (see Table 6.1). All recordings were transcribed in their entirety using Jeffersonian conventions (Jefferson, 2004). Translations into English are provided in bold face, and interlinear glosses are provided in cases where the translation significantly changed the syntax or word order of the original turn, or when an idiomatic expression without a suitable English equivalent was used. For this study, the datasets were trawled for sequences in which participants displayed orientations to severity/leniency. These sequences were subsequently transcribed in more detail, and Swedish translations were added.

As Table 6.1 shows, the NEST 9 recordings are longer than the NEST 6 recordings, and while the NEST 9 meetings centered on one paired test for assessment, the two NEST 6 groups discussed several different paired tests across three meetings in the same day.

3.3 Methods of analysis

For the analysis of orientations to rater identities as severe or lenient, CA (Sacks *et al*., 1974; Sidnell & Stivers, 2013), combined with some observational tools (or 'keys') from membership categorization analysis (MCA; Stokoe, 2012: 280–281), were used. With the descriptor *rater identity*, we broadly refer to displayed orientations to aspects of rater severity and leniency that participants draw upon in the rater discussions, and where such identity orientations constitute participants' 'displays of, or ascription to, membership of some feature-rich category' (Antaki & Widdicombe, 1998: 2). According to Silverman (1998: 77), Harvey Sacks' take on addressing categories in interaction was to 'try to understand when and how members' descriptions are properly produced'. By producing membership categories (or invoking categories more implicitly through descriptions and reference forms), an interactant can 'strengthen the social action that he or she is performing' (Liu, 2015: 1). Categories, performed through various categorical practices in interaction, can be examined from a sequential, participants' perspective (Stokoe, 2012). The present chapter, while principally adopting a CA approach, also analytically examines participants' descriptions as part of sequentially organized action from the lens of membership categories. The two datasets differ in the sense that NEST 6 participants had been explicitly instructed to talk about their individual rater feedback from the pretest day, whereas for NEST 9 participants issues related to rater severity were volunteered in connection with their assessment talk. As such, we believe the two datasets complement each other in uncovering how rater categories are drawn upon in assessment talk.

4 Analysis

Two main sequential environments in which participants oriented to their severity/leniency identities, namely *rater identities in relation to absent 'others'* (Section 4.1) and *rater identities in assessment negotiations* (Section 4.2) were identified in the two datasets. For the sake of illustration, three sequences are analyzed in detail.

4.1 Positioning rater identities in relation to absent 'others'

In our first analytic presentation, we will examine sequences in which the teachers display orientation to rater severity, and mobilize membership in a particular category of raters-as-professionals by reference to relativeness to non-present others. We begin with an extended excerpt from the first session of the rater training day, as raters here explicitly describe themselves along the severity continuum by using the bird metaphors, and account for the self-identification with explanations for their rater performance at the pretest. In this sequence, accounts following membership descriptions contribute to managing the delicacy of having to reveal a

professionally problematic category. Subsequently, we examine a sequence from a group discussion from the NEST 6 training program in which orientations to severity also occur in connection with talk about non-present others, but where group members position themselves as affiliating with each other and against non-present others. Because of the length of the sequence, the presentation has been divided up into two segments, and analytic comments are presented in conjunction with each. In what follows, Lines 3–46 are presented first.

As per the instructions for the group work, the raters were told to reveal something to their group members about the pretest feedback they had received by email prior to the rater training day, and talk about their reactions to their rater profiles, before proceeding with discussing the new assessments of tests they had just made individually prior to the rater group meeting. As we enter the group's talk, Rater 201 orients to the instructions, and asks the group where they should begin (Line 1, not shown) before suggesting (in a question format, Line 3) that they begin by 'outing' their rater profiles to each other.

Excerpt 6.1a 'One of the doves'; NEST6_2, Lines 3–46

3	201	ska vi: outa oss först. **should we: out ourselves first.**
4	208	**kHHHHHhhhh**
5		(0.8)
6	208	j[ia↓. **Y[eah↓**
7	201	[hur det gick me: [**how it went with**
8	208	[me'ren,] [**with the,**]
9	201	[inte-] inte nu me dagens utan me [not-] not now with today's but with [**Not-**] **not the ones today but with**
10		förra gångens (.) förtest(et) **last time's (.) the pretest**
11	201	[ah HUR Ä: vi som be[dömare [**ah how are we as ra[ters** **ah how we are as raters**

```
12  204  [∘Ja just de∘          [ja just de
         [∘oh yeah right        [oh yeah right=

13  208  =ja:=
         =yes:=

14  201  =ja?
         =yes:=

15       (0.8)

16  202  ja: är ju då: en utav domdäre:h (.) duvor↑na
         I: am PRT then one of those         doves
         So I'm one of the doves then

17  201  m:?

         (1.4)

18  202  .pt e:h lite mer givmild och jag tr↑o:r
         .pt e:h a little more generous and I ↑think

19       att de be↑ror på att när vi har gjort
         that it depends on that when we have done

         + gaze up, circling finger movements
20       +sambedömningar¿ (0.4) så ha ja förstått att
         +co-assessments¿ (0.4) I have learned that

                    +202 both palms pressing down twice
                                + 201 nods twice
21       att många +(.)dömer+ (.) eller bedömer såe:h .pt
         that many +(.) judge (.) or assess so e:h .pt

22  201  hårt.=
         Harshly.=

23  202  =hårt¿
         =harshly¿

24  202  så då har jag varit den som liksom har
         so then have I been the one who kind of has
         then I have been the one to kind of

                       +hands raised
25       ställt frågan +men*e::h* (0.5) kan vi inte
         asked the question but         can we not
         ask bute::h (0.5) can't we see

26       se:r vi inte de hä:r (.) as[så
         see we not this I mean
         can't we see this (.) I    [mean

                                        +nods
27  201                               [a:? +
```

```
28  202  för å lätta upp de
         PREP lighten up it
         to lighten things up

29  201  m:?

30  202  å det fortsätter ja me när jag bedömer
         and that continue I with when I assess
         an' I continue with that when I assess

31       själv också¿
         self too
         on my own too¿

32  201  ja?
         Yes?

33  202  å de:- de visar sej precis i de här också .hh
         an' it it shows-RFL precisely in these too
         and it- it is evident in these too.hh

34  201  ja?
         Yes?

35  202  ja har (0.2) bedömt tjejen <li:te>
         I have       assessed girl-the little
         I have (0.2) assessed the girl a little

36       för högt
         too high

37  201  m:

38  202  .h ock- men killen har jag på rätt nivå
         .h and- but guy-the have I on right level.
         .h an- but the guy I have at the right level.

39  201  ↑ja.
         ↑Yes.

40  202  m: (0.3) men så att ja: måste
         m: (0.3) but so that I have to

41       tänka mig för att
         phrasal verb  to
         be careful to

42       inte va riktigt så generös.=
         not be quite so generous.=

43  201  =m:=

44  202  =som ja (1.1) har varit
         =as I (1.1) have been

45  201  m:.

46  202  m:.
```

Rater 201 continues in Lines 7–10 by specifying that this is in reference to the pretest performances rather than to the assessment work they have just conducted, and uses the description 'how we are as raters' (Line 11) to further describe the proposed activity. Rater 204 displays recognition of the suggested activity, and so does Rater 208, with a minimal agreement response. Rater 202, then, volunteers a self-categorization in Line 16. In her turn, Rater 202 categorizes herself as 'one of the doves' – using the bird

metaphor used to symbolize rater leniency. By doing so, she explicitly mobilizes the rater severity continuum, with the severe hawks at one end and the lenient doves at the other, and reveals that in her pretest assessments her performance had clearly placed her on the lenient side in comparison with the others. Note, however, that while the researchers had offered the categories, the individual feedback did not contain classifications of the participants as such, but only showed scores of individual ratings compared to benchmarks, and the severity distribution within the rater group (see Figure 6.2). As such, participants may opt to recruit these categories based on their own interpretations of their individual scores, which are unknown to the group. Rater 202 thus volunteers this categorization of herself.

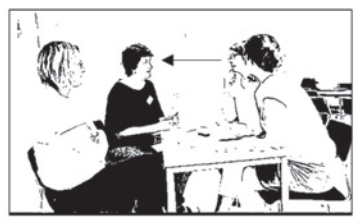
Co-participants (except for the minimal 'm:?' from 201 in Line 17) do not comment on or assess this revelation, but appear to await further elaboration from Rater 202. In Line 18, she elaborates on her revelation, specifying a category-bound predicate (Stokoe, 2012: 281) of a dove rater as 'a little more generous'. She immediately embarks on an account where she reflects upon possible reasons behind her leniency, which indicates that self-categorization is an accountable action. This account (Lines 19–31) centers on her rating experiences outside of the training, where other teachers she has co-assessed with have judged the learners 'harshly' (Line 202, description first provided by Rater 201 in response to Rater 202's possible word search in Line 21), and where she positions herself as someone who has attempted to counter 'harsh' assessments by highlighting positive aspects of learner performances. The account offers an explanation for being lenient that casts Rater 202's approach in a more positive light – in contrast to 'many' of her colleagues, she is the one to highlight strengths by asking her colleagues to 'see this' in order to 'lighten things up'. From just this revelation of a rater profile, we can see that being placed on the far end of the lenient side of the continuum is treated as problematic and accountable. By sharing past experiences, where severe raters are described as rating 'harshly' (rather than described as being 'to the point'), Rater 202 depicts a scenario where her leniency accomplishes an important balance, and thus casts her own 'dove status' as a result of her also paying attention to the strengths in learners' performances. By invoking her leniency as a result of her experience of striving for a more holistic approach, she also invokes severity as paired with actions of excessive strictness in judging learners. She sums up the connection between her pretest performance and her past experience in Lines 30–31, where she states that she tends to continue with the same approach when she grades tests on her own. As co-participants only display receipt of her account, Rater 202 continues in Lines 33 and onwards by specifically referring to the feedback sheet in front of her, stating that her approach to rating is also visible in a particular test from the pretest where she had assessed 'the girl a little too high' (Lines 35–36) but the boy 'at the right level' (Line 38, below). However, the slow production and elongated vowel on <li:te> (a little, Line 35) emphasizes the qualifying adverbial 'a little', which works

to downplay the severity of her rater error in relation to the benchmark. Thus, in the evidence she supplies for her own analysis of her rater profile, her turn indicates that it is designed for a specific hearing: that even though she was marginally over-lenient for one learner, she was on the benchmark for the other. Thus, her turn serves to pre-empt recipient understandings of her as *always* being overly lenient or 'wrong' in her professional assessment work. In Lines 40–44, Rater 202 formulates what she needs to think about in her future assessment work as a result of the feedback: she has to be 'careful to not be quite so generous' as she has been in the past, to which Rater 201 provides an acknowledging 'm:'. As such, the self-categorization, followed by an analytic account, ends with a reflective and forward-oriented formulation of desirable future conduct.

Moving forward to the second part of this sequence, presented in Excerpt 6.1b below, Rater 208 follows the self-revelation path set by Rater 202, using the categorization device *hawk* (Line 48) to reveal that she was, in fact, on the other end of the continuum at the pretest:

Excerpt 6.1b 'I am a hawk'; NEST6_2, Lines 48–74

		+gaze up to 202
48	208	å +ja är en hök.
		and I am a hawk.



49	202	**m:**
50	201	**m::,**
		+gaze down to documents
51	208	+>ja.< .pt .hhh
		+>yeah<. .pt .hhh

52		(1.4)
53	208	ocke:h (0.8) a: när ja skulle försöka
		ande:h (0.8) a: when I was trying to
54		fundera på varför ja va: en hök för ja
		think about why I was a hawk cuz I

55 (0.3) upplever mig inte va en hök (.)
(0.3) don't see myself as a hawk (.)

56 annars¿
otherwise¿

57 vid bedömning men men hä:r var jag ju de
in assessment but but he:re I was

58 helt klart. |
no doubt.

59 (0.8)

60 208 **e::h**

61 (1.7)

62 208 och jag tro:r att ja va: (.) att
and I think that I was (.) that

+gaze up
63 det ä: lite grann det här jag är +ju själv
it's a little this that I'm +*ju* on my own

64 på min skola (.) ja är alldeles själv när ja
at my school (.) I'm all on my own when I

65 bedömer+ och att de finns en sån
assess + and that there is such an

+ 201 nods
66 <o+säkerhet>
<in+security>

67 å när ja då är osäker (.) så sätter ja mej på
an' if I'm unsure (.) I'll place myself on

68 nån slags (1.2)
a kind of (1.2)

69 .hha de e lite riskfritt att
.hha it's sort of risk-free to

70 va ↑hök då
be a ↑hawk then

71 201 [ja: just d e t]
[Ye:ah that's right]

72 208 [än å va:- än å va
[than to be:- than to be a

73 208 du:[↑va på nåt sätt
a ↑do[ve in some ways

74 201 [m:

As Rater 208 begins speaking, she shifts her gaze to Rater 202, as if responding specifically to her as the previous speaker. At Line 51, after having produced the description, she shifts her gaze down to the documents on the table and produces a 'yeah.' with falling intonation. In combination with her facial expression, this is a confirming response to the two acknowledgment tokens produced by Raters 201 and 202, but also indicates that there is something problematic about having performed as a hawk. The confirming 'yeah' in combination with the gaze shift, lip smack and inbreath seems also to appeal to a shared sentiment about performing at the extremes of the severity–leniency continuum. As with the prior dove categorization, co-participants await further elaboration, and Rater 208 reports on her own reflection process at the time of receiving the feedback: she was trying to think about why she was a hawk here, because her results did apparently not match her own perception. Her surprise at the feedback is expressed as, 'I don't see myself as a hawk otherwise in assessment', but also acknowledges that her results tell a different story (Lines 55–58).

Having pre-announced an upcoming reflection about severity on the pretest, she prefaces her candidate explanation with 'I think' (Line 62) which, just as Rater 202 did previously, relates her performance to past experiences at her local school, where she happens to be the only English teacher (Lines 63–64). She continues her account by formulating what these conditions generate: that there is such an 'insecurity' as a result of having to make all decisions on her own. She then returns to the severity/leniency metaphors, and proposes that it is 'sort of risk-free' to exercise severity rather than leniency (Lines 69–70, 73). This yields an affiliative response from Rater 201 before Rater 208 explicitly offers the contrast to a dove (Line 73). In Rater 201's description, then, it is safer, professionally speaking, to exercise severity than risk facing accusations of contribution to grade inflation by awarding high scores.

In this sequence, identification at either end of the continuum is treated as being a problematic rater category, warranting accounts of prior experiences and contexts outside of the current interaction. At the same time, the benchmarks are oriented to as normative, with the implication that displaying category membership in a group that is close to benchmarks is non-accountable. As such, a scale of categories is occasioned – both institutionally and interactionally – where acknowledged membership at either end of the continuum is morally accountable in relation to treatment of benchmark grades as the norm. Scales often operate 'together with notions of normality and markedness' (Bilmes, 2019: 82). Similarly, the understanding of how particular descriptions fit into a given scale requires cultural knowledge as well as 'attention to what scales are relevant and how a particular scale is constructed within the local interaction' (Hauser & Prior, 2019: 76). Raters in our data orient to shared information (figures and images in the feedback documents) as well as to culturally shared norms about professionalism. Their orientations to the benchmarks as normative are evident, for example, in descriptions of rater performance in *relative* terms ('too high', Line

35; 'on the right level', Line 38) and the self-reported conclusions about future conduct, and in the very production of the categorical description (gaze shifts, gaps, intonation) that projects a problem associated with such membership. Also, the categories of hawk and dove, while provided by the workshop organizers, are also locally occasioned in the sequential context of each speaker's analysis of the feedback received, where participants themselves link the contents of the feedback to a rater category metaphor. The two consecutive self-categorizations in the first and second parts of Excerpt 6.1 place the two raters on each end of the scale. While both participants deal with their 'problematic' rater identities in similar ways using accounts, the second account (i.e. the hawk) also *modifies* the scale so that a view of severity as slightly more preferable than leniency (while still problematic in relation to the norm) emerges. In both self-categorizations, individual performance is accounted for in terms of the conduct of others (or lack of others), and these external circumstances are assigned part of the blame for a particular rater's performance.

In the NEST 9 dataset, we also observed orientations to severity, which surfaced in talk about the assessment criteria or in relation to colleagues at their schools, as exemplified in Excerpt 6.2. Here, the discussion has centered on the benchmark tests provided by the test constructors (here referred to as *Skolverket*, i.e. The Swedish National Agency for Education), and in Lines 1–3 Ann announces that she 'often' does not agree with the benchmark grades set as reference points. She presents her claim rather neutrally, thus not revealing whether there is any systematicity in the difference between her views and those of the test constructors, but continues to reveal that this perceived discrepancy is because she feels the benchmark grades are too lenient. Her turn is left incomplete in Line 6, but her 'I think they pass way too-' clearly displays an orientation to the test constructors' set grades as too lenient:

Excerpt 6.2 'went down a notch'; NEST 9_2, Lines 3–18, 19–31

```
1    ANN  men där e:: jag tycker de e ofta de e::m
          but there e:: I think it is often it e::m
          but I often find that

2          (1.1)

           +shaking head repeatedly
3          +skolverket å ja tycker inte samma sak
           +skolverket and I don't think the same

4          ofta asså
           often really

5    KAR   .hhhnä:.hh
           .hhhno: .hh

6    ANN   ◦ja tycker dom godkä[nner alldeles-,  ◦]
           ◦I think they pass way too-, ◦
```

7 KAR [ja och sen bara m-]
[yes and also just b-

8 mellan kolleger ↑mä
between colleagues also

9 LIS m::?

10 ANN ja:?
ye:s?

11 KAR e:h den ja jobbar närmast med nu: vi är
e:h the one I work closest with now we are

12 väldit överrens:
very much in agreement

13 (.)

14 KAR ∘men sen:: en annan kollega∘
∘but then another colleague∘

15 >hon har gått i pension nu< ↑hon (.) .hhh
>she has retired now< ↑she (.) .hhh

16 (.)

17 KAR hade nog lite >vi tog< över lite:
probably had a little >we took< over some

18 klasser efter henne å d-
classes from her and d-

19 (.)

20 ∘var nåra∘ s(hh)om(hh) åkte ne:r
∘were a few∘ who went down

21 ett hack.
a notch

22 ANN m::

23 KAR [ja↓]
[yes↓]

24 LIS [du:] ä på högstadiet¿
[you] are at secondary school

25 KAR +ja::
+yes
+nods twice, gaze at LIS

26 KAR .hh ocke:h (.) ↑a:?
.hh andu:h (.) well?

27 (.)

28 KAR ä: det hon eller vi som har (.) .hh [rätt
is it her or us who is (.) right

29 ANN [JA DE Ä
[YES IT IS

30 JU ↑svårt
PRT
***JU* really difficult**

Ann thus positions herself as a more severe rater, and also provides an (albeit incomplete) assessment of the benchmarks as 'too' lenient, and making it possible to pass 'way' more students than she would. Consequently, leniency in relation to the lowest passing grade (E) is depicted in a negative light, and a higher level of severity, then, is recruited for displaying professionalism. While Ann's turn challenges the epistemic primacy of the norms set by the educational authority, she nevertheless treats it as a norm, albeit a problematic one. In overlap, Kari offers an agreeing 'yes', but instead of exploring the issue of the benchmark grades, she brings forth a parallel context in which differences in severity can arise – between colleagues at one's local school (Line 8). She exemplifies this issue further with an account of how she and her current colleague are 'very much in agreement' (Line 12) but that when a former colleague retired, they took over some of her classes, at which point they had 'a few who went down a notch' (Line 20). As with Ann's example of the benchmark grades, Kari's account is based on a narrative about a third party who, apparently, graded more leniently, resulting in some students' grades being lowered one step when new teachers came in. Kari delivers her account factually, but later acknowledges some uncertainty as to whether the former colleague, or Kari and her current colleague, were 'right' (Line 28). While Ann's contributions positions her own severity as somewhat superior to the benchmark grades from the test constructors, Kari's account, while revealing that she is obviously more severe than a former colleague, mainly functions to assert the presence of discrepancies in assessment between different raters. Ann then provides an agreeing assessment, that it is 'really difficult' (presumably to know which assessment is 'right').

Across both datasets, issues of rater severity is frequently brought up in the context of acknowledging rater differences and preferences. We now turn to the second context in which orientations to rater profiles frequently surface, namely in sequences where assessment decisions are to be made, and disagreement about a particular grade has been revealed.

4.2 Rater identity displays in making collaborative assessment decisions

Unsurprisingly, raters frequently orient to their own perceived position along the severity–leniency continuum when a discrepancy between individually assigned assessments have become evident. In Excerpt 6.3 from the NEST 9 dataset, group members Katherine, Victoria and Alison are discussing a grade for one of the two boys, having revealed their individual grades earlier in the rater meeting and now returning to them in order to agree on a joint grade. In Line 1, Victoria delivers the scope of her preferred grade – a C or a D (Line 3) – and as there is no response apart from the minimal acknowledgment from Katherine, she asserts, using the extreme case formulation *never* (Edwards, 2000), that this is as high as she could go (Line 6). In her rather adamant claim, combined with the formulation 'I think', she is invoking a degree of severity as a property of her assessment decision with regard to the learner as it makes clear that a grade above C would be out of the question for her:

Excerpt 6.3 'maybe it's me who's too strict'; NEST 9_3, Lines 1–37

1	VIC	men jag skulle- (0.2) om ja: s: (0.3) **but I should- if I s-**
2		↑HAN skulle ja no va lite mer att jag tänker **↑HIM I would probably be more that I think**
3		ce: elle de: **cee:: or dee::**
4	KAT	°m:¿°
5		(0.3)
6	VIC	<aldrig högre> än- än dä: **<never higher> than- than that**
7	KAT	ne:↑j ehmen ja sa dä ja är benägen å **no:↑y uh but I said that I'm inclined to**
8		sätta ce: pl[us **put a ce:: pl[us**
9	VIC	[ja:¿ **[ye:s¿**
10	KAT	istället för be: **instead of be::**
11		(1.0)
12	KAT	[(ja tycker)] **[(I think)]**

13 VIC [men ↑DU tycker] ↑↑BE:.
[but ↑YOU think] ↑↑BE:.

14 (0.7)

15 ALI jamen grejen är att ja::g n- nhhh (0.3) ja
wellbut the thing is that I:: n- nhhh (0.3) I

16 har <aldrig rättat> nått sånt här .h för↑ut
have <never graded> anything like this .h be↑fore

17 å ja tänker (0.6) j- n(hh)(0.8)ja (.) liksom
an' I'm thinking (0.6) I- n(hh)(0.8)I (.) kinda

+palm held flat in the air, lowering movement
18 +sch- >sänker min: >mina krav (0.7) eftersom
+ sch- >lower my: >my standards (0.7) since

19 det inte äre:hm (0.4) tvåspråk[iga
it isn't hm (0.4) bilingu[al

20 VIC [a:¿ a:¿=
[ye:s¿ ye:s¿

21 KAT **=m:.**

22 ALI så ja tänker ja måste sänka det ganska
so I'm thinking I have to lower it quite

23 mycke'[rå
a lot [then

24 KAT **[m::.**

25 ALI men hh [kanske ja sänker det för mycke
but hh [maybe I'm lowering it too much

26 KAT [m:.

27 (2.3)

28 KAT **m:**

29 (2.1)

30 KAT mene:h,=
Bute:h,=

31 VIC =kanske är ja som är för sträng
=maybe it's me who's too strict

32 [också?]
[too?]

33 KAT [men är du]
[but are you]

34 VIC så kan de ju va.
so can it *ju* be.
that could be the case.

35 KAT °annars betraktas ja: som st(hh)rä(hh)ng hh°
°usually I:'m viewed as st(hh)ri(hh)ct hh°

36 brukar ja ju göra=
normally I am=

37 VIC =ja[HA::]
=a:[HA::]

In response, Katherine produces an initial 'no' token, which functions as an initial agreement with Victoria's claim (Pomerantz, 1984), but then announces her own grading preference in the shape of a dispreferred disagreeing action ('uh but', Line 7). The formulation 'inclined to put a ce:: plus instead of be::' positions Katherine's preferred grade at a higher level than Victoria's D or C. Here, a grading discrepancy has become publicly available, where Victoria's grade indicates greater severity. Katherine's 'I think' (Line 12) is overlapped by Victoria, who turns to the third rater, Alison. Victoria formulates Alison's stance on the grade for confirmation, emphasizing 'YOU' and the grade 'BE:' with a 'surprised' intonation (Line 13, cf. Wilkinson & Kitzinger, 2006). By displaying surprise although the B grade had been previously revealed and thus is no actual news to Victoria, she also displays some doubt or disbelief at Alison's professional opinion, which is more lenient than Victoria's and even Katherine's. Alison's account in response shows that Victoria's turn projected that she is accountable for explaining her grade, and her account centers on her *inexperience* with rating the NEST: she teaches English as a mother tongue and is therefore used to bilingual learners rather than foreign language learners, and her suggested B grade was the result of her lowering her standards to fit with non-bilinguals (Lines 15–19, 22, 25). Consequently, she is projecting a connection between her lack of experience with a perceived leniency in the graded test. However, while mobilizing a temporary identity as lenient, she is also invoking a higher standard in her everyday professional practice. In Line 25, Alison opens up for deviant views by acknowledging that she is perhaps 'lowering it too much'.

In response, Katherine initiates a disagreeing turn (Line 30), but stops as Victoria produces a self-reflective categorization: 'maybe it's me who's too strict too', which in a way mirrors Alison's indication that she may have been too lenient. In acknowledging that she may just as well be the reason for the discrepancy, she mobilizes the severity–leniency characteristics in affiliating with Alison's displayed uncertainty. Up to this point, then, Victoria has positioned her (more severe) grading view in relation to Katherine and Alison rather strongly, but after Alison's account, she mitigates her earlier claims and treats misplaced severity as equally

problematic. In self-identifying as a severe rater who may be too severe in this particular case, public self-reflection is initiated. Interestingly, this occasions another self-categorization from Katherine in Line 35: 'usually I:'m viewed as st(hh)ri(hh)ct hh'. With the emphasis on 'I'm' and the use of 'usually', which recruits non-present others as perception evidence, her turn challenges Victoria's self-categorization as severe by claiming a severe identity for herself. Her production of 'strict' contains laugh particles, and is followed by another reference to her rater profile outside the current context: 'normally I am' (Line 36). Victoria treats this as surprising news in Line 37, and the discussion continues with additional arguments about the particular student they are jointly grading (not included).

The sequence reveals that when diverging perceptions of the learner's performance have been made publicly available, a space for explaining the divergence opens up. This is done through public self-reflection on reasons underlying each rater's view, which is partially accomplished through self-categorization and reflection on the accuracy of these approaches. As Logren *et al*. (2017) have noted, self-reflection in interaction can be identified in 'utterances in which speakers report their own behaviour and experiences and *mark them as a target of reflection*' (Logren *et al*., 2017: 426, italics in the original), which in this case relates to their grading. While Alison is cast as lenient and accounts for her inexperience as an explanation, Victoria reflects on her possible excessive severity, and Katherine, consequently, claims membership in the severity group of raters by drawing on her experiences in other contexts. Katherine's positioning thus rejects Victoria's indication that it is her general severity that underlies the current discrepancy, since Katherine herself has been viewed as severe in all other contexts. Alison's account, which is accepted, and the 'competition' for membership in the 'severe rater' group also indicate a view of leniency as more problematic than severity, as Katherine displays unwillingness to be identified as lenient, even though she initially proposed a higher grade than Victoria.

5 Discussion

In our analytic section, we have examined three selected sequences in which teachers-as-raters orient to rater severity or leniency in two distinct sequential contexts. Across these three and others in our datasets, participants display an orientation to leniency as a slightly more problematic professional rater identity than severity. Whether using rater metaphors provided, or orienting to severity/leniency in the context of diverging views on particular learner performances, rater leniency is accounted for in relation to inexperience or attributed more positive predicates such as 'generous' in accounts of the excessive severity of others. Severity is linked to excessive strictness, but also to rater insecurity, where severity is accounted for as a safer option, which in turn implies that leniency faces the risk of

accusations of unprofessionalism. A scale with a continuum from severe to lenient is not only occasioned from the institution of assessment (in this case, through individual feedback on pretest rating performance, which in itself placed each rater along this continuum), but is also occasioned and made relevant in the situated rater interactions. In their talk, only category membership far away from the benchmark grades is treated as accountable. However, this is also evident when participants question the accuracy of the benchmarks. By critiquing the benchmarks as overly lenient, participants show orientation to them as the norm, but also tilt the moral implications of the scale in favor of the severity category. Consequently, even though both extreme positions are treated as problematic, the scale occasioned in the raters' treatment of the benchmark as the norm allows for professionalism to be displayed through critique of lenient benchmark grades. In all, rater self-categorizations strengthen preferred identity positionings as professionals and/or invite further justifications for rater performance on the extremes of the continuum.

In the NEST 6 project from which data for the present study were drawn, participants returned a month after rater training for a posttest. The posttest analysis revealed that the group scored even closer to the benchmarks than at the pretest, and made greater use of the full range of grades available after participating in training, revealing that changes in assessment practice from pretest to posttest did take place (Sundqvist *et al.*, 2020). For the NEST 9 project, which mainly centered on collaborative assessment rather than rater training, no scoring data were collected at a later occasion. While the present study has focused specifically on interactional trajectories during two types of L2 assessment training events, it is possible that the category memberships formulated by our participants, and the subsequent treatment of them, constitute a core aspect of the development of rater awareness, which in turn contributed to the posttest change. It remains for further research to examine more carefully how such identity positionings and participants' stance towards them may gradually change and even (temporarily) stabilize through participation in rater training activities over time and, in turn, how self-categorizations may form pivotal moments in calibrating assessment practices.

6 Conclusion

Rater variability is naturally a problem in high-stakes language assessment and, as McNamara (1996) notes, rater bias and variations in severity are two of the factors underlying problematic variability. However, these issues have mainly been explored in quantitative studies of rater performance rather than as socially and interactionally constructed and negotiated identities in accomplishing professional activities. Likewise, research on rater training has principally centered on either self-reported

experiences or measurable effects and less on the reflective practices involved, such as how raters formulate, negotiate and mobilize their own rater identities in assessment talk. This chapter has targeted how rater identity positionings in situated talk between professionals, frequently adopted through self-categorizations and accounts, enforce, justify or mitigate past assessment performances. Through a CA lens, we have demonstrated some ways in which raters' reflection-in-action can be accessed in descriptions and accounts, partly accomplished through categorization practices (Evans & Fitzgerald, 2016; Hauser, 2011; Sacks, 1992), which can be examined sequentially (Stokoe, 2012). The two sequential contexts examined in moderation interactions between teachers-as-raters – in relation to non-present others and in disagreements about grades – revealed how identity positionings contribute to the establishment of lay/expert roles, and to the shared construction of severity as 'more professional' than leniency. As such, rater positionings taken in interaction have moral implications. This observation is central, as a more positive view on severity may reveal an assessment bias that could hinder equity in high-stakes assessment. We argue that sequential analysis of rater identities in interaction can offer a window into teachers' stepwise modification of rater cognition, and thus holds promise for further studies on assessment (cf. Jönsson & Thornberg, 2014).

References

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

Antaki, C. and Widdicombe, S. (1998) Identity as an achievement and as a tool. In C. Antaki and S. Widdicombe (eds) *Identities in Talk* (pp. 1–14). London: Sage.

Benwell, B. and Stokoe, E. (2006) *Discourse and Identity*. Edinburgh: Edinburgh University Press.

Bilmes, J. (2019) Regrading as a conversational practice. *Journal of Pragmatics* 150, 80–91. See http://www.sciencedirect.com/science/article/pii/S0378216617308007. doi:10.1016/j.pragma.2018.08.020

Bloxham, S., Hughes, C. and Adie, L. (2016) What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices. *Assessment & Evaluation in Higher Education* 41 (4), 638–653. doi:10.1080/02602938.2015.1039932

Boud, D., Keogh, R. and Walker, D. (1985) Promoting reflection in learning: A model. In D. Boud, R. Keogh and D. Walker (eds) *Reflection: Turning Experience into Learning* (pp. 18–40). London: Routledge Falmer.

Chapelle, C.A. and Brindley, G. (2002) Assessment. In N. Schmitt (ed.) *An Introduction to Applied Linguistics* (pp. 267–288). New York: Oxford University Press.

Davis, L. (2016) The influence of training and experience on rater performance in scoring spoken language. *Language Testing* 33 (1), 117–135. doi:10.1177/0265532215582282

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

Eckes, T. (2009) On common ground? How raters perceive scoring criteria in oral proficiency testing. In A. Brown and K. Hill (eds) *Tasks and Criteria in Performance Assessment* (pp. 43–73). Frankfurt am Main: Peter Lang.

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

Elder, C., Knoch, U., Barkhuizen, G. and von Randow, J. (2005) Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly* 2 (3), 175–196.

Erickson, G. (2009) Nationella prov i engelska – en studie av bedömarsamstämmighet. See https://www.gu.se/nationella-prov-frammande-sprak/rapporter-och-skrifter#Studie-av-bed%C3%B6marsamst%C3%A4mmigheti-engelska-%C3%A5k-9

Evans, B. and Fitzgerald, R. (2016) 'It's training man'! Membership categorization and the institutional moral order of basketball training. *Australian Journal of Linguistics* 36 (2), 205–233. doi:10.1080/07268602.2015.1121531

Fulcher, G. (2003) *Testing Second Language Speaking*. Harlow: Pearson Education.

Grainger, P., Adie, L. and Weir, K. (2016) Quality assurance of assessment and moderation discourses involving sessional staff. *Assessment & Evaluation in Higher Education* 41 (4), 548–559.

Hauser, E. (2011) Generalization: A practice of situated categorization in talk. *Human Studies* 34 (2), 183–198. doi:10.1007/s10746-011-9184-y

Hauser, E. and Prior, M.T. (2019) Editorial. Introduction to topicalizing regrading in interaction. *Journal of Pragmatics* 150, 75–79. See http://www.sciencedirect.com/science/article/pii/S0378216619304928. doi:10.1016/j.pragma.2019.07.001

Jefferson, G. (2004) Glossary of transcript symbols with an introduction. In G.H. Lerner (ed.) *Conversation Analysis: Studies from the First Generation* (pp. 13–31). Amsterdam: John Benjamins.

Jølle, L.J. (2014) Pair assessment of pupil writing: A dialogic approach for studying the development of rater competence. *Assessing Writing* 20, 37–52. See http://www.sciencedirect.com/science/article/pii/S1075293514000038. doi:10.1016/j.asw.2014.01.002

Jönsson, A. and 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.

Knoch, U. (2011) Investigating the effectiveness of individualized feedback to rating behavior – a longitudinal study. *Language Testing* 28 (2), 179–200. See http://ltj.sagepub.com/cgi/content/abstract/28/2/179. doi:10.1177/0265532210384252

Leclercq, P. and Edmonds, A. (2014) How to assess L2 proficiency? An overview of proficiency assessment research. In P. Leclercq, A. Edmonds and H. Hilton (eds) *Measuring L2 Proficiency: Perspectives from SLA* (pp. 3–23). Bristol: Multilingual Matters.

Linacre, J.M. (2017) Facets® (Version 3.80.0) [computer software]. Beaverton, OR: Winsteps.com.

Linn, R.L. (1993) Linking results of distinct assessments. *Applied Measurement in Education* 6 (1), 83–102.

Liu, R.Y. (2015) Invoking membership categories through marked person reference forms in parent-child interaction. *Working Papers in TESOL & Applied Linguistics* 15 (1), 1–13.

Logren, A., Ruusuvuori, J. and Laitinen, J. (2017) Self-reflective talk in group counselling. *Discourse Studies* 19 (4), 422–440. doi:10.1177/1461445617706771

Lumley, T. and McNamara, T. (1995) Rater characteristics and rater bias: Implications for training. *Language Testing* 12 (1), 54–71. doi:10.1177/026553229501200104

Lundahl, C. (2016) Nationella prov – ett redskap med tvetydiga syften [National tests – a tool with ambiguous aims]. In C. Lundahl and M. Folke-Fichtelius (eds) *Bedömning i och av skolan – praktik, principer, politik* (pp. 243–261). Lund: Studentlitteratur.

Mann, S. and Walsh, S. (2013) RP or 'RIP': A critical perspective on reflective practice. *Applied Linguistics Review* 4 (2), 291–315. doi:10.1515/applirev-2013-0013
May, L. (2011a) *Interaction in a Paired Speaking Test*. Frankfurt am Main: Peter Lang.
May, L. (2011b) Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly* 8 (2), 127–145. doi:10.1080/154303.2011.565845
McGraw, K.O. and Wong, S.P. (1996) Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1 (1), 30–46.
McNamara, T. (1996) *Measuring Second Language Performance*. New York: Longman.
Moss, P.A. and Schultz, A. (2001) Educational standards, assessment, and the search for consensus. *American Educational Research Journal* 38 (1), 37–70.
NAFS Project (2021a) Ämnesprov i engelska för årskurs 6 [English National Test Year 6]. See https://www.gu.se/nationella-prov-frammande-sprak/prov-och-bedomningsstod-i-engelska/engelska-arskurs-1-6/nationellt-prov-i-engelska-for-arskurs-6
NAFS Project (2021b) Ämnesprov i engelska för årskurs 9 [English National Test Year 9]. See https://www.gu.se/nationella-prov-frammande-sprak/prov-och-bedomningsstod-i-engelska/engelska-arskurs-7-9/nationellt-prov-i-engelska-for-arskurs-9
Pomerantz, A. (1984) Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In J.M. Atkinson and J. Heritage (eds) *Structures of Social Action* (pp. 57–101). Cambridge: Cambridge University Press.
Popham, W.J. (2009) Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice* 48, 4–11.
Popham, W.J. (2011) Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator* 46 (4), 265–273. doi:10.1080/08878730.2011.605048
Rasch, G. (1980) *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
Richards, K. (2006) *Language and Professional Identity: Aspects of Collaborative Interaction*. Basingstoke: Palgrave Macmillan.
Sacks, H. (1992) *Lectures on Conversation, Vols I and II* (ed. G. Jefferson; Introduction by E.A. Schegloff). Oxford: Blackwell.
Sacks, H., Schegloff, E.A. and Jefferson, G. (1974) A simplest systematics for the organization of turn-taking in conversation. *Language* 50 (4), 696–735.
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. and Sundqvist, P. (2019) Doing versus assessing interactional competence. In M.R. Salaberry and S. Kunitz (eds) *Teaching and Testing L2 Interactional Competence: Bridging Theory and Practice* (pp. 357–396). Abingdon and New York: Routledge.
Schnurr, S. and Chan, A. (2011) When laughter is not enough: Responding to teasing and self-denigrating humour at work. *Journal of Pragmatics* 43 (1), 20–35. doi:10.1016/j.pragma.2010.09.001
Sidnell, J. and Stivers, T. (eds) (2013) *The Handbook of Conversation Analysis*. Chichester: Wiley-Blackwell.
Silverman, D. (1998) *Harvey Sacks: Social Science and Conversation Analysis*. Cambridge: Polity Press.
Stokoe, E. (2012) Moving forward with membership categorization analysis: Methods for systematic analysis. *Discourse Studies* 14 (3), 277–303. doi:10.1177/1461445612441534
Sundqvist, P., Wikström, P., Sandlund, E. and 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
Sundqvist, P., Sandlund, E., Skar, G.B. and Tengberg, M. (2020) Effects of rater training on the assessment of L2 English oral proficiency. *Nordic Journal of Modern Language Methodology* 8 (1), 3–29. doi:10.46364/njmlm.v8i1.605

Swedish National Agency for Education (2009) *Bedömaröverensstämmelse vid bedömning av nationella prov* [*Rater Agreement in the Assessment of National Tests*]. Dnr/Reg no 2008:286. Stockholm: Swedish National Agency for Education.

Swedish National Agency for Education (2014) *English. Ämnesprov, läsår 2013/2014. Lärarinformation inklusive bedömningsanvisningar till Delprov A. Årskurs 9* [*English. National Test 2013/2014. Teacher Information Including Assessment Instructions for Part A. Year 9*]. Stockholm: Swedish National Agency for Education.

Swedish National Agency for Education (2015) *English. Ämnesprov, läsår 2014/2015. Lärarinformation inklusive bedömningsanvisningar till Delprov A. Årskurs 6* [*English. National Test, 2014/2015. Teacher Information Including Assessment Instructions for Part A. Year 6*]. 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.

Walters, F.S. (2007) A conversation-analytic hermeneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing* 24 (2), 155–183. doi:10.1177/0265532207076362

Weigle, S.C. (1998) Using FACETS to model rater training effects. *Language Testing* 15 (2), 263–287. doi:10.1177/026553229801500205

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, IN: Solution Tree.

Wilkinson, A. (1968) The testing of oracy. In A. Davies (ed.) *Language Testing Symposium* (pp. 117–132). Oxford: Oxford University Press.

Wilkinson, S. and Kitzinger, C. (2006) Surprise as an interactional achievement: Reaction tokens in conversation. *Social Psychology Quarterly* 69 (2), 150–182. doi:10.1177/019027250606900203