Eyes on the ball: Teachers' rating processes when assessing a national L2 speaking test through the lens of their scoring document

Liliann Byman Frisén[1]
Karlstad University

The administration of national second language (L2) tests in schools has increased during the last decades. Speaking is often missing from tests, possibly because it is particularly challenging to assess in a reliable and standardized way, not least when communicative competence is part of the test construct. This study examines how teachers attend to the challenges of L2 speaking assessment, by studying why and how scoring documents are used in the rating process of a national L2 English speaking test in Sweden. Data consist of stimulated recall interviews with 13 secondary-school teachers. The method of analysis was qualitative thematic analysis guided by the *Anthropological Theory of the Didactic* (ATD, Chevallard, 2007). Findings show that teachers reported on a three-step rating process, where scoring documents were used in relation to the purpose of each step. In Step One, the documents were regarded as beneficial for a focus on relevant criteria and for quick notetaking. In Step Two, students' spoken performances were reflected in notes to which teachers referred to analytically decide the score. A third step was added after the score was determined, when notes were passed on to stakeholders. Analysis of interview data indicates washback effects from the test on teaching, and illuminates teachers' assessment responsibilities in a system increasingly affected by accountability.

Key words: L2 speaking assessment, rating process, language assessment literacy, notetaking, standardized tests

[1] Email address for correspondence: liliann.frisen@kau.se

Introduction

In an increasingly multilingual world, foreign language learning, teaching and assessment has gained attention during the last decades, and consequently, a majority of European countries administer national tests in English as a foreign/second language (L2)[2] (European Education and Culture Executive Agency, 2023). However, not even half of the countries include all four language skills (reading, listening, speaking and writing) in the tests (European Commission et al., 2015). Speaking is most frequently missing from national tests, despite a growing recognition that this skill is a catalyst in language proficiency (Figueras, 2019), and "at the very heart of what it means to be able to use a foreign language" (Alderson & Bachman, 2004, p. ix). Previous research shows that language skills in focus for high-stakes testing are also what is being taught in classrooms (Qi, 2005). Therefore, there are reasons to believe that when speaking is missing from national tests, there will be an imbalance in how language skills are taught in the classrooms, to the detriment of explicit teaching of speaking skills (Pakula, 2019).

A plausible reason as to why speaking is missing from national L2 English tests in some countries is that this skill is particularly challenging to assess (European Commission et al., 2015; Hughes & Szczepek Reed, 2017). Firstly, speaking is the language skill that, according to Alderson and Bachman (2004), is the most difficult to assess reliably. An explanation is that oral language use is multicomponential, leading to a plethora of possible criteria that raters can attend to (Hughes & Szczepek Reed, 2017). Secondly, the communicative approach to language teaching and learning inherent in the *Common European Framework of Reference for Languages* (CEFR, Council of Europe, 2020), has led to an inclusion of social aspects of language in the construct underlying speaking tests, as well as the use of testing formats such as peer or group conversations (Borger, 2019). Interlocutor variables (for instance, participants' gender, personalities, and proficiency levels) have been shown to affect assessment and scoring decisions (Borger, 2019; May, 2009, 2011; Sandlund & Sundqvist, 2016). As such, assessment of speaking is context-dependent, which makes it particularly difficult to standardize (Bachman, 2007; Sundqvist et al., 2018). Although several studies indicate that there is broad consensus between raters

[2] In this article, L2 is used to refer to foreign/second/additional languages.

regarding what criteria to apply for assessment of speaking (Borger, 2019; Byman Frisén et al., 2021; Bøhn, 2015; Hasselgren, 1997), research on the various practices and processes involved in high-stakes assessment of speaking and interactional skills still has gaps (Youn & Chen, 2021).

Teachers are often engaged as examiners and raters of speaking tests (Figueras, 2019; Sundqvist et al., 2018). In Sweden, which constitutes the empirical case for the present study, students' own teachers assess speaking in L2 English for all students in Year 6 (ages 12–13) and 9 (ages 15-16) in the National English Speaking Test (NEST), annually administered by the Swedish National Agency for Education (SNAE). It is a high-stakes, summative proficiency test, where students are divided into pairs and are instructed to react to and discuss topics from the test material. There is no specific rater training for teachers acting as raters of the NEST, but extensive assessment guidelines are provided from SNAE. Figueras (2019) states that when teachers are engaged as raters of speaking tests, they "will be making the decisions based on their interpretation of the performances through the lense(s) [*sic*] of a marking scheme" (Figueras, 2019, p. 143). Teachers who act as raters of the NEST sometimes create and use their own scoring document when operationalizing assessment (Byman Frisén et al., 2021), indicating that self-generated scoring documents are helpful when attending to the rating task. This study sets out to examine *how* and *why*.

Since raters' cognitive processes when assessing L2 speaking are underresearched (Han, 2016), studies of the mediating effect of scoring documents, that is, what raters notice when rating, how they distil their observations into a score, and the role of the document in this process, are few (Seedhouse & Satar, 2021; Thai & Sheehan, 2022). Likewise, studies of how raters of peer/group conversations take notes and use the mediating effect of a scoring document of their own choosing are scarce. Given that raters' decisions, as reflected in scores, have real-life consequences for test-takers, studies targeting the very processes of scoring, such as raters' notetaking practices when dealing with the challenges of L2 speaking assessment, are timely and relevant.

Using assessment of the NEST as an empirical case, the present study contributes to such insights, since Swedish teachers (a) rate their own students, (b) rate peer groups of L2 English speakers, and, (c) are autonomous in their scoring decisions (Gustafsson & Erickson, 2018). Thus, there is room for Swedish teachers to assess and score this high-stakes, national test in different ways, which lends a lens through which rating processes when attending to challenging rating tasks could be understood.

The aim of this study is to contribute to a clearer understanding of teachers' rating processes when assessing and scoring L2 English speaking, as processes emerge from teachers' reports of their notetaking practices when rating the NEST. A second aim is to gain new knowledge about what value teachers assign to the use of a scoring document in the rating process.

The following research questions guided the study:

- RQ1: In what way(s) do teachers report taking notes when assessing students' speaking skills?
- RQ2: In what way(s) are notes reported to be used when reaching a conclusion on a summative score?
- RQ3: What reasons are reported for the creation and use of self-generated notetaking documents for assessment of speaking in L2 English?

Additionally, results from the study might shed light on practical issues related to speaking assessment, such as what conditions teachers deem beneficial for solving the rating task.

Background

Raters' conceptualizations of L2 speaking assessment

In a literature review on differences and similarities regarding raters' cognitive processes when assessing L2 speaking (Han, 2016), it is demonstrated that studies of what factors raters heed when assessing are the most prevalent. For example, studies have shown that raters disagree on the relative importance of rating criteria in scoring documents (Ang-Aw & Goh, 2011), attend to different aspects of spoken performance (Orr, 2002; Seedhouse & Satar, 2021), and focus on different aspects depending on candidates' levels of performance (Sato, 2012). Other studies have shown that raters generally agree on what criteria should be in focus for assessment (Borger, 2014;

Byman Frisén et al., 2021; Bøhn, 2015; Hasselgren, 1997), although raters may differ when it comes to how they orient to these.

Interaction-based tests, where test-takers are conducting the test with peers, generally result in interaction more in line with naturally occurring conversations compared to the oral proficiency interview (OPI), where an individual candidate is interviewed by a trained native speaker rater (Ducasse & Brown, 2009; van Lier, 1989). Interaction-based tests are therefore preferred when communicative ability is assessed. Studies of rating processes when assessing L2 peer groups show that raters differ both in conceptualizations of rating criteria, as well as how to apply these to student performances (e.g., Borger, 2019; Ducasse & Brown, 2009; Frisch, 2021; May, 2011; Sandlund & Sundqvist, 2019, 2021). For instance, May (2011) found that raters' understanding of *co-construction* differed, i.e., to what extent speaking performance is co-constructed by participants in the assessment situation. Also, both Ducasse and Brown (2009) and May (2009) found that raters of interaction-based tests attended to criteria not included in scoring documents, such as *gaze* or *body language*, indicating that non-verbal features of speaking are aspects of interactional skills that raters notice.

An interview study on teachers' assessment practices showed that teachers' conceptualizations of *speaking proficiency* as outlined in Swedish policy documents varied (Frisch, 2021). However, there was room for all differences observed since policy was expressed in broad terms. Findings also showed that teachers' rating processes of the NEST were affected by a perceived lack of time, both for preparing for their roles as raters and for collegial discussions about what score to award.

The use of scoring documents in rating processes

For any assessment, there is a number of potential criteria available, commonly divided into *manifest* and *latent* criteria (Sadler, 1989). The former relates to criteria that are consciously attended to, whereas the latter are criteria raters are unconscious of but when triggered can temporarily become part of the set of manifest criteria. An example of a latent criterion is *body language* in assessment of speaking skills; although not part of the manifest criteria, students' body language might affect decisions about the quality of their performance. For human-mediated rating of writing and speaking proficiency, criteria are usually expressed holistically or analytically (Khabbazbashi & Galaczi, 2020). Scoring documents list manifest criteria that reflect the test construct under scrutiny. An analytic document lists all criteria separately, and a separate score is given for each of the analytic criteria, while a holistic document contains criteria to be considered simultaneously (Brookhart, 2018; Brown, 2012). Whether one is to be preferred over the other is still debated (see Panadero & Jönsson, 2020), but holistic assessment is usually preferred over analytic for standardized tests since making one scoring decision is considered less cognitively demanding for raters than making several (Brookhart, 2018; Brown, 2012; Davis, 2018; Xi, 2007). However, when raters assess holistically, they may differ in how they conceptualize criteria, and as a consequence, criteria used to inform the score might be weighted differently by the raters involved. Also, a holistic score cannot be used to communicate test-takers' attainment of separate criteria, and is therefore not suited for formative feedback (Ma, 2022).

Although there is a paucity of research on rater cognition in rating processes of L2 speaking (Han, 2016; Seedhouse & Satar, 2021; Thai & Sheehan, 2022), the descriptive framework for assessment of constructed responses developed by Bejar (2012) presents a model of cognitive processes raters go through. In this framework, the rating process is seen as consisting of two phases: the "assessment design phase" and the "scoring phase" (Bejar, 2012, p. 5). In the first phase, raters form a mental representation of the document. In the second phase, raters form a mental representation of the text they have read and compare and contrast both mental representations to decide a final score. A recent study (Thai & Sheehan, 2022) examined the rating process of 13 raters of a L2 speaking test through data collected via think-aloud protocols (cf. also May, 2009; 2011). Findings showed that raters experienced similar stages in their rating processes, but there were differences between experienced and novice raters, where experienced raters were more likely to provide a verbal justification of the score they awarded by relating their scores for all the assessment criteria in the order they were presented in the document, whereas this was less frequently done by novice raters. Also, when discussing their findings, the authors compared their results to a model for L2 writing and found that the main difference was that L2 speaking raters attended to several criteria simultaneously (Thai & Sheehan, 2022). A reason proposed by the authors was that raters of writing can refer to the visible text features in the texts they are assessing, while this is not possible for raters of speaking.

Studies of how scoring documents affect scores awarded in speaking assessment show conflicting results. In a study by Khabbazbashi and Galaczi (2020) on the effect of three different scoring models (holistic, analytic and part marking) on measurement properties and CEFR classifications, the choice of model was found to significantly impact the CEFR levels awarded to candidate performances. In contrast, Ma (2022) compared analytic and holistic scoring of 127 international teaching assistants' speaking skills assigned by ten raters and found that both types of scores differentiated candidate performances in a similar way. The study supports findings from Xi (2007) where candidates taking the TOEFL Academic speaking test were rated analytically. Xi (2007) concluded that analytic scores were not superior to holistic, as they did not provide enough additional information beyond what holistic scores could offer. Also, for tests where more variability in criteria attainment can be expected, Xi argues that analytic assessment would entail reliability problems. A common denominator between the three studies (Khabbazbashi & Galaczi, 2020; Ma, 2022; Xi, 2007) is that only the outcome of rating (i.e., the score) was examined. What still remains unanswered is in what way raters interacted with the scoring documents they used, and the process raters went through when rating.

As demonstrated in previous research, raters' preferences for using particular types of scoring documents differ (Heidari et al., 2022; Horák & Gandini, 2021), and even when the same scoring document is used by raters, they vary in how they understand and adapt it (Ang-Aw & Goh, 2011; Seedhouse & Satar, 2021). As noted by Thai and Sheehan (2022), further investigation into the rating processes experienced by raters is called for, something that can be achieved "if the researchers could have interviewed the raters about their opinions on the rating scale, their approach to it and reflection about their rating practice" (Thai & Sheehan, 2022, p. 49). The present study is a response to such a call.

Method

The setting – the National English Speaking Test

In Sweden, the national test of L2 English includes speaking, receptive skills (reading and listening), and writing. A holistic approach to assessment is adopted by the Swedish National Agency for Education (SNAE) in their guidelines for the assessment of L2 English proficiency. The focus of this study is on Part A, *Speaking* (NEST), where guidelines include recommendations regarding administration (for instance, how to pair students), instructions on how to conduct the test (for instance, for the teacher to stay in the background of the students' conversation), as well as guidelines for assessment that include a one-page document (see Appendix A), a notetaking document (see University of Gothenburg, 2023), and audio files with benchmarks. The benchmarks are assessed and commented on by SNAE's expert raters together with a summative score.

Data

Data consist of stimulated recall interviews (Gass & Mackey, 2016) with teachers of English in Sweden who act as raters of the NEST (Year 6/Year 9). All teachers were defined as female following external body characteristics and self-reported first names. The study constitutes one part in a broader project on assessment of the NEST in compulsory schooling; thus, teachers from both Year 6 and Year 9 were selected. Data were retrieved in two steps: Step One (*N*= 5) and Step Two (*N*= 8). Participants in Step One were selected in connection with a previous project on content and design of teacher-generated scoring documents (Byman Frisén et al., 2021). Findings from the project indicated the use of analytic assessment when teachers used their own scoring document for notetaking and scoring decisions. Step Two was initiated to examine the role of notetaking documents in the assessment process further, by recruiting new participants from professional networks of teachers in Years 6 and/or 9. Semi-structured interview guides (Kvale & Brinkmann, 2014) were used in both steps (see Appendix B).

Interviews were conducted with the teachers during the time of the school year when the speaking part of the national test was conducted (i.e., in the middle of the assessment period, that spanned six weeks), or when they had just finished assessment of the NEST. Interviews were audio recorded and transcribed orthographically. Length of interviews varied between 33 and 78 minutes, with a mean length of 52 minutes.

Teachers were asked to bring the scoring document they used for notetaking during the NEST to the interview, which was used as stimulus and visual aid (Bryman, 2008) when describing their notetaking and scoring practices. Scoring documents commonly contained teachers' authentic notes taken during assessment, with information about students' names and notes about students' spoken production. For ethical reasons, scoring documents could not be collected as data. However, the researcher took notes on the design of the scoring document, type of notes written (e.g., *words/phrases*, *errors*), and how the teacher oriented to their document (i.e., pointing to different parts) to complement spoken interview data. In the following sections, scoring documents reported to be used by teachers in the study will be referred to as notetaking documents (NTDs).

Participants

All thirteen participants were certified teachers of L2 English working in schools situated both in urban and rural areas across Sweden. Twelve of the teachers were experienced in teaching English, with a range of 11–26 years of experience, whereas one had taught English for five years. Since none of the teachers assessed the NEST annually, the numbers for teacher experience and times as rater differed (see Table 1). However, as teachers found it difficult to report an exact number of times acting as rater of the NEST, all teachers made an estimate (see Table 1). All participant names have been pseudonymized for ethical purposes.

Table 1. Participant experience of teaching and NEST

Teacher	Teaching experience (years)	Teaching Year 6	Teaching Year 9	Estimated number of years acting as rater for the NEST
Beata	20	x	x	17
Carol	16	x		4
Cecilia	24		x	12
Céline	11	x	x	5
Hannah	17		x	5
Harriet	22		x	15
Julia	15	x	x	13
Laura	26	x		15
Mary	17		x	11

Miriam	25	x		15
Monica	13		x	13
Susanne	16	x		8
Tina	5	x	x	4

Several of the participating teachers worked in schools with both Year 6 and Year 9 and thus had experience from teaching and assessing English for both groups, although they predominantly assessed only one of the groups within the same academic year.

Analytic process and approach

Data were analyzed using qualitative thematic analysis (Braun & Clarke, 2006), for which the software program NVivo 12 (QSR International, 2018) was used. Analysis was guided by the research questions for the study and by the framework *Anthropological Theory of the Didactic* (ATD, Chevallard, 2007). According to ATD, knowledge is a changing reality, formed and affected by the institution and participants within which it exists. The idea of *praxeologies,* which consists of *praxis* and *logos,* is part of the ATD framework. Praxis is a type of *task* as well as the *technique* used to carry out the task, whereas logos is the logic behind using that particular technique for that particular task (the *technology* of the technique) as well as *theory* justifying the technology. Viewing NTD as the technique used to carry out the task of assessing L2 English speaking, the ATD framework is used in this study to analyze how teachers use this technique as well as analyzing the logos behind it, i.e., the discourse of why and how NTDs are beneficial for carrying out the task.

For the analysis of interview data, the first five phases in Braun and Clarke's (2006) approach to thematic analysis were followed: 1: Familiarizing oneself with the data, 2: Generating initial codes, 3: Searching for themes, 4: Reviewing themes, and 5: Defining and naming themes (Braun & Clarke, 2006, p. 87). An inductive process was aimed for in order to code for as many themes as possible. Coding was done by reading through all transcripts to search for patterns and/or interesting features. At the end of an iterative process between phases 2–4, a mind-map was drawn of the first tentative themes. A coherent narrative for each of the themes generated was created, in which data extracts illustrative for the themes were selected, to provide context and relevance to themes (Kiger & Varpio, 2020). The second step in this phase was to take a more deductive approach to the data, guided both by the research questions for the study as well as the idea of *praxeologies* inherent in ATD. Teachers' reports of what kind of notes they take in the rating process, how notes are taken, as well as how notes of rating criteria are considered for the summative score were seen as *praxis* (Chevallard, 2007), whereas teachers' reports of why notes are taken in this particular way and why they are beneficial for scoring decisions were seen as *logos* (Chevallard, 2007). Themes in the data were then generated to reflect *praxis* and *logos* (see Figure 1).

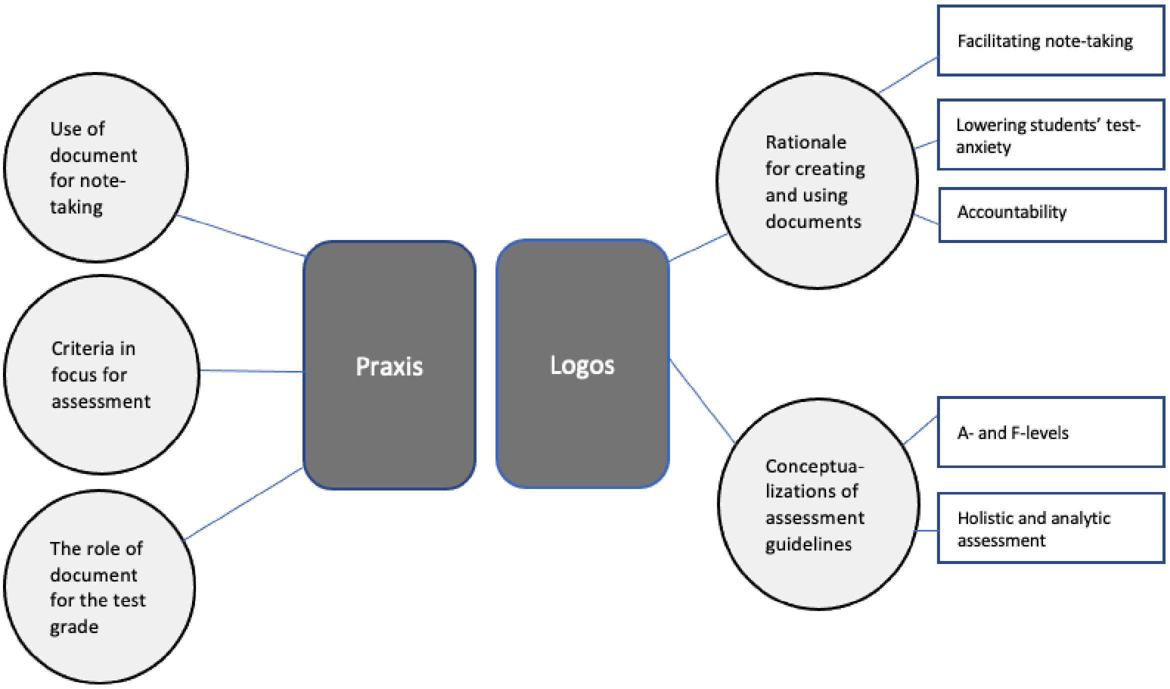
Use of document for note-taking
Criteria in focus for assessment
The role of document for the test grade
Praxis
Logos
Rationale for creating and using documents
Facilitating note-taking
Lowering students' test-anxiety
Accountability
Conceptua-lizations of assessment guidelines
A- and F-levels
Holistic and analytic assessment

Figure 1. Themes generated in phase 5 of thematic analysis (Braun & Clarke, 2006).

Three themes were found to reflect *praxis* and two themes were found to reflect *logos*. Also, for the two themes reflecting *logos*, a total of five subthemes were generated.

Findings

In this section, results will be presented in relation to each theme[3] generated in the analytic process.

[3] Except for the sub-theme *A- and F-levels,* since the data is not part of this particular study.

Use of documents for notetaking

All teachers interviewed reported that they took notes as part of their rating process of the NEST, but they differed when it came to what kind of document they used for this task. No NTDs were collected, but from the teachers' descriptions, NTDs that were teacher-generated could be categorized as Type A, B or C (see Tables 2–4).

Table 2. Example of NTD categorized as Type A – Grid-like NTD with performance level descriptors in separate columns

Criterion	**Low level**	**Medium level**	**High level**
Criterion 1 (e.g., *Richness and variation*)	Description of Criterion 1 for the low level	Description of Criterion 1 for the medium level	Description of Criterion 1 for the high level
Criterion 2 (e.g., *Vocabulary*)	Description of Criterion 2 for the low level	Description of Criterion 2 for the medium level	Description of Criterion 2 for the high level
Criterion 3 (e.g., *Grammar*)	Description of Criterion 3 for the low level	Description of Criterion 3 for the medium level	Description of Criterion 3 for the high level
Criterion 4 (e.g., *Communicative strategies*)	Description of Criterion 4 for the low level	Description of Criterion 4 for the medium level	Description of Criterion 4 for the high level
Criterion 5 (e.g., *Adaptation, and engagement*)	Description of Criterion 5 for the low level	Description of Criterion 5 for the medium level	Description of Criterion 5 for the high level
Criterion 6 (e.g., *Comprehension and clarity*)	Description of Criterion 6 for the low level	Description of Criterion 6 for the medium level	Description of Criterion 6 for the high level

Table 3. Example of NTD categorized as Type B – Grid-like NTD with no level descriptors

Criterion	+	-
Criterion 1 (e.g., *Richness and variation*)		
Criterion 2 (e.g., *Vocabulary*)		
Criterion 3 (e.g., *Grammar*)		
Criterion 4 (e.g., *Communicative strategies*)		
Criterion 5 (e.g., *Adaptation, and engagement*)		
Criterion 6 (e.g., *Comprehension and clarity*)		

Table 4. Example of NTD categorized as Type C – Grid-like NTD with no level descriptors and students' names

Criterion	**Student 1** + /-	**Student 2** + /-
Comprehension and clarity		
Richness and variation		
Context and structure		
Adaption to purpose, recipient, and situation		
Communicative strategies		
Fluency and ease		
Breadth, variation, clarity, accuracy		
Adaption to purpose, recipient, and situation		

Two teachers reported using the notetaking document provided by the SNAE (which resembles a Type C document; for comparison with the original document, see University of Gothenburg, 2023), one teacher reported using two blank papers, and the remaining ten teachers reported using a document that was teacher-generated, either by themselves alone, by the group of English teachers at their school, or by a colleague. As a result, NTDs differed in appearance as well as in what they afforded the user in terms of notetaking (see Table 5).

Table 5. Notetaking document and scoring procedure used by each teacher

Teacher	Type of NTD	NTD author	Timing of notetaking	Symbols	Audio recordings of student pairs	Score decision
Beata	Type A	Team of English teachers at the school	During and after	Underlines text. Additional comments.	Yes	With colleagues directly after
Carol	No rubric. Instead, two blank papers.	n/a	Mostly after	On one paper: Colours, clouds, and arrows. Additional comments. On the other paper: Only the score for each student.	Yes	Alone but can turn to colleagues for advice if need be
Cecilia	Type B	Generated by colleague	Mostly during	Checks boxes. Plus-signs. Additional comments.	No	With colleagues directly after
Céline	Type A	Team of English teachers at the school	During and after	Puts crosses in boxes and underlines text.	No	With colleagues directly after
Hannah	Type A	Self-generated	During	Puts crosses in boxes. Additional comments.	No	With colleagues directly after
Harriet	SNAE's note-taking document	Generated by SNAE	During	Plus-column: letters to indicate score level (E, C or A). Minus-column: Comments.	No	With colleagues directly after
Julia	Type A	Self-generated	During	Puts letters in boxes. Additional comments.	Yes	Alone but listens to some audio recordings with colleagues to "calibrate" scoring
Laura	Type C	Self-generated	During	Plusses and minuses. Signs. Comments.	Yes	Assesses alone but turns to colleagues for difficult score decisions
Mary	Type A	Self-generated	During	Arrows that indicate level (within a box). Additional comments.	No	With colleagues directly after
Miriam	Type A	Self-generated	Mostly after	Crosses in boxes. Additional comments.	Yes	With colleague when listening to all audio recordings together
Monica	Type A	Team of English teachers at the school	During and after	Puts crosses in boxes. Underlines text. Additional comments.	No	With colleagues directly after
Susanne	Type A	Generated by a colleague online	During	A cross in a box ”here and there”. Additional comments.	Yes	With colleagues directly after
Tina	SNAE's note-taking document	Generated by SNAE	Mostly during	Comments.	Yes	Alone but brings up all difficult score decisions with colleagues later

Teachers reported putting crosses and/or other symbols in the squares of NTDs of Type A to mark qualitative level. For some teachers, the position of the cross within the square marked even more “fine-grained” quality: a cross to the left of the low level-square was close to the score F (i.e., the lowest score possible = fail), whereas a cross to the right was close to the score D (the third lowest score). Another method was underlining, or colouring, pre-written text (wholly/partly) illustrative of students’ attainment of criteria.

A common practice among the interviewed teachers was to make up one’s own symbolic system, something that afforded quick notetaking. For instance, “plus” and “minus” signs were used to mark good and bad examples of language use. Some teachers also used these symbols to mark the score, where for instance the number (1–3) of plus signs (+) indicated the score, where one + equalled a low score, and three + signs equalled a high score. In addition to plusses and minuses, abbreviations, arrows, clouds and different colours were used to mark quality, an example being “SE” for “Swenglish”. Thus, teachers developed their own personalized way of taking notes that would make sense mostly to themselves. Even when the same notetaking document was used (as was the case with the two teachers using the notetaking document from the SNAE), the teachers’ use of it for notetaking differed (cf. Ang-Aw & Goh, 2011; Seedhouse & Satar, 2021).

All teachers reported that they were present in the room with the students during the NEST, and all but one took notes then. However, they differed when it came to how extensive their notes were. Most of the interviewed teachers took extended notes, while some said that they tried to refrain from taking notes while students were talking. These teachers believed that their notetaking practices made students nervous, and therefore prioritized taking notes after the students had left the room. Easing students’ nervousness was therefore an important factor for decisions on *how* to take notes as well as *when* to take notes.

Criteria in focus for assessment

Rating criteria in focus for assessment differed in the NTDs. Two teachers used the National Agency for Education’s document where *assessment factors* (see Appendix A) were listed, and others used these assessment factors as inspiration for their own creation of a NTD. Teachers using Type C documents listed analytic criteria from the SNAE in their NTDs, whereas teachers using Types A and B documents listed their own choice of criteria (see Tables 2–4).

According to some of the participants, other sources that inspired rating criteria in teacher-generated documents were assessment guidelines for the national test of spoken proficiency in L1 Swedish, and scoring documents for speaking assessment found online in different teacher fora. One of the teachers rating the NEST6 decided to include the criterion "argue" in her document, as she was inspired by the notetaking document for the speaking part of the national test in the subject Swedish, where, according to her, students' ability to argue for a standpoint or view is assessed[4]:

> That's where I got it from because I saw that *argue* was included in that one. So in order to have something here [...], I believed that *argue* should be included. Even if it's not part of the *knowledge requirements*, one of the things one assesses there, it is actually part of this test. (Miriam)

All teachers, even one with a very detailed NTD of Type A, wrote additional notes in the margins. Examples of what was noted were errors in students' production, such as wrong intonation, or the use of "Swenglish". According to the participants, these comments were sometimes meant for their own use only, to aid in the rating process, but the most common answer was that it enabled formative feedback to students. Many teachers also reported how they noted good use of vocabulary in students' productions, and one of the teachers gave an explanation as to why:

> If you don't have the words, you won't be coherent anyway. All parts are important really, but vocabulary is pretty easy to separate. If you have a rich vocabulary, you will automatically be placed higher up, and that is where you show signs in a way how far you have come in your English. It gives clear signals. And that is often the reason why one starts by assessing it. (Miriam)

Other examples included body language (such as nodding, shifts in body posture, eye contact) and how eager (or not) students seemed to be when it came to participating in the conversation. Also, some teachers reported noting whether students seemed nervous.

[4] All quotes from interviews have been translated from Swedish into English by the author.

Content (i.e., *what* students say) seemed to be an important part of assessment, since a majority of the interviewed teachers said that students' ability to develop their thoughts was something they paid particular attention to in assessment. To assess this ability, different strategies were used. One of the teachers sometimes noted how many cards students used for their conversation as a sign of how "developed" their conversation was; the more cards used, the less they had developed their thoughts. Another teacher rather viewed the test as consisting of three different rounds. For a student performance to be awarded the highest score, the student would have to complete the first two and reach round three, since that is where the "deepest conversations" take place. This finding might reflect the fact that when assessing speaking proficiency, assessment decisions need to be taken quickly while considering numerous factors simultaneously (cf. Thai & Sheehan, 2022). In this situation, it might be difficult to focus on both what students say (i.e., content) and how they say it (i.e., linguistic aspects). Teachers therefore developed discernible and concrete signs of attainment of the criterion *content*.

The role of NTDs for the test score

There are differences between the interviewed teachers regarding how they reported using the NTDs in the process of deciding the final score that stem from different rating procedures. Co-assessing the NEST with another colleague, who was also present in the room when students took the test, was reported to be the most common procedure (see Table 5). These teachers reported coming to a score decision in two ways. One was to use a NTD of their own preference in the discussion with their co-rater. The other was to take notes individually on a blank paper, and to fill in a NTD together when students had left the room. For both approaches, individual notes were used as a mnemonic device in the score discussion with their colleague. Almost all said that the score was decided by having a 'what-did-you-hear'-dialogue first. Then they discussed back and forth what the score was going to be:

> We kind of start with "what was your impression?", and then we look at the impressions and "this was good I think, this is what I heard". And then one says, "did you also hear that? This is something I found less good, what do you think? Did you also see that?" And then we discuss: "this is probably a -, or is on a C-level, what do you think?", and we discuss like that back and forth. (Cecilia)

Some teachers described how they discussed rating in terms of talking about the score:

> But then I had circled Cs for the student, and she, well, I guess this is a C, ok, that's good, then we think alike in a way. (Julia).

However, most teachers described how their discussions were based on the different rating criteria:

> We usually go bit by bit when we discuss "so what do you think about this student's content?", and then we move on to "well, I think that this student uses good strategies. Yes, I think so too". So we go bit by bit in a way. (Tina)

A few teachers reported coming to a score decision about their students' scores mostly on their own. For these teachers, the assessment guidelines, and specifically the benchmark examples, seemed to play a particularly important role, as sole raters more often than co-rating teachers reported how they consulted this material whenever they were in doubt.

Teachers using a Type A NTD reported how it helped them discern a pattern of what score to award, and one teacher said that the pattern that emerged through crosses in boxes made her see what the score was. Another teacher reported how she assigned different weights to criteria in her NTD, where the first three criteria of her document were seen as more important than the others. Although the criteria (*comprehension and clarity*, *richness and variation*, and *context and structure*) were separated in her document, she saw them as sub-criteria of an overarching criterion that she called *content*, which took precedence over the seven remaining criteria in her decision of what score to award. These strategies seem to have helped teachers in discerning different score levels.

Rationale for creating and using NTDs

Several reasons were given by the teachers when asked why they decided to create their own NTD, or why it was beneficial to take notes in this way. Three sub-themes were identified through the thematic analysis.

Facilitating notetaking

Teachers reported that having access to one's own NTD, particularly one with pre-defined criteria, made it easier to take notes without losing track of the students' conversation, since it was hard to take (extended) notes and listen to the students at the same time.

> The one from the Agency [SNAE] encourages you to write and that is something that you don't have time for while you listen [...] as soon as you write sentences you lose what they talk about. You lose the students when writing sentences. (Laura)

Teachers expressed the view that a Type A NTD provided an overview and a structure of assessment criteria. This, in turn, was beneficial for a more straightforward rating process, as the teacher got an idea of the different levels and what distinguishes these. Furthermore, the pattern that emerged through teachers' notetaking practices when using these documents made the rating process more efficient:

> I get a score right away. Wham, bam, done! (Mary)

Thus, according to the teachers, Type A NTDs made it easier to see the score and/or what level the student's speaking skills were on.

Lowering students' test anxiety

The importance of a safe and non-threatening atmosphere in the rating situation was mentioned by many of the interviewed teachers. In order to avoid making students nervous, participants reported being discreet when it came to the teachers' role as raters, which included concealing notetaking. Teachers reported that notetaking was imperative, and therefore some teachers offered verbal accounts for why they needed to take notes, such as "I am getting old and forgetful". One teacher reported how she accustomed her students to her notetaking practices prior to the test so they would be prepared for this situation. However, most of the interviewed teachers either refrained from taking notes while students were in the room, took as few notes as possible, or placed themselves in a position so that their notetaking was not as visible. Having access to a document with pre-printed text (for example, Type A) made it easier to take notes in a discreet and quick manner. This is also the reason given for the creation of symbols as a notation system commonly used by participants.

Accountability

All of the interviewed teachers stated that notetaking was imperative for mnemonic purposes when deciding the score; notes helped them remember the conversation that students had, what stood out to them as raters in the rating situation, and aspects complementing audio recordings (such as body language and signs of eagerness/nervousness) when these were made.

One of the main benefits reported of using a NTD with pre-printed criteria was that it made rating criteria visible. Several participants stated that words and phrases inherent in their document helped them focus just on the criteria for rating. Some teachers also reported carefully reading the assessment guidelines each year and adjusting their NTD to include any changes. Thus, NTDs helped teachers feel that their assessment and scoring decisions were as valid as possible. Some of the teachers using a document that included performance level descriptions reported that, although they were happy with their own creation, they would have liked to have a similar one from SNAE. This way, teachers stated, rating would be more reliable and they would not have to put time and effort into making their own version:

> I look to see if my document is comparable to what is written there [assessment guidelines] or have they made their own document now so I don't have to use mine, because I would like to have the same document in the whole country, but as long as they don't make one I think that, well, I'll use my own. (Miriam)

Most of the participants expressed the opinion that an added value of visible rating criteria was that assessment became clear to students. Although the actual document used in the summative situation was seldom handed out to students afterwards, it is clear from the interviews that the teachers routinely provided formative feedback. An analytic document made it easier to provide this feedback in a more detailed way, to explain why students were awarded a specific score, what they could develop in the future, and/or what progression looked like:

> Well, then I can explain to this student in the end that this is something you need to practice more. You're good at, let's say, your pronunciation is very, very good [...] I feel that your fluency maybe is not quite there, but your pronunciation is good, and your vocabulary is quite ok, so you need to practice more words. (Laura)

This seemed particularly beneficial for low-performing students, as it could help teachers communicate what they, in fact, did manage to accomplish despite a low score. Moreover, two teachers specifically addressed the fact that rating criteria became visible also to parents or guardians. Since a document that included performance level descriptions clarified what aspects of oral proficiency each student needed to develop, parents could help their children develop these specific aspects.

Conceptualizations of assessment guidelines

Holistic/analytic

The score-pattern that emerged for teachers using a grid-like document with performance level descriptions was reported to be used for a discussion of strengths and weaknesses shown for each criterion. Several of the participants stressed holistic assessment over analytic, where rating criteria were seen as indicators of quality in student productions (or lack thereof) and subsequently used for holistic assessment:

> If you were to follow this [her document] adamantly all the time then it might be something like THIS [points to the crosses in boxes] and then it cannot be anything but a D, but then you need to go back to the whole picture, because I think that the whole picture is always more important. (Julia)

At the time of data collection, SNAE instructed teachers to assess holistically by consulting the holistically expressed *knowledge requirements* (Swedish: *kunskapskrav*). However, none of the participants referred to these when reporting on holistic score decisions. Instead, teachers referred to holistic assessment as holistic in relation to the different rating criteria inherent in their NTDs. One of the teachers described rating criteria in focus for the NTD that all teachers at her school used when acting as raters of the NEST, and the rationale behind its creation:

> We have broken down the knowledge requirement into the parts that one actually looks at. Because it is not evident when one looks at the knowledge requirements what exactly is measured. Here [her document] it becomes clear. As you can see, we have used the bullet points [SNAE's *assessment factors*], we look at the same things. This is much easier to show students and what we actually look at. (Monica)

When providing feedback, one of the teachers who used SNAE's notetaking document for assessment (Harriet) reported that she transformed her own notes into an analytic document where feedback was given for each criterion. In comparison, Laura, who used an analytic notetaking document in the form of a grid during rating, provided feedback by writing a coherent text, which according to her was a holistic account of students' strengths and weaknesses.

Discussion

From teachers' reports of their notetaking practices and scoring decisions, a two-step rating process similar to the conceptual framework proposed by Bejar (2012) was identified. In the first step, teachers listened (and watched) students attentively and took notes of aspects they found noteworthy. In the second step, teachers drew on their notes to decide on a score. However, findings also show that a third step was added, where notes were transformed into information that could be passed on to stakeholders. Since the three steps had different purposes and outcomes, answers to the logos of NTDs for the rating process need to be related to the respective purpose for each step.

The logos of notetaking in the first step was to capture students' speaking skills, and in some cases, conditions affecting the expression of students' skills in the test situation. The discourse behind using the technique of a notetaking document in this step was that a document where manifest criteria were listed helped teachers "keep an eye on the ball", which a) helped teachers attend to what was seen as relevant only and b) lessened the cognitive burden. Although NTDs afforded teachers a list of manifest criteria, all teachers reported taking additional notes, signaling that certain context-specific aspects could not be included in pre-written criteria. Additional notes that teachers took were either taken to exemplify manifest criteria (such as examples of vocabulary, or grammatical errors), or in the form of latent criteria such as whether students seemed nervous. Both Ducasse and Brown (2009) and May (2009) found that raters of peer interaction paid attention to criterion-irrelevant aspects, such as *gaze* and *body language*, which led Ducasse and Brown (2009) to conclude that different types of non-verbal language might signal evidence of interactional ability or lack thereof. This is a finding also in the present study, since teachers took notes of latent criteria such as *body language* and *eye contact*, and reported that these were discernable signs of students' eagerness to participate in the test, and thus, students' ability to communicate. Likewise, some of the teachers reported noting students' nervousness, which is also a latent, and seemingly a non-criterion-relevant aspect. Although none of the teachers explained why students' nervousness was noted, a possible reason might be that it is one step in attempting to create a non-threatening rating situation when acting as rater. Downplaying the high-stakes characteristics of the situation and their role as raters by concealing their notetaking could also help create a safe atmosphere. Teachers who used NTDs with performance level descriptors particularly stressed that their documents were beneficial for the purpose of easing students' nervousness. For these teachers, the *theory* supporting the use of *technology* (Chevallard, 2007) was therefore that less notetaking during assessment (with students present in the room) led to less nervous students.

The logos of using NTDs in the second step was that they represented an account of students' speaking skills deemed sufficient for distilling the observed performances into a score. As such, it functioned as a "text" that could be used for comparisons and contrasts with their mental representations of what quality should look/sound like (Bejar, 2012). Having noted verbatim what students said (e.g., specific formulaic phrases) could be particularly helpful for this purpose. In line with Seedhouse and Satar (2021), teachers in the present study reported coming to score decisions in different ways. For instance, step two of the process could be facilitated by taking notes in the first step so that differentiation between score levels emerged. However, most teachers reported on a rather time-consuming second step that for most of them included discussions with colleagues about what score to award. Findings show that criteria were attended to analytically in this process, although teachers reported taking a holistic approach to the score decision, where the different criteria were seen as pieces of a puzzle. These findings differ from Frisch (2021), where teachers reported not having time for collegial score discussions, and that the score decision was first taken holistically before the analytic criteria were attended to. A possible explanation is that Frisch's interview study was conducted in 2013. Since then, the NEST has been moved to the autumn term, while the remaining parts of the national test in English (receptive skills and written proficiency) are conducted during the spring term, which seems to have freed time for teachers to attend to the NEST in a more thorough way.

Although previous research on speaking assessment based on peer interaction has shown that interlocutor variables might affect assessment and scoring decisions (see e.g., Borger, 2019), none of the interviewed teachers stated that awarding an individual score to a production that is jointly constructed by the pair was problematic (cf. May, 2011). However, teachers reported that pairing of students required much time and effort. Pairing students according to level of proficiency is something teachers are advised to do in the guidelines from SNAE. This was essential, according to the interviewed teachers, since it helped create a safe atmosphere and as such eased students' production of speech. Teachers also reported noticing whether one student took more responsibility for carrying the conversation forward than the other, and the extent to which students helped each other, even though it was unclear if and how these factors affected scores awarded.

Findings shed light on a rating process that can be said to be both cognitively and administratively demanding with a blurry line between *formative* and *summative*, as well as between *holistic* and *analytic* assessment. A plausible reason is the fact that the raters were students' own teachers. The way in which teachers reached a score decision can be defined as holistic assessment, but where *holistic* was conceptualized in relation to manifest and latent criteria included in each teacher's NTD. Also, since holistic assessment is not suitable for formative assessment (Ma, 2022), teachers gave formative feedback to students, for the most part in an analytic way. Findings from the study do therefore not support Xi's (2007) argument that holistic assessment is less cognitively challenging than analytic assessment for L2 speaking raters, at least when raters are students' own teachers and might be held accountable for the scores they award. Although the NEST is a summative test, formative assessment was central in teachers' reports of their rating processes. Being actively engaged in the rating process gave teachers access to information that could be used for the purpose of subsequent teaching of speaking skills, and thus, could lead to positive washback from the test on teaching (Alderson & Wall, 1993). The administration of the NEST during the autumn term leaves room for several months of teaching where formative comments on students' speaking skills can be attended to before the final score of the subject English is to be decided. Findings indicate washback from the NEST on teaching, also in the form of teaching-to-the-test activities prior to the test, but further studies are needed in order to examine washback effects, and in what ways teachers' conceptualizations from being engaged as raters transform into teaching activities.

Limitations of the present study need to be addressed. Even though teachers' own NTDs were used as stimuli during interviews, data are self-reported. It is therefore possible that the data do not reflect real-life notetaking practices and scoring behaviors. One factor could be time related, as Gass and Mackey (2016) advise researchers to conduct stimulated recall interviews within 48 hours. Due to practical reasons, not all interviews in the present study were conducted within that time frame.

Another factor that might have affected the reactivity of the data is that interviewees might have said what they thought the researcher wanted to hear, rather than reporting their actual behavior. In response to such limitations, future studies might adapt additional methods and/or use video-recordings of teachers' notetaking practices in the assessment situation as stimuli during interviews. Despite these limitations, findings from the present study show that raters develop their own personalized way of taking notes to capture the gist of what students say and/or do. A recommendation to school leaders is therefore to free enough time for teachers who are engaged as raters of high-stakes L2 speaking tests to discuss their notes and understanding of criteria, discuss scoring decisions and/or co-assess students, as was indeed done in some of the schools where participating teachers worked. Rater training for the NEST is next to impossible, considering the vast number of teachers involved. However, affordances and constraints of the three-step rating process found in the study can be taken into consideration when preparing teacher students for the task of assessing and rating L2 speaking proficiency.

Conclusion

The aim of this study was to contribute to a clearer understanding of teachers' rating processes when assessing and scoring L2 English speaking, as processes emerged from teachers' reports of their notetaking practices when rating the NEST. One of the main concerns of testing speaking is how criteria encompass speech as realised (also) outside of the walls of instruction, while at the same time being conceptualised in the same way by raters (Hughes & Szczepek Reed, 2017). The present study shows how teachers attended to this concern. It also shows how teachers dealt with the assessment responsibilities put on them in an education paradigm increasingly affected by accountability (Fulcher, 2012). Teachers reported that their NTDs, as well as their notetaking practices, helped them attend to aspects regarded to be in focus for assessment, they facilitated collegial discussions and score decisions, and they were helpful in lowering students' test anxiety (thereby facilitating students' production of speech as authentically as possible). In addition, NTDs provided teachers with a tool that can be shared with colleagues, students and parents – making it possible to provide detailed formative assessment as well as to account for attending to the rating task in an accurate and thorough way.

References

Ang-Aw, H. T., & Goh, C. C. M. (2011). Understanding discrepancies in rater judgement on national-level oral examination tasks. *RELC Journal, 42*(1), 31–52. https://doi.org/10.1177/0033688210390226

Alderson, J. C., & Bachman, L. F. (2004). Series editors' preface. In S. Luoma, *Assessing speaking* (pp. ix–xi). Cambridge University Press. https://doi.org/10.1017/cbo9780511733017.001

Alderson, J. C. , & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*(2) 115–129. https://doi.org/10.1093/applin/14.2.115

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41–71). University of Ottawa Press. https://doi.org/10.2307/j.ctt1ckpccf.9

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement*: *Issues and Practice, 31*(3), 2–9. https://doi.org/10.1111/j.1745-3992.2012.00238.x

Borger, L. (2014). *Looking beyond scores. A study of rater orientations and ratings of speaking*. Licentiate thesis. University of Gothenburg. http://hdl.handle.net/2077/38158

Borger, L. (2019). Assessing interactional skills in a paired speaking test: Raters' interpretation of the construct. *Apples – Journal of Applied Language Studies, 13*(1), 151–174. http://dx.doi.org/10.17011/apples/urn.201903011694

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Q*ualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education, 3*, 1–12. https://doi.org/10.3389/feduc.2018.00022/full

Brown, J. D. (2012). *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages*. National Foreign Language Resource Center, University of Hawaii.

Bryman, A. (2008). *Samhällsvetenskapliga metoder* (2. uppl.). Liber.

Byman Frisén, L., Sundqvist, P., & Sandlund, E. (2021). Policy in practice: Teachers' conceptualizations of L2 English oral proficiency as operationalized in high-stakes test assessment. *Languages, 6*(4), 1–23. https://doi.org/10.3390/languages6040204

Bøhn, H. (2015). Assessing spoken EFL without a common rating scale. *SAGE Open, 5*(4), 1–12. https://doi.org/10.1177/2158244015621956

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

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing.

Davis, L. (2018). Analytic, holistic, and primary trait marking scales. In J. I. Liontas (Ed.), *The TESOL encyclopaedia of English language teaching* (pp. 1–6). Wiley. https://doi.org/10.1002/9781118784235.eelt0365

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26*(3), 423-443. https://doi.org/10.1177/0265532209104669

European Commission/EACEA/Eurydice. (2015). *Languages in Secondary Education: An Overview of National Tests in Europe – 2014/15*. Eurydice Report. Publications Office of the European Union. https://data.europa.eu/doi/10.2797/364184

European Education and Culture Executive Agency, Eurydice. (2023). *Key data on teaching languages at school in Europe: 2023 edition*, Publications Office of the European Union. https://data.europa.eu/doi/10.2797/529032

Figueras, N. (2019). Developing and using tasks for the assessment of speaking. *Apples – Journal of Applied Language Studies, 13*(1), 133–149. https://doi.org/10.17011/apples/urn.201903011693

Frisch, M. (2021). Hur betygsätts muntlig språkfärdighet i engelska? En studie av lärares resonemang kring bedömning av det nationella provet för årskurs 9. In C. Bardel, G. Erickson, J. Granfeldt, & C. Rosén (Eds.), *Forskarskolan FRAM*

– *lärare forskar i de främmande språkens didaktik* (pp. 157–175). Stockholm University Press. https://doi.org/10.16993/bbg.h

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132. https://doi.org/10.1080/15434303.2011.642041

Gass, S. M., & Mackey, A. (2016). *Stimulated recall methodology in applied linguistics and L2 research* (2nd ed.). Routledge.

Gustafsson, J. E., & Erickson, G. (2018). Nationella prov i Sverige – tradition, utmaning, förändring. *Acta Didactica Norge*, *12*(4), 1–20. https://doi.org/10.5617/adno.6434

Han, Q. (2016). Rater cognition in L2 speaking assessment: A review of the literature. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *16*(1), 1–24. https://doi.org/10.7916/salt.v16i1.1261

Hasselgren, A. (1997). Oral test subskill scores: What they tell us about raters and pupils. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment - Proceedings of LTRC 96* (pp. 241–246). University of Jyväskylä and University of Tampere.

Heidari, N., Ghanbari, N., & Abbasi, A. (2022). Raters' perceptions of rating scales criteria and its effect on the process and outcome of their rating. *Language Testing in Asia*, *12*, Article 20. https://doi.org/10.1186/s40468-022-00168-3

Horák, T., & Gandini, E. A. M. (2021). Going off-grid: Multiple benefits of rethinking the marking criteria format. *Research Notes, 80*, 7–20, Cambridge Assessment English. https://www.cambridgeenglish.org/Images/630216-research-notes-80.pdf

Hughes, R., & Szczepek Reed, B. (2017). *Teaching and researching speaking* (3rd ed.). Routledge.

Khabbazbashi, N., & Galaczi, E. D. (2020). A comparison of holistic, analytic, and part marking models in speaking assessment. *Language Testing*, *37*(3), 333–360. https://doi.org/10.1177/0265532219898635

Kiger, M.E., & Varpio, L. (2020). Thematic analysis of qualitative data: *AMEE Guide no. 131, Medical Teacher, 42*(8), 846–854. https://doi.org/10.1080/0142159X.2020.1755030

Kvale, S. & Brinkmann, S. (2014). *Den kvalitativa forskningsintervjun* (Third [edited] edition). Studentlitteratur.

Ma, W. (2022). What the analytic versus holistic scoring of international teaching assistants can reveal: Lexical grammar matters. *Language Testing, 39*(2), 239–264. https://doi.org/10.1177/02655322211040020

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

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly, 8*(2), 127–145. https://doi.org/10.1080/15434303.2011.565845

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System, 30*(2), 143–154. https://doi.org/10.1016/s0346-251x(02)00002-7

Pakula, H-M. (2019). Teaching speaking. *Apples – Journal of Applied Language Studies, 13*(1), 95–111. https://doi.org/10.17011/apples/urn.201903011691

Panadero, E., & Jönsson, A. (2020). A critical review of the arguments against the use of rubrics. *Educational Research Review*, *30,* 1–19. https://doi.org/10.1016/j.edurev.2020.100329

Qi, L. (2005). Stakeholders’ conflicting aims undermine the washback function of a high-stakes test. *Language Testing*, *22*(2), 142–173. https://doi.org/10.1191/0265532205lt300oa

QSR International Pty Ltd. (2018). NVivo (Version 12). https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144. https://doi.org/10.1007/bf00117714

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. https://doi.org/10.35360/njes.365

Sandlund, E., & Sundqvist, P. (2019). Doing versus assessing interactional competence. In R. Salaberry & S. Kunitz (Eds.), *Teaching and testing L2* *interactional competence: Bridging theory and practice* (pp. 357–396). Routledge. https://doi.org/10.4324/9781315177021-14

Sandlund, E., & Sundqvist, P. (2021). Rating and reflecting: Displaying rater identities in collegial L2 English oral assessment. In M. R. Salaberry & A. R. Burch (Eds.), *Assessing speaking in context. Expanding the construct and its applications* (pp. 132–162). Multilingual Matters. https://doi.org/10.21832/9781788923828-007

Sato, T. (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing, 29*(2), 223–241. https://doi.org/10.1177/0265532211421162

Seedhouse, P., & Satar, M. (2021). VEO IELTS Project Report: Which specific features of candidate talk do examiners orient to when taking scoring decisions? *IELTS Research Reports Online Series, 5*. British Council, Cambridge Assessment English and IDP: IELTS Australia. https://www.ielts.org/for-researchers/research-reports/online-series-2021-5

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. https://doi.org/10.1177/0265532217690782

Thai, T., & Sheehan, S. (2022). The processes of rating L2 speaking performance using an analytic rating scale – A qualitative exploration. *Language Education and Assessment, 5*(1), 34–51. https://doi.org/10.29140/lea.v5n1.777

University of Gothenburg (2023). *Noteringsunderlag till delprov A. Nationellt prov i engelska för årskurs 9.* https://www.gu.se/nationella-prov-frammande-sprak/prov-och-bedomningsstod-i-engelska/engelska-arskurs-7-9/nationellt-prov-i-engelska-for-arskurs-9

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, *23*(3), 489–508. https://doi.org/10.2307/3586922

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® academic speaking test for operational use. *Language Testing, 24*(2), 251–286. https://doi.org/10.1177/0265532207076365

Youn, S. J., & Chen, S. (2021). Investigating raters' scoring processes and strategies in paired speaking assessment. In M. R. Salaberry & A. R. Burch (Eds.), *Assessing speaking in context. Expanding the construct and its applications* (pp. 107–131). Multilingual Matters. https://doi.org/10.21832/9781788923828-006

Appendix A

Author translation of the assessment guidelines for the National English Test (Year 6) from the Swedish National Agency for Education as they were phrased when data was collected (school year 2019–2020).

OVERVIEW ASSESSMENT PART A

Aim of subject

This part of the test relates first and foremost to four of the long-term goals in the statement of subject aims in the statements of subject aims in the syllabus. Pupils should be given opportunities to develop their ability to

- understand and interpret the content of spoken English
- express themselves and communicate in speech
- use language strategies to understand and make themselves understood
- adapt language for different purposes, recipients and contexts

Assessment of oral production and interaction

Assessment of oral proficiency presupposes that the student, based on the given task and on the syllabus for English, has the willingness and ability to produce and develop topical content, on their own and together with others. The following assessment factors are rooted in the communicative and action-oriented approach to language of the Swedish syllabi for English and Modern Languages. The factors are meant to function as support in the analysis forming a holistic assessment, and should be viewed as different aspects of quality in spoken language.

Assessment factors

Content

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

Language and ability to express oneself

- 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 accuracy
    - Vocabulary, phraseology, idiomaticity
    - Pronunciation and intonation
    - Grammatical structures
- Adaption to purpose, recipient and situation

Grading of part A is first and foremost related to the following parts of the knowledge requirements, which focuses on oral production and interaction in particular.

Knowledge requirements

Grade E	**Grade C**	**Grade A**
In oral … production of different kinds, pupils can express themselves simply and understandably in phrases and sentences. To clarify and vary their communication, pupils can … make some simple improvements to their communications*. In oral…interaction, pupils can express themselves simply and understandably in words, phrases and sentences. In addition, pupils can choose and use a strategy to solve problems and improve their interaction*.	In oral … production of different kinds, pupils can express themselves simply, relatively clearly and to some extent coherently. To clarify and vary their communication, pupils can … make simple improvements to their communications*. In oral…interaction, pupils can express themselves simply and relatively clearly in words, phrases and sentences. In addition, pupils can choose and use some different strategies to solve problems and improve their interaction*.	In oral … production of different kinds, pupils can express themselves simply, relatively clearly and relatively coherently. To clarify and vary their communication, pupils can … make simple improvements to their communications*. In oral…interaction, pupils can express themselves simply and clearly in words, phrases and sentences. In addition, pupils can choose and apply several different strategies to solve problems and improve their interaction*.

* Not in focus specifically, but this part of the national test makes it possible to assess this ability

Appendix B

Interview guide

- **Background questions**
    - What is your teacher degree? From when? What subjects are included in the teacher degree? How many credits do you have from studying English? (Make an estimation of the teacher's age or ask if it's difficult to know)

    - Certified teacher?

    - Years of experience from teaching English?

    - Experience from teaching English for other levels than secondary level? Which ones? How many years?
        - How many times have you been assessing the speaking part of the national test? Do you assess it every year?
        - Experiences from assessing the speaking part of the national test at other levels (for instance year 6, English in upper secondary school)?
        - Experiences from assessing speaking skills in other subjects?

This study is about the speaking part of the national test in English in years 6 and year 9. **The first part of the interview focuses on the material about the national test that you get from the Agency of Education.**

1) *Can you describe how you (and your colleagues) get access to the material at your school?*
    a. *Do all teachers get their own version of the material?*
    b. *What kind of documents do you usually get? Other resources?*
    c. *What kind of documents are included in the material? (Instructions, extra material etc.)*
    d. *Do you get any information from your principal in addition to the material from the Agency of Education regarding instructions on how to administer and assess the national test?*
    e. *Do you find the information that you get to be sufficient?*
        i. *What parts of the information do you particularly appreciate? Why?*
        ii. *According to you, is the information inadequate in any way? How?*

The second part is about your preparations for administering and assessing the speaking part of the national test (how teachers conceptualize and transform the guidelines to something operable)

a. *What is included in your preparations for assessment of the test?*
b. *Do you include the material from the Agency of Education in your preparation of the test? If so; in what ways?*
c. *Do you know what to specifically listen for during the speaking test? Describe!*
d. *Do you feel well prepared before the administration of the test? Why/why not?*

The third part is about the note-taking document that you use during the test.

1. *Did you develop this note-taking document yourself, or have you taken part in developing it? If yes; describe how you did. If no; where did it come from?*

2. *For how long have you been using this note-taking document (estimate a number of times that you have used it for assessment of the speaking part of the national test)*
3. *What are the reasons to why you use this particular note-taking-document?*
4. *Imagine that you are going to assess your students' spoken proficiency in English, you are so to speak in the assessment situation. Describe and tell how you take notes using your note-taking document.*
5. *What would you say is the number one reason to why you use this note-taking document for assessment?*
6. *Are there any other advantages related to using your note-taking document for assessment? If so; what are they?*
7. *Do you experience any disadvantages in using your note-taking document for assessment? If so; what are they?*
8. *Is there anything that the note-taking document fail to acknowledge or miss?*
9. *Is there a need for developing your note-taking document? If so; how?*
10. *If you have used other note-taking documents before, or have made changes to the current one; what are the reasons to why you now use another document/another version of your document?*
11. *Do you use the note-taking document when teaching speaking (i.e., in a formative way)?*

The fourth part is about how you score the test

1. *Can you tell me how you decide what score the student's production represents?*
    a. *What do you have in front of you? Paper, pen, guidelines, note-taking document?*
    b. *How do you use your note-taking document when deciding the score of the speaking part of the national test? Show me by the help of your note-taking document what and how you do.*
    c. *Has it ever, to your recollection, been difficult to reach a decision about the score and if so, what did these difficulties consist of?*
2. *When do you decide the score? Directly after/ In close proximity to the test or later?*
3. *Can you tell me how you reach a decision, is it primarily done individually, or together with colleagues?*
    a. *At your school, do you usually assess your own students in the speaking part of the national test? (Have you ever assessed another teacher's students in this test?)*
    b. *Do you usually decide the score yourself first and then take help of colleagues for the final decision?*
    c. *At your school, are there routines for co-assessment? Describe!*
    d. *If you co-assess – how do you together decide the score? How is the note-taking document used in that process?*

4. *The Agency for Education published a revised version of the general guidelines for grading in which a discussion of scoring rubrics for assessment is included – have you changed anything in the way that you score or assess in relation to these guidelines?*

    a. *Have you for instance made any changes to the note-taking document that you use?*

5. *What assessment criteria are important for the score you award?*

    a. *What signifies a student production on a very high level (the score A or higher)?*

    b. *What signifies a student production on a very low level (beneath the score E)?*

    (or: Describe the differences between a production on the E-, C- and A-level. With this question I hope to learn whether any specific criteria are deemed more important than others).

6. *Lastly, I would like for you to fill in this very short questionnaire (hand out on paper).*