languages

MDPI

Article

Policy in Practice: Teachers' Conceptualizations of L2 English Oral Proficiency as Operationalized in High-Stakes Test Assessment

Liliann Byman Frisén [1,*], Pia Sundqvist [2] and Erica Sandlund [1]

[1] Department of Language, Literature and Intercultural Studies, Faculty of Arts and Social Sciences, Karlstad University, 65188 Karlstad, Sweden; erica.sandlund@kau.se
[2] Department of Teacher Education and School Research, Faculty of Educational Sciences, University of Oslo, 0317 Oslo, Norway; pia.sundqvist@ils.uio.no
* Correspondence: liliann.frisen@kau.se

Abstract: Assessment of foreign/second language (L2) oral proficiency is known to be complex and influenced by the local context. In Sweden, extensive assessment guidelines for the National English Speaking Test (NEST) are offered to teachers, who act as raters of their own students' performances on this high-stakes L2 English oral proficiency (OP) test. Despite guidelines, teachers commonly construct *their own* NEST scoring rubric. The present study aims to unveil teachers-as-raters' conceptualizations, as these emerge from the self-made scoring rubrics, and possible transformations of policy. Data consist of 20 teacher-generated scoring rubrics used for assessing NEST (years 6 and 9). Rubrics were collected via personal networks and online teacher membership groups. Employing content analysis, data were analysed qualitatively to examine (i) what OP sub-skills were in focus for assessment, (ii) how sub-skills were conceptualized, and (iii) scoring rubric design. Results showed that the content and design of rubrics were heavily influenced by the official assessment guidelines, which led to broad consensus about *what* to assess—but not about *how* to assess. Lack of consensus was particularly salient for interactive skills. Analysis of policy transformations revealed that teachers' self-made templates, in fact, lead to an analytic rather than a holistic assessment practice.

Keywords: language assessment; oral proficiency; interaction strategies; English as a second/foreign language; holistic assessment; analytic assessment; scoring rubrics; high-stakes testing; teachers-as-raters; ATD

Citation: Byman Frisén, Liliann, Pia Sundqvist, and Erica Sandlund. 2021. Policy in Practice: Teachers' Conceptualizations of L2 English Oral Proficiency as Operationalized in High-Stakes Test Assessment. *Languages* 6: 204. https://doi.org/10.3390/languages6040204

Academic Editors: Dina Tsagari and Henrik Bøhn

Received: 20 September 2021
Accepted: 24 November 2021
Published: 9 December 2021

1. Introduction

Oral proficiency in a second and/or foreign language (L2) is "at the very heart of what it means to be able to use a foreign language" (Alderson and Bachman 2004, p. ix), but it is also the language skill that is the most difficult to assess in a reliable way (ibid.). One of the challenges raters of L2 oral proficiency face is the fact that numerous aspects of quality need to be considered simultaneously by raters (Bachman 1990). Furthermore, as assessment of L2 oral proficiency (henceforth OP) is generally tested in different social interactional formats, such as paired/small group conversations or interviews with an examiner, standardizing testing conditions for learners' L2 OP is particularly challenging (Bachman 2007). Likewise, as convincingly demonstrated in the literature, the indisputable co-constructedness of the product for *assessment*, the actual interaction in L2 oral tests (see e.g., Young and He 1998; May 2009), entails a realization that each interaction is unique in terms of context and co-participants and that, as such, generalizations regarding individual proficiency are virtually impossible. Adding to these difficulties, raters will inevitably make different professional judgments of the same candidate's performance in an interaction-based test, to the extent that some scholars have

argued that "aiming for perfect agreement among raters in paired speaking assessment is not feasible" (Youn and Chen 2021, p. 123). However, a strive for equity and fair grading in high-stakes contexts is, and should be, of central concern to stakeholders in language testing and assessment.

While many well-known large-scale OP tests rely on trained examiners for their assessment (e.g., the Test of English as a Foreign Language (TOEFL) and the Cambridge ESOL Examinations, including the International English Language Testing System (IELTS)), some national assessment systems use teachers for assessing their students' performance (Sundqvist et al. 2018). In New Zealand, an assessment reform—*interact*—has been implemented in which teachers are required to collect several instances of interaction output from students for grading purposes, using guidelines for the assessment (East 2015). Norway implemented a system where teachers could choose between assessing their students' language proficiency (OP included) in a testing or classroom context (Hasselgren 2000). In Sweden, which constitutes the empirical case of the present study, the National Test of English is compulsory in years 6 and 9, and teachers both administer and assess the test of OP with assessment guidelines provided by the Swedish National Agency for Education (SNAE) (Borger 2019; University of Gothenburg 2021). While the setup has many advantages, voices have also been raised about the role of teachers as examiners for achieving equity and assessment standardization (Sundqvist et al. 2018). Furthermore, as studies focusing explicitly on uncovering the rating process for speaking assessment have demonstrated, raters differ in both their understandings of assessment criteria and in their application of those understandings to authentic speech/interaction samples (e.g., Borger 2019; Ducasse and Brown 2009; May 2011; Sandlund and Sundqvist 2019, 2021). In line with such observations, Youn and Chen (2021) call for additional research on the rating process as such. Similarly, Ducasse and Brown (2009) and Sandlund and Sundqvist (2019) emphasize the need for empirical work uncovering how raters orient to interaction as they conceptualize assessment. In response to these calls, the present study examines how teachers, as raters of L2 OP in the Swedish national test of English in compulsory school, conceptualize and "do" assessment of L2 OP. In contrast to studies examining post-assessment rating protocols, stimulated recall, rater interviews (May 2009, 2011; Youn and Chen 2021), or recorded rater discussions (Sandlund and Greer 2020; Sandlund and Sundqvist 2019, 2021), we target scoring rubrics created by raters themselves for notetaking during actual test situations. The study thus contributes to filling a research gap in terms of uncovering raters' conceptualizations of the construct at hand as they prepare and organize their actual assessment work.

In Sweden, students in school years 6 (ages 12–13) and 9 (ages 15–16) take a high-stakes, summative, three-part proficiency test in English, whose aim is to support equity in assessment and the grading of students' knowledge and skills (Swedish National Agency for Education 2021). Our focus is on part A (Speaking)—the National English Speaking Test (NEST), where L2 English OP is tested by the test construct *oral production and interaction*. Students are divided into pairs or small groups and instructed to discuss pre-set topics from the test material with their peers. As mentioned, the students' own teacher also administers and assesses the NEST (Sundqvist et al. 2018). Teachers do not receive specific rater training; instead, they are provided with extensive rater guidelines and benchmark examples, which they are instructed to review before operationalizing assessment of the test. In the present article, the term *teacher-as-rater* is used to reflect the dual role of the teachers, as they serve as raters/examiners of the NEST while, at the same time, they are the test-takers' own teacher. Despite extensive rater guidelines and benchmark examples from SNAE, some teachers construct their own scoring rubric that they use when assessing the NEST. As teachers-as-raters are faced with the task of assessing a standardized test of L2 English OP in a fair, valid, and reliable manner, while also taking into consideration different local conditions, examining their self-made scoring rubrics might reveal how teachers-as-raters adapt to this particular assessment situation and, consequently, whether alterations of the test construct are made when operationalizing assessment in this particular context. An

issue arising from this research problem is how assessment policy (in this case, SNAE's holistic view) is conceptualized and operationalized in scoring rubrics created by teachers-as-raters. Thus, the aim of the study is to unveil teachers-as-raters' conceptualizations of the NEST construct as these emerge from teachers' self-made scoring rubrics, to examine whether, and possibly how, policy is *transformed* (Chevallard 2007) in the process.

The following research questions (RQs) guided the study:

- RQ1: What sub-skills of oral proficiency are in focus for assessment in teachers' own scoring rubrics, and in what ways is transformation of the test construct possibly reflected in sub-skills chosen for assessment?
- RQ2: How are oral proficiency sub-skills to be assessed organized in teachers' own scoring rubrics? In what ways are teachers' conceptualizations of the test construct reflected in this organization?
- RQ3: What similarities and differences are there between conceptualizations as reflected in the scoring rubrics when it comes to sub-skills to be in focus for assessment?

The present case study, in a high-stakes context in Sweden, can shed light on teachers-as-raters' conceptualizations of the process of assessing L2 OP, as well as on whether raters in other settings conceptualize assessment of L2 OP in a similar way, despite its multi-faceted and context-dependent nature. Furthermore, the study aims to provide insights into whether the use of rubrics for assessment of L2 OP affects conceptualizations and, thus, knowledge on what stands out as salient to teachers-as-raters as they operationalize policy for assessment. These insights, in turn, can inform test constructors and contribute to construct development as well as rater training efforts.

1.1. Assessing L2 Oral Proficiency

For the assessment of "complex" performances, such as free-written or spoken production, the "individualized uniqueness and complexity" (Wang 2010, p. 108) of the tasks and performances to be assessed present a challenge in terms of ensuring that raters understand and apply rating scales in the same way. As Papajohn (2002) observes, raters may arrive at the same score on the same learner response for different reasons, and, with such inherently subjective scoring of complex language abilities, disagreement between raters is to be expected (Meadows and Billington 2005). Additional challenges with test constructs for linguistic and interactional skills include the deeply collaborative nature of conversations between individuals (Sert 2019), as raters are forced to attempt to separate the performances of test-takers, although they are, in fact, inter-dependent.

Traditionally, the assessment of L2 speaking performance has been based on several core components making up L2 proficiency, such as the CAF framework of *complexity*, *accuracy*, and *fluency* (Skehan and Foster 1999; Housen et al. 2012), or divided up into further analytic criteria such as *fluency*, *appropriateness*, *pronunciation*, *control of formal resources of grammar*, and *vocabulary* (McNamara 2000). However, whether raters are asked to assess these different components separately (analytic rating) or to consider all in assessing a single impression of performance (holistic rating, see McNamara 2000, p. 43) varies between different tests. While there is agreement in testing research that OP is multi-faceted, the jury is still out on exactly which components should be weighed in and relative importance for proficiency measurements of these components. The communicative movement in language teaching and testing (Canale and Swain 1980; Bachman 1990) brought about an increased focus on learners' use of language for communicative purposes, and more recently, the framework of interactional competence (Kramsch 1986; Young and He 1998; Kasper and Ross 2013; Salaberry and Kunitz 2019; Salaberry and Burch 2021) has also made its way into speaking tests, and many speaking tests now include interactional abilities along with proficiency in their constructs.

Needless to say, as most speaking tests are conducted in the form of interaction between either an interviewer and a candidate (*oral proficiency interviews*, OPIs) or in pairs or small groups (see Sandlund et al. 2016; or Ducasse and Brown 2009 for an overview), raters of L2 speaking tests face a challenging task when considering multiple aspects of proficiency

and interaction when assessing an individual speaking performance. As McNamara (1996) noted, judgments of complex language skills "will inevitably be complex and involve acts of interpretation on the part of the rater and thus be subject to disagreement" (p. 117). Ducasse and Brown (2009) point out that studies of raters' perceptions of L2 test interactions are essential in understanding how assessment plays out in practice, "because it is their view of interaction which finds its reflection in the test scores" (Ducasse and Brown 2009, p. 425).

1.2. Raters' Conceptualizations of L2 OP Assessment

As the act of speaking an L2 involves not only purely linguistic but also pragmatic competence, raters have at their disposal a range of aspects of learner performances to be weighed against criteria or which contribute to an overall impression of learner's L2 linguistic and interactional skills. Empirical studies of raters' understandings of scoring rubrics have emphasized this complexity for assessment. Ang-Aw and Goh (2011), for instance, show that raters have conflicting ideas about the importance of the various criteria in the rating scales and that aspects of learner performances both within and at the outskirts of the scoring rubrics may be weighed into assessment decisions, such as effort (Ang-Aw and Goh 2011). Other studies demonstrate how raters of oral L2 tests pay attention to different aspects of performance (see, e.g., Orr 2002) and that raters focus on different aspects of performance depending on what level the test-takers are at (Sato 2012). As such, the salience of particular aspects may vary—between raters but also in their application to test-taker proficiency levels.

In a study of raters' perceptions of interaction in tests using test discourse, written rater protocols, rater discussions, and stimulated recall, May (2011) demonstrates that many features of L2 interaction that were salient to raters fell under the scope of what test-takers co-construct in a test, such as understanding and responding to the interlocutor's message, cooperating interactionally, and contributing to the quality and perceived authenticity of the interaction. The issue of co-construction and assessment was also raised by May (2009), who points out that, given raters' apparent struggles in separating the individual performances of interlocutors when scoring, "it seems counter-intuitive to ask raters to award separate marks to candidates for interaction" (ibid., p. 417). Relatedly, both May (2011) and Sandlund and Sundqvist (2016) note that raters sometimes tend to compare test-takers against each other rather than against criteria. In a Swedish high-stakes L2 English testing context, Borger (2019) examined rater perceptions of interactional competence based on raters' NEST scoring and notes about test-takers. Three aspects of interaction seemed to stand out to raters in their assessment: topic development moves, turn-taking management, and interactive listening strategies. Raters also considered test-takers' interactional roles. Further, Borger observes that the raters attended to features both within and beyond the assessment guidelines, indicating that guidelines for raters "have to be elaborated, including conceptually grounded reasoning as well as commented examples" (2019, p. 167).

In most research studies, raters use a pre-defined set of criteria, or scoring rubric, from which they draw their conclusions about what, or what not, to include in assessment of OP. However, there are also examples of research on the assessment of oral proficiency where no common rating scale is used (Bøhn 2015). Results from this study show that raters generally have similar thoughts regarding what aspects to include in the assessment of OP but that they differ when it comes to the relative importance of these aspects and results that are in line with studies where common rating scales are being used.

In Sweden, an examination of raters' orientations to L2 English oral tests shows that raters using national performance standards and raters using the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2018) seem to understand and interpret the categories to be assessed in a similar way, and therefore a broad level of agreement regarding the test construct *oral production and interaction* seems to be reached (Borger 2014). However, another study in the area (Frisch 2015) indicates that teachers have different orientations toward OP and that differences in orien-

tations stem from "which criteria from the national test guidelines they mostly referred to" (Frisch 2015, p. 102). In other words, there seems to be broad consensus among Swedish teachers about what categories assessment of L2 OP should encompass; however, there also seem to be conflicting ideas about what is stated about such assessment in the national test guidelines and, as a consequence, what to focus on when assessing these categories. The present study aims at generating new knowledge when it comes to how teachers-as-raters conceptualize the categories to assess, as well as whether there are differences or similarities between their conceptualizations. In relation to Frisch's study, an interesting question is whether teachers-as-raters' conceptualizations, as reflected in their scoring templates, mirror different views of what is stated about assessment of the test construct in the NEST guidelines.

In summary, raters face a tough task when assessing L2 OP, and there is often a "lack of meaningful descriptors" for guidance, particularly with asymmetric dyads (May 2009, p. 416). Adding to previous research on rater perceptions, this study focuses on some ways in which raters deal with such complexity. This study also aims to add to previous research on how raters' conceptualizations of assessment are possibly affected by the actual practice of carrying out assessment. It contributes as such to filling a knowledge gap identified by Tsagari (2021, p. 27): "there is a lot yet to be learnt about the protagonists of assessment—students and teachers, and how they enact assessment policy mandates in their daily practices".

1.3. Scoring Rubrics

A *rubric* is generally viewed as an assessment tool that consists of two parts: criteria in focus for assessment and descriptive text for different performance levels for each criterion (Brookhart 2018; Brown 2012; Sadler 2009). The descriptive text typically consists of a "qualitative description of the corresponding 'standard', often with reference to sub-attributes of the main criterion" (Sadler 2009, p. 163). *Rubrics* are often confused with other rubric-like instruments such as *checklists* and *rating scales* that also list criteria. Although definitions of *checklists* and *rating scales* vary somewhat in the literature (e.g., Brookhart 2018; Brown 2012), a common denominator is that they, unlike *rubrics*, contain little descriptive text to inform the raters what varying qualities of each criterion are. Instead, a *checklist* typically consists of criteria that target details of a performance, where raters make rough estimations whether the criteria listed are absent/present or rate the performance according to a scale that indicates the quality of the feature listed (e.g., *ok*, *good*, *great*). A *rating scale* lists criteria (although they are usually fewer compared with those in checklists) and also ask raters to assess the performance along a scale, but "characteristics of performance at each level are not described" (Brown 2012, p. 40).

Rubrics can be categorized as *analytical* or *holistic*, where the former consists of criteria that are considered one at a time during assessment and the latter of all criteria considered simultaneously (Brookhart 2018; Brown 2012). The content of an analytic rubric and a holistic rubric might be identical; what distinguishes the two is mainly the format, something that might fundamentally change both the purpose and utility of the rubric used (Brown 2012). Therefore, the choice between using one or the other depends on the assessment situation (Davis 2018), as they are "representations of what is considered important in performance" (Khabbazbashi and Galaczi 2020, p. 336). Each has its strengths and weaknesses. Since analytic rubrics list different criteria of a multi-dimensional performance, using them for assessment can give an indication of test-takers' potentially jagged performance profiles (Davis 2018; Khabbazbashi and Galaczi 2020), and analytic rubrics are therefore beneficial for formative feedback purposes (Jönsson and Svingby 2007). The main advantage of a holistic rubric is that it is practical: only one grading decision needs to be made (Brookhart 2018; Brown 2012; Davis 2018), which makes it less cognitively demanding for raters (Xi 2007). Holistic rubrics are often used in large-scale assessment and standardized testing (Brown 2012; Jönsson and Svingby 2007).

Particularly in relation to assessing written and spoken proficiency, holistic and analytic scoring seem to be widely used human-mediated marking methods, and their respective strengths and weaknesses have been both documented and discussed (Khabbazbashi and Galaczi 2020). In comparison, analytic scoring rubrics allow for a more systematic assessment process where qualities are made explicit. Further, raters are given a clearer picture of the focus for assessment, which may improve reliability (ibid.). However, both Lee et al. (2010) and Xi (2007) found a clear correlation between analytic and holistic scores and concluded that analytic scoring might be psychometrically redundant. Moreover, since analytic assessment is cognitively more challenging, Xi (2007) argues that rigorous rater training is necessary if raters are to reliably distinguish between criteria listed in an analytic scoring method.

According to Khabbazbashi and Galaczi (2020), choice of marking method has been shown to affect grades awarded, but there is little empirical research on how different marking methods compare. In their study (Khabbazbashi and Galaczi 2020), they examined the impact of holistic, analytic, and part marking models on measurement properties and CEFR classifications in a speaking test. Although there were strong correlations between the three different marking models, the choice of model impacted significantly on the CEFR levels awarded to candidates, with half of the candidates being awarded different levels depending on the scoring model used. The authors conclude that, of the three models examined, the part marking model had superior measurement qualities.

In relation to rubrics for assessment of L2 spoken interaction, Ducasse (2010) reports on a study of scale development based on raters' perceptions of salient interactional features. Scale development was thus empirically developed by raters following a method for developing task-specific rubrics. However, although rubrics are extensively used as guides for scoring (Lindberg and Hirsh 2015), no studies have to our knowledge examined what is included or excluded from assessment in teacher-generated scoring rubrics and in what ways content is organized. The present study aims to contribute to filling this gap of knowledge, particularly when it comes to what *content* teachers decide to include in an L2 OP assessment rubric. Since *content* is a representation of what is of importance in the assessment situation (Khabbazbashi and Galaczi 2020) and since organization of content can greatly affect the outcome of scoring (Brown 2012; Davis 2018; Khabbazbashi and Galaczi 2020), studying teacher-generated rubrics used for standardized testing of L2 OP can shed light on *what* is being assessed as well as *how*. We use the unifying term *scoring templates* for all the teacher-generated assessment documents analysed, including documents where performance level descriptors are lacking. We categorize the data by applying Brookhart's (2018) definitions of *rubric*, *rating scale*, and *checklist*. She makes a distinction between the three different scoring instruments as follows: "A *rubric* articulates expectations for student work by listing criteria for the work and performance level descriptions across a continuum of quality [. . .]. *Checklists* ask for dichotomous decisions (typically has/doesn't have or yes/no) for each criterion. *Rating scales* ask for decisions across a scale that does not describe the performance" (Brookhart 2018, p. 1).

1.4. Transformations of Subject Content: The ATD Framework

As we are interested in possible transformations of the construct of *oral production and interaction* in teachers' operationalization of policy, we draw on the theoretical framework Anthropological Theory of Didactics (ATD) (Chevallard 2007) and the *didactic transpositions* that the content to be taught and learnt is subject to. According to ATD, there is a dialectic relationship between institutions and the people within these institutions in which content is co-determined on a hierarchy of levels (Achiam and Marandino 2014). At the core of didactic transposition theory lies the assumption that knowledge is a changing reality, as it is affected and formed by the conditions prevalent in the institution in which knowledge is taught and learnt. Thus, knowledge is adapted to the institution it exists within, and in the ATD framework, this is called *the institutional relativity of knowledge* (Bosch and Gascón 2014).

Figure 1 shows how knowledge, according to ATD, undergoes a transformation process in order to become (institutionally) operable. Knowledge generated in academia thus needs to be transformed by stakeholders and authorities (for instance, the SNAE) to become policy (for instance, assessment guidelines for the NEST). Policy, in turn, needs to be interpreted by stakeholders at the institutional level and transformed to become operable (for instance, into teaching material). Finally, each student's knowledge will be formed by the teaching and assessment activities they are being subjected to and partake in. Here, the ATD framework is used to explain potential transformations of the test construct *oral production and interaction* as it emerges when influenced and formed by operationalization.

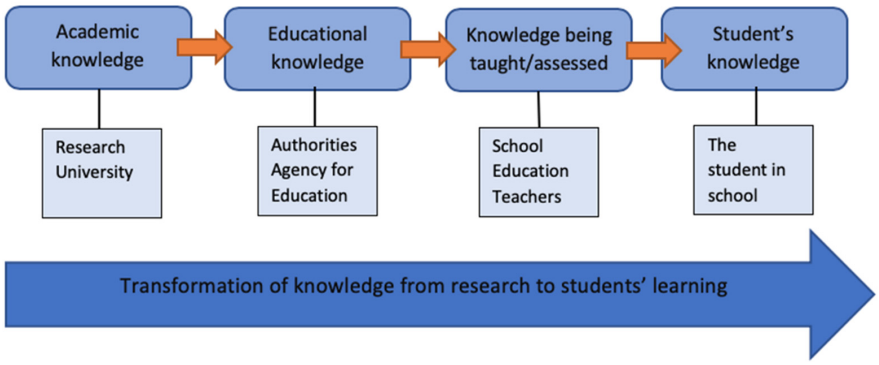

Figure 1. Our illustration of Anthropological Theory of Didactics (Chevallard 2007).

The idea of *praxeologies* is part of the ATD framework. Praxeologies are entities consisting of *praxis* and *logos*. Both these parts need to be taken into account "in order to explain the fate of 'true' knowledge" (Chevallard 2007, p. 133). *Praxis* consists of a type of task and a technique for carrying out the task, whereas *logos* consists of technology (i.e., the discourse of the technique, such as why a particular technique is beneficial for carrying out a task) and a theory supporting the use of technology. Following this idea, the scoring templates generated by teachers-as-raters can be seen as the *technique* supporting the *task* of assessing L2 OP. Teachers-as-raters' reasons for creating the scoring templates as well as the theory supporting their use are viewed as the *logos* of the praxeology. Data generated in the present study do not allow us to analyse the *logos* behind the use of scoring templates for the assessment of L2 English OP; however, examining the *technique*, and conceptualizations of the test construct as reflected therein, should reveal teachers-as-raters' *relative institutional knowledge* of *oral production and interaction*. This *knowledge*, generated at the "school level" in Figure 1, is compared and contrasted to *knowledge* of L2 English OP generated at the "authority" level (which is reflected in the assessment guidelines for the test). By comparing knowledge generated at these two different levels, transformations of the test construct when using scoring templates for assessment can emerge.

2. Materials and Methods

2.1. Setting the Scene: Assessment of L2 English Oral Proficiency in Sweden

As previously mentioned, Swedish teachers-as-raters are provided with extensive guidelines for the operationalization of the NEST and the test construct *oral production and interaction*. The guidelines include instructions to teachers about test organization (e.g., divide students into pairs), administration (e.g., remain in the background and let the students do the talking), and preparation (e.g., listen to accompanying NEST sample recordings). The pilot students' productions are commented on and holistically assessed (in the form of a grade) by SNAE's expert raters in the form of benchmark examples. As guidelines as well as benchmark examples are extensive, they are reviewed by teachers-as-raters when preparing for assessment of the NEST or consulted after assessment is done. In addition, teachers-as-raters are provided with a one-page assessment document (see Supplementary Materials S1) beneficial for use during the test situation. Teachers are also advised to copy this document and distribute it to students taking the test. This document states that

assessment of students' oral production and interaction should be made holistically, taking *all* aspects of students' performance into account (University of Gothenburg 2021). Besides qualitative descriptions of three grade levels that the holistic assessment should be based on (the so-called *knowledge requirements* for passing grades E, C, and A), some factors are also listed in this assessment document ("analytic assessment factors", Borger 2019). These include linguistic qualities (e.g., *grammatical structures and vocabulary*), ability to produce content (e.g., *giving examples and providing different perspectives*), and ability to interact (e.g., *adaptation to recipient and situation*) (University of Gothenburg 2021). There are no explicit instructions about what students should master in terms of, for example, grammar, vocabulary, or strategies, nor about how to summarize or relate aspects to the holistic assessment. However, teachers-as-raters might consult the benchmark examples to read expert raters' considerations of these factors when commenting on the pilot students' performances.

2.2. Data

Data consist of scoring templates used for assessing NEST in grades 6 and 9. To be included in the study, scoring templates (STs) had to be (i) designed/created by teachers-as-raters, (ii) used for assessment of NEST 6 or NEST 9, and (iii) unique (that is, not identical to any other collected ST). ST data generally corresponded to a one-page document used for notetaking and assessment during the actual test situation.

2.3. Data Collection

Data were collected via our professional networks and two closed teacher groups on Facebook: (i) English grades 4–6 (with 4406 members) and (ii) English grades 6–9 (4394 members). Those interested were asked to submit material used when documenting assessment of the NEST (i.e., their STs). In total, 28 STs were obtained (6 via networks, 22 via Facebook), of which 20 fulfilled the criteria for inclusion (8 for year 6, 12 for year 9). These 20 STs turned out to be detailed and sufficed to answer our research questions. Each ST was coded Y6 or Y9, plus a number (e.g., Y601 and Y912).

2.4. Data Analysis

The method of analysis was qualitative, complemented with quantitative frequency counts of categories (see below), following a summative approach to qualitative content analysis (Hsieh and Shannon 2005). Content analysis was guided by the terms CONSTRUCT, CRITERION, and SUB-CRITERION, following Bøhn (2015). In his analysis of interview data where Norwegian teachers' conceptualizations of L2 English OP were studied, Bøhn (2015) categorizes features that teachers reported to include in assessment in a hierarchical structure. CONSTRUCT was used for broader categories of concepts and CRITERION and SUB-CRITERION for more narrowly defined aspects of performance. We adopted a similar hierarchical structure to categorize the content of the STs; however, we used CONSTRUCT to describe the language ability specifically stated in the assessment guidelines by SNAE to be in focus for the test (i.e., the test construct *oral production and interaction*). We reserved CRITERION and ASPECT for broader categories, while SUB-ASPECT, SUB-SUB-ASPECTS, etc. were used for narrowly defined aspects of performance. We used CRITERION for features in focus for assessment aligns with its use in literature about scoring rubrics (Brookhart 2018; Brown 2012; Sadler 2009). Content analysis was conducted in four analytical steps.

2.4.1. Step One

Step One comprised classifying the sub-skills of OP listed in each ST using a coding scheme in which the categories CRITERION, ASPECT, SUB-ASPECT, SUB-SUB-ASPECT, and so forth were used. All STs listed criteria that are in focus for assessment of the test construct, and these criteria were categorized as CRITERION in the coding scheme, (for example, *Content*). In several STs, descriptive texts inform what standards characterize the criterion in question, and/or what to look/listen for in relation to that particular criterion when

assessing. These standards and descriptions were labelled ASPECT (for example, *different perspectives and examples*). Additionally, given descriptions of ASPECTS were labelled SUB-ASPECTS, and yet further descriptions SUB-SUB-ASPECTS, et cetera. This procedure resulted in a tree diagram for each ST representing its hierarchical structure as well as how many CRITERION, ASPECTS, and so forth each ST encompassed (see Figure 2).

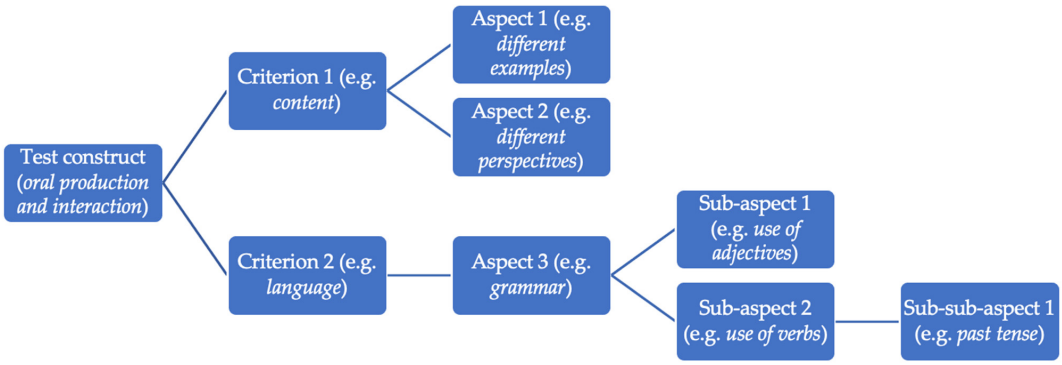

Figure 2. Tree Structure of Categories Criterion, Aspect, Sub-aspect, and Sub-sub-aspect.

2.4.2. Coding Reliability

In order to evaluate intra-rater reliability, two STs (Y608 and Y906) were re-classified six months after Step One was completed. For the purpose of inter-rater reliability, a colleague with long experience of assessing NEST coded the same two STs using the same coding scheme along with an instruction to count how many CRITERION, ASPECTS, SUB-ASPECTS, etc. she could identify. The results from the three different classifications of oral sub-skills into categories (original coding, recoding, and colleague's coding) are presented in Table 1 (Y608) and Table 2 (Y906).

Table 1. Inter-rater reliability: Number of codes per classification of oral sub-skills into categories for ST Y608.

Version of Coding	CRITERION	ASPECT	SUB-ASPECT	SUB-SUB-ASPECT	SUB-SUB-SUB-ASPECT
Original coding	2	8	16	12	1
Re-coding	2	8	16	9	1
Colleague's coding	2	8	16	14	2

Table 2. Inter-rater reliability: Number of codes per classification of oral sub-skills into categories for ST Y906.

Version of Coding	CRITERION	ASPECT	SUB-ASPECT	SUB-SUB-ASPECT	SUB-SUB-SUB-ASPECT
Original coding	7	26	4	0	0
Re-coding	7	23	6	0	0
Colleague's coding	7	28	2	0	0

As the colleague was only instructed to provide numbers for each category—CRITERION, ASPECT, and so forth—the data do not allow us to analyse possible reasons for discrepancies between her coding and the original one. However, since tree structures were drawn in both the original coding and the recoding of ST Y608, analysis of the differences between these two versions yielded some answers. Discrepancies stemmed from different interpretations regarding whether descriptive texts put forth one or several SUB-SUB-ASPECTS. An example is *vocabulary*, described in three quality levels according to the ST: grade E (use of *standard vocabulary*), grade C (*extended vocabulary*), and grade A (*large vocabulary*) (our translations from Swedish). In the original coding, these were seen as three different aspects, while it was interpreted as one aspect (*vocabulary, varying degrees of*) in the recoded version.

Analysis of discrepancies between the original coding and re-coding showed that, also for ST Y906, it was difficult to determine whether descriptive texts contained one or several ASPECTS and SUB-ASPECTS (see Table 2). However, there was an additional problem

causing discrepancies between the different versions of the coding due to difficulties in interpreting whether a factor listed in the ST was an ASPECT or a SUB-ASPECT. For example, in ST Y906, the way the student uses *linking words* was an ASPECT listed for the grades E and C (grade E: "student uses some linking words", grade C: "student uses several linking words"), but for the grade A, it said, "student uses a bigger repertoire of connecting phrases", making it difficult to determine whether "connecting phrases" was an additional ASPECT or an example of "linking words" (and thus a SUB-ASPECT of the ASPECT *use of linking words*).

An analysis of the three different versions of coding revealed that a general coding scheme was difficult to apply on all STs, since STs differed in design and were most likely used differently by teachers-as-raters. For instance, in Y608, the content was relatively well-structured and thus easier to code than ST Y906, which consisted of extensive coherent descriptive text leaving more room for interpretation. However, both inter-rater and intra-rater were reliability deemed satisfactory since the number of identified features rendering a code for CRITERION, and, for the most part, also for ASPECT, were identical across all three coding instances. As descriptions became more detailed (as was the case for SUB-ASPECTS, etc.), discrepancies between coders were found, which is to be expected. In the subsequent analytical work, it was considered important to regard the quantifiable number of CRITERION, ASPECTS, SUB-ASPECTS, etc. for each ST only as a means for sorting data for the subsequent content analysis, in line with suggestions by Hsieh and Shannon (2005).

Since a clear-cut line between the different categories could not be made, we decided to continue the analytical process by regarding the data as consisting of *two types of categories*: one type that described more OVER-ARCHING COMPETENCIES or abilities where proficiency was described in a broad, generic way (e.g., *language, content*) and the other that described more NARROWLY DEFINED COMPONENTS of language use or tokens of understanding (e.g., *asks questions, agrees with conversation partner*). The categories CRITERION and ASPECT belong to the first type (OVER-ARCHING COMPETENCIES), and it was deemed suitable to analyse the data further by applying content analysis to these two categories (i.e., CRITERION and ASPECT for Y6 STs as well as CRITERION and ASPECT for Y9 STs) in order to see what themes emerged.

2.4.3. Step Two

Step Two in the data analysis was to list all CRITERIA/ASPECTS/SUB-ASPECTS, etc. used to describe assessment of the test construct. One list was drawn for the categories inherent in teachers-as-raters' STs for year 6 and a similar list for year 9.

The computer software program NVivo12 was used to conduct content analysis of the two categories CRITERIA and ASPECT. A frequency count of the occurrence of the same words or phrasings (such as *vocabulary* or *comprehension*) was used to detect key thoughts or concepts (Hsieh and Shannon 2005), and as such, themes for STs from Y6 and Y9 emerged (see Results). After comparing these themes, six themes were selected for further analysis mainly based on the frequency counts. (See Results for considerations made when establishing the six themes).

2.4.4. Step Three

After having established the six themes, we also applied content analysis on the more narrowly defined categories (SUB-ASPECTS/SUB-SUB-ASPECTS, etc.) in order to study how the themes were exemplified and characterized. Content analysis was also applied to compare similarities and differences between conceptualizations made for each theme for Y6 and Y9.

2.4.5. Step Four

The last analytical step, Step Four, included content analysis to categorize the design of the STs into *rubrics*, *rating scales*, and *checklists* following Brookhart's (2018) definitions. STs where standards for each criterion were described were categorized as a *rubric*, whereas

STs that listed CRITERION/ASPECTS/SUB-ASPECTS, etc. to be assessed but did not describe what to look and/or listen for when it comes to these were categorized as *checklists*. STs categorized as *rating scales* would include one scale per criterion along which raters graded students' performances for that particular criterion.

3. Results

3.1. RQ1: What Sub-Skills of Oral Proficiency Are in Focus for Assessment in Teachers' Own Scoring Rubrics, and in What Ways Is Transformation of the Test Construct Possibly Reflected in Sub-Skills Chosen for Assessment?

The content analysis resulted in nine themes for Y6 STs and thirteen themes for Y9 STs (see Tables 3 and 4, which also include the frequency counts). As themes emerged from concepts used in the STs, theme phrasings are similar to ST phrasings.

Table 3. Themes STs Y6—Number of times mentioned in categories CRITERION and ASPECT.

Theme	*N* (Percentage)
Adaptation to purpose, recipient, and situation (in Swedish: anpassning till syfte, mottagare och situation)	14 (17.9)
Comprehensibility and clarity (in Swedish: begriplighet och tydlighet)	15 (19.2)
Strategies/communicative strategies (in Swedish: strategier/kommunikativa strategier)	10 (12.8)
Richness and variation (in Swedish: fyllighet och variation)	6 (7.7)
Context and structure (in Swedish: sammanhang och struktur)	10 (12.8)
Grammatical structures (in Swedish: grammatiska strukturer)	7 (9.0)
Pronunciation and intonation (in Swedish: uttal och intonation)	6 (7.7)
Language/Language and ability to express oneself (in Swedish: språk/språk och uttrycksförmåga)	5 (6.4)
Content (in Swedish: innehåll)	5 (6.4)
Total	**78 (100)**

Note. All theme phrasings are identical to phrasings in the SNAE guidelines.

Table 4. Themes STs Y9—Number of times mentioned in categories CRITERION and ASPECT.

Theme	*N* (Percentage)
Adaptation to purpose, recipient, and situation * (in Swedish: anpassning till syfte, mottagare och situation)	17 (11.3)
Comprehensibility and clarity * (including **intelligible** and **comprehensible** (in Swedish: begriplighet och tydlighet)	15 (9.9)
Strategies/communicative strategies * (in Swedish: strategier/kommunikativa strategier)	16 (10.6)
Richness and variation * (in Swedish: fyllighet och variation)	21 (13.9)
Context and structure * (in Swedish: sammanhang och struktur)	11 (7.3)
Grammar (in Swedish: grammatik)	8 (5.3)
Pronunciation and intonation * (in Swedish: uttal och intonation)	10 (6.6)
Language/Language and ability to express oneself * (in Swedish: språk/språk och uttrycksförmåga)	9 (6.0)
Content * (in Swedish: innehåll)	12 (7.9)
Engagement/initiative (in Swedish: engagement/initiativ)	8 (5.3)
Fluency * (in Swedish: flyt)	13 (8.6)
Vocabulary * (in Swedish: ordförråd)	7 (4.6)
Interaction/interact (in Swedish: interaction/interagera)	4 (2.6)
Total	**151 (100)**

Note. Asterix (*) marks phrasings that are identical to phrasings used by the SNAE in their guidelines to teachers-as-raters in relation to assessment of the NEST.

Some concepts appeared more than once in an ST. For example, *adaptation to purpose, recipient, and situation* was found twice in several STs. A plausible explanation is the fact that this phrase also appears twice in the *assessment factors* from SNAE: as an ASPECT of the CRITERION *content* and as an ASPECT of the CRITERION *language and ability to express oneself*. Thus, several teachers-as-raters seem to have adopted this approach in their own ST. Additionally, many phrasings in the STs were clearly modified versions of the SNAE phrasings (see Supplementary Materials S1). Although speculative, we believe that the frequency with which concepts were mentioned in the STs are indicative of how often these concepts are used in the SNAE guidelines.

As mentioned, six themes were selected for further analysis to see whether the themes expressed in similar ways in the STs also were conceptualized in a similar way by the teachers-as-raters. The themes *engagement/initiative* and *interaction* were included, despite the fact that they were not mentioned at all on the CRITERION and ASPECT levels for Y6 and mentioned relatively few times for Y9 (*engagement* and/or *initiative* eight times, *interaction* and/or *interact* four times). However, from Step One in the analysis, these themes seemed to be prominent in the second type of categories, where more narrowly defined components of language use or tokens of understanding were described. The six themes selected for further analysis (reported on in relation to RQ3) were:

1. Adaptation to purpose, recipient, and situation
2. Comprehension and clarity
3. Strategies/communicative strategies
4. Richness and variation
5. Engagement/initiative
6. Interaction

In answer to RQ1, then, it can be concluded that the criteria chosen were heavily influenced by the analytic *assessment factors* from SNAE. Not only were phrasings from these frequently used in the STs, but teachers' own phrasings clearly appeared influenced by sub-skills described in SNAE's *assessment factors*. Some phrasings used by teachers-as-raters were the exact same as those used by SNAE, but there were also numerous examples of how SNAE phrasings were modified to become CRITERION and/or ASPECTS to assess in the STs:

- Dividing long phrasings into shorter: *communicative strategies to develop and carry the conversation forward* became *communicative strategies and carry the conversation forward* (Y603).
- Focusing on only one part of a long phrasing: *breadth, variation, clarity, and accuracy* became *variation and breadth* (Y606).
- Reformulating phrasings: *Content—richness and variation—different examples and perspectives* became *treatment of subject—in depth/simplistic* (Y901)
- Constructing new criteria based on parts from different phrasings: *fluency and ease* and *pronunciation and intonation* became *pronunciation, intonation, and fluency* (Y910).

Although most of the criteria included in the STs bore close resemblance to phrasings from SNAE's *assessment factors*, there were also examples of criteria that were more in line with the so-called *knowledge requirements* (see Supplementary Materials S1). Y605, for instance, consisted of four criteria (*oral presentation*, *oral interaction*, *strategies*, and *adaptation to purpose, recipient, and situation*), of which three were identical to the foci of different *knowledge requirements*. In other words, a "label" was put on three of the knowledge requirements and used as criteria for assessment in the ST.

Even though they were few, some criteria were not in line with the NEST assessment guidelines. One example was the CRITERION *argue* found in ST Y607. It stands out given the fact that the passing level of English of Y6 students is equivalent to CEFR beginner level A2 and nowhere in the assessment guidelines, nor in the Y6 syllabus for English is it stated that students at this level should be able to argue for or against anything or even encounter argumentative texts (neither spoken nor written).

3.2. RQ2: How Are Oral Proficiency Sub-Skills to Be Assessed Organized in Teachers' Own Scoring Rubrics? In What Ways Are Teachers' Conceptualizations of the Test Construct Reflected in this Organization?

When studying how OP sub-skills were organized in the STs, we considered both the extratextual and the textual organization of OP sub-skills.

3.2.1. Extra-Textual Organization

Brookhart's (2018) definitions of *checklist*, *rating scale*, and *rubric* guided our categorization of the scoring templates (see Table 5).

Table 5. Categorization of STs for Y6 and Y9.

Level	*Checklist*	*Rating Scale*	*Rubric*
Y6: Number of scoring templates categorized	**4** (Y601, Y602, Y603, Y604)	**0**	**4** (Y605, Y606, Y607, Y608)
Y9: Number of scoring templates categorized	**4** (Y901, Y903, Y910, Y911)	**0**	**8** (Y902, Y904, Y905, Y906, Y907, Y908, Y912, Y913)

Table 5 shows that a majority of STs (12/20) could be categorized as *rubrics*, as these contained descriptive text pertaining to different qualities of criteria. The remaining eight STs were *checklists*, as they listed CRITERION/ASPECTS/SUB-ASPECTS, etc. to be assessed but not what to look and/or listen for. These eight STs instead left room for the teacher-as-rater to comment on the different criteria, either in the shape of an empty box or a plus and minus sign.

The twelve *rubrics* STs differed regarding the number and explicitness of criteria listed for assessment, but one common denominator was the design: a grid containing four to five columns. Criteria to be assessed were placed in the left-hand column, one criterion per row. The remaining columns consisted of performance level descriptions. Nine of the twelve *rubrics* STs had labelled the columns "E", "C", and "A", which mirrors instructions for grading in the Swedish school system. Two *rubrics* STs had, in addition to the three E/C/A columns, a fourth column labelled F (fail), and one had no column labels at all.

When analysing the design of the STs categorized as *rubrics*, we concluded that most boxes in the grids included descriptive text, even though it might be difficult to actually define several levels of quality for a specific criterion. The rubric design therefore seemed to invite teachers-as-raters to fill in descriptive text in all boxes. The fact that all but one of *rubrics* STs had labelled the columns with letter grades suggests that the STs were designed to simplify assessment decisions in the specific situation. Moreover, the small or no space for comments (as boxes were usually filled) suggests that when teachers-as-raters use *rubrics* STs, thoughts, and ideas about students' performances are expressed in the form of highlighting, underlining, or marking (parts of) the pre-written text rather than in the form of making comments.

Using Brookhart's (2018) definition of analytic versus holistic rubrics, the STs can be said to be analytic since all listed sub-skills of OP that were considered in the assessment situation. Although as many as seven of twenty STs only listed two criteria (*Content* and *Language*), these general criteria were always exemplified by ASPECTS and SUB-ASPECTS, which suggests that analytic rather than holistic assessment of students' oral production is taking place when these STs are used for assessment.

3.2.2. Textual Organization

The number of criteria listed in the STs varied (Y6 STs: 2–10; Y9 STs: 2–11). However, sub-skills that were listed as a CRITERION in one ST could be listed as an ASPECT in another. All STs that listed only two CRITERION mentioned *Content* and *Language* (alternatively *Language and ability to express oneself*). In the SNAE guidelines, exactly these two criteria are

mentioned, and therefore, when it also came to how to organize the sub-skills in their STs, teachers-as-raters seemed to use SNAE's assessment factors as a template.

Although the STs often expressed OP sub-skills generally (such as *adaptation to recipient*), several described performance in a much more detailed way. For example, Y905 spanned over three pages and, not surprisingly, very long STs included many SUB-ASPECTS filled with examples (although they did not always present them as such) of what students do or say to demonstrate proficiency relating to a specific criterion. In the example ST Y902, concrete examples of the use of specific words for each of the three grade levels were included, and to assess whether students' content is *rich and varied*, students should put forward different perspectives, shown when using the following phrase: "If . . . I would and I don't, but . . . ".

3.3. RQ3: What Similarities and Differences Are There between Conceptualizations as Reflected in the Scoring Rubrics When It Comes to Sub-Skills to Be in Focus for Assessment?

To understand teachers-as-raters' different conceptualizations of what to include in their assessment of the test construct, we drew tree-diagrams (see Figures 3–5):

Figure 3. ST Y602.

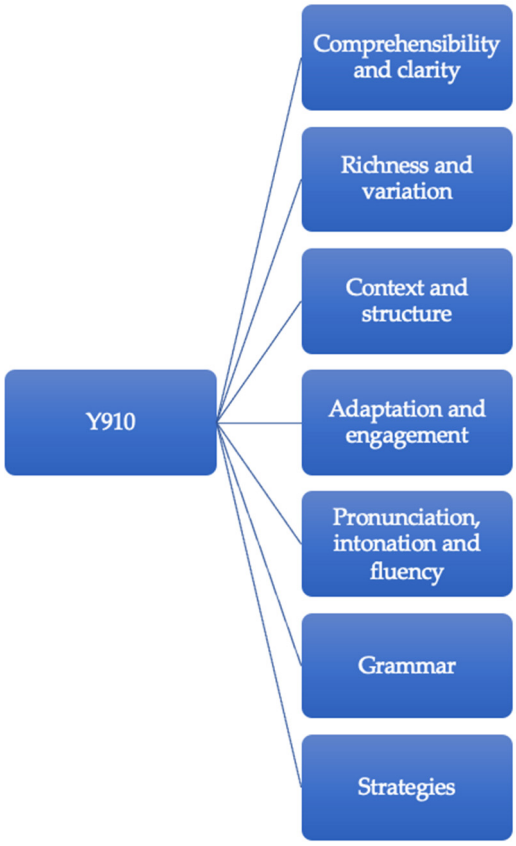

Figure 4. ST Y910.

This first example (Y602, Figure 3) is for an STs categorized as a *checklist*. Two CRITERION were listed (*Content* and *Language and ability to express oneself*), each exemplified by four ASPECTS (e.g., *Content* was exemplified by *comprehension and clarity, richness and variation, context and structure,* and *adaptation to purpose, recipient, and situation*). Some of the ASPECTS were further exemplified/described by SUB-ASPECTS (e.g., *different examples and perspectives*), and in total, this ST consisted of two CRITERIA, eight ASPECTS, and five SUB-ASPECTS. In the analysis, it became evident that CRITERIA, ASPECTS, and SUB-ASPECTS were almost identical to the ones listed in SNAE's *assessment factors*. Presumably, SNAE's *assessment factors* were copied and used in assessing the test construct.

ST Y910 (Figure 4) listed seven criteria. None of them were described or exemplified, so this ST had the lowest number of ASPECTS/SUB-ASPECTS, etc. in our data.

Figure 5 (ST Y907) serves as an example of a detailed ST. It encompassed descriptions on different levels of what students do or do not do to display their level of OP. It had 6 CRITERIA, 16 ASPECTS, 22 SUB-ASPECTS, and 5 SUB-SUB-ASPECTS. The hierarchical structure for the most detailed/deconstructed ST (Y904), which was too extensive to render here, included 7 CRITERIA, 11 ASPECTS, 10 SUB-ASPECTS, 19 SUB-SUB-ASPECTS, 30 SUB-SUB-SUB-ASPECTS, 14 SUB-SUB-SUB-SUB-ASPECTS, and 3 SUB-SUB-SUB-SUB-SUB-ASPECTS.

By studying the more narrowly defined categories, teachers-as-raters' conceptualizations of the themes emerged (for more examples, see Supplementary Materials S2). Content analysis of the themes revealed both similarities and differences between how they were interpreted by the teachers-as-raters. One finding is that some of the themes yielded few SUB-ASPECTS (*adaptation to purpose, recipient, and situation*, for example), while others yielded more (e.g., *communicative strategies*). As a consequence, the amount of data for which we could apply content analysis differed between the themes. For this reason and due to space limitations, selected results are presented below.

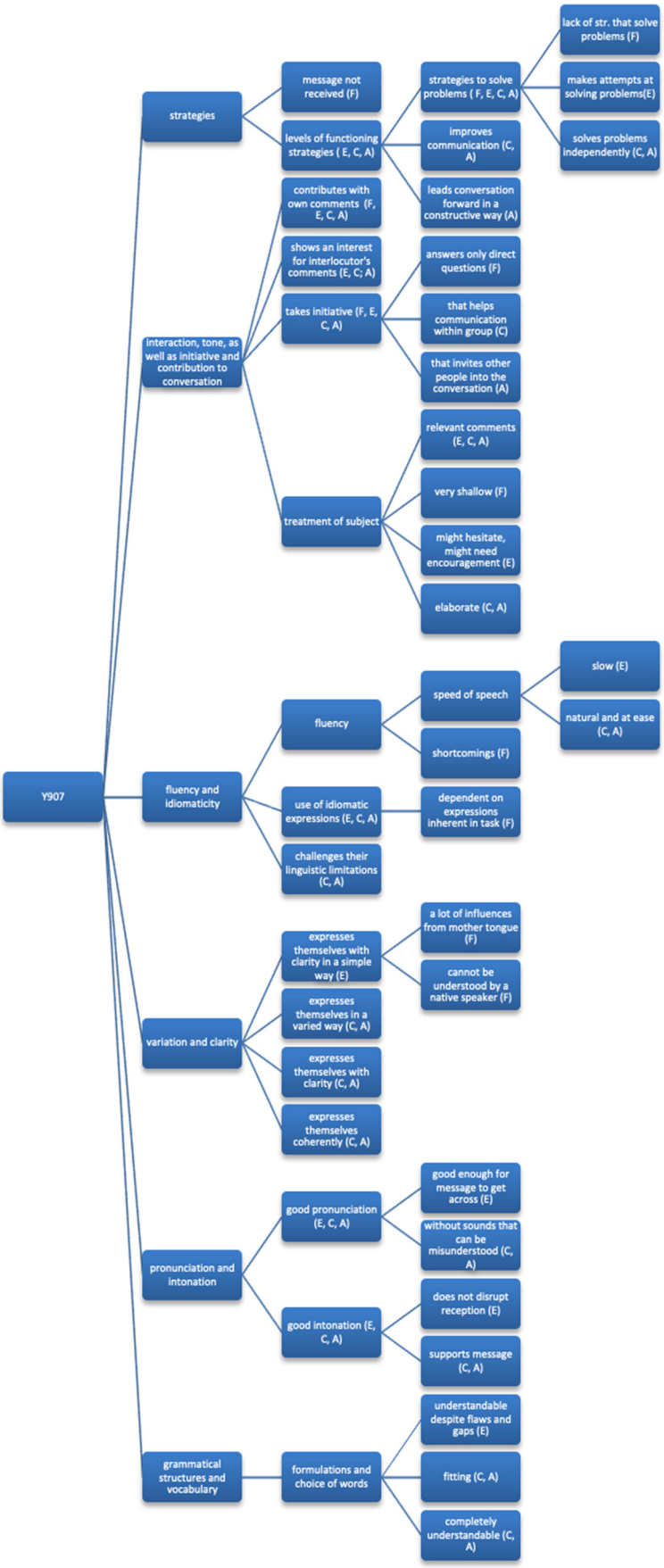

Figure 5. ST Y907.

For two themes, *comprehension and clarity* and *richness and variation*, there was notable consensus. The former was interpreted as a CRITERION assessing students' pronunciation and intonation, although some STs also mentioned to what extent students code-switched to L1. For the latter theme, linguistically, conceptualizations focused on students' use of vocabulary, whereas content-wise, richness and variation meant that students used a variety of examples and perspectives when talking about the topic of the test. The theme *adaptation to purpose, recipient, and situation* was present in as many as 16 rubrics. However, further explanations of what the theme entailed were only found in seven rubrics, which might indicate that this theme is viewed as self-explanatory. Those STs that exemplified the theme further connected it closely to interactive skills and to playing one's part in the test situation, such as "show great interest in participating in the conversation" (grade A, Y605) (cf. testwiseness, Bachman 1990).

The themes with the most diverging conceptualizations were *(communicative) strategies*, *engagement/initiative*, and *interaction*. Conceptualizations of *communicative strategies* differed between STs for Y6 and Y9. For Y6, it was indicated that students should be able to make corrections to their utterances, whereas for Y9, *communicative strategies* encompassed being good interlocutors and conversation partners. However, there was broad consensus regarding how to spot *(communicative) strategies* in students' oral production and how to differentiate between the grades E, C, and A (in essence, based on the number of strategies used, thereby mirroring SNAE's differentiation between grades in the *knowledge requirements* for Y6).

The themes *engagement/initiative* and *interaction/interact* were inter-connected in the STs to the extent that it was difficult to determine whether teachers-as-raters interpreted them as being separate or the same. However, there was a clear difference between the levels. In Y6, assessment of *interaction* was only found in one ST, and the word *engagement* was not found at all (although in two STs, the CRITERION *willingness and ability in speech and conversation to* . . . was used). In addition, in ST Y607, students' ability to "actively participate in conversation" as well as "making other people part of the conversation" suggested that students' level of engagement in the test situation was assessed. In the Y9 STs, the words *engagement/initiative* and/or *interaction/interact* were present in ten of twelve STs, five of which were listed as CRITERION (see Table 6).

Table 6. Occurrence of *Engagement/Initiative* and/or *Interaction/Interact* as CRITERION in the STs for Y9.

In What ST?	Title of Criterion Used in the ST:
Y901	*Interaction/initiative*
Y902	*Interaction/strategies*
Y903	*Willingness and ability in speech and conversation to . . .*
Y907	*Interaction, tone as well as initiative and contribution to conversation*
Y910	*Adaptation and engagement*

Table 6 shows how *engagement*, *interaction*, *(communicative) strategies*, and *adaptation to purpose, recipient, and situation* were conceptualized as being interchangeable criteria. The fact that interaction is part of assessment in a test construct called *oral production and interaction* was not surprising. However, Y9 teachers-as-raters appeared to stress assessment of *interaction* more than Y6 teachers-as-raters did. As mentioned, many themes included phrasings that resembled phrasings from SNAE's guidelines, but this was not the case for interaction (and only very little so for engagement).

4. Discussion

The purpose of this study was to unveil teachers-as-raters' conceptualizations of the NEST construct *oral production and interaction* as these emerge from teachers' self-made scoring rubrics and to examine whether, and possibly how, policy is transformed in the process. To avoid confusion, we have referred to these self-made scoring rubrics as *scoring*

templates. Since teachers-as-raters are provided with extensive assessment guidelines for the test in lieu of specific rater training, it is particularly interesting to analyse how teachers-as-raters conceptualize and transform guidelines when preparing for test assessment. When applying the ATD framework to our empirical findings, teachers-as-raters' *institutional relativity of knowledge* emerged regarding assessment of L2 English OP. The *technique* (i.e., the scoring templates) teachers-as-raters have created to perform the *task* (assessment of a high-stakes test) displayed conceptualizations of the *praxis*. We clearly saw how policy documents, in our case the NEST assessment guidelines for the NEST, influenced teachers-as-raters' decisions on what to include in their scoring templates. We also saw how policy is transformed by teachers-as-raters when scoring rubrics/templates were used for operationalization of the test, since only parts of the SNAE guidelines, more specifically the *assessment factors*, were in focus for assessment. As Khabbazbashi and Galaczi (2020) point out, the content of scoring rubrics mirrors what is made important in assessment. The results of our study suggest that SNAE's *assessment factors* played a particularly important role for the content of the scoring templates examined and, thus, assessment of L2 English proficiency when teacher-generated scoring templates were used in this context. As a result, our study shows that teachers-as-raters were in broad agreement on what aspects should be in focus for assessment, which is in line with both national and international studies (Borger 2014, 2019; Bøhn 2015; Frisch 2015). SNAE's *assessment factors* are introduced to teachers in relation to the assessment of the NEST specifically, which can be a reason why they caught teachers' attention. Even though SNAE instructs teachers to assess holistically based on the *knowledge requirements*, the inclusion of the analytic *assessment factors* in the NEST assessment guidelines may send a signal to teachers-as-raters that the assessment of the test construct in fact is analytic, as our analysis of the teacher-generated STs suggests. If so, instructions to teachers-as-raters to grade students' L2 OP on the *knowledge requirements* expressed in holistic terms, while at the same time considering a number of analytic criteria specific to this test situation, might conflate different meanings of what holistic assessment is. The end-result of the assessment—the grade—is indeed holistic, yet the underlying decision is based on a number of criteria and/or qualities not necessarily mirrored in the grade awarded.

When applying the ATD framework, our results do not allow us to analyse the *technology* (i.e., the *logos*) behind the *technique*. However, our study indicates possible explanations as to why this particular *technique* might be beneficial for carrying out the *task*. All STs in our study were analytic, despite the fact that all were used for summative assessment in an assessment situation where quick decisions about grades need to be taken (Sundqvist et al. 2018). In addition, holistic scales are quicker and less cognitively demanding to use when grading compared to analytic scales (Brookhart 2018; Brown 2012; Davis 2018; Xi 2007). Since analytic rubrics are more beneficial for formative feedback to students (Jönsson and Svingby 2007), it is possible that teachers are more familiar with analytic rather than holistic STs in their daily work and therefore prone to use an analytic rubric as a template when designing their own. SNAE's analytic *assessment factors* can, therefore, fit this type of ST well. An explanation as to why teachers-as-raters decide to construct their own analytic scoring rubric/template could be to improve inter-reliability, as a more systematic assessment process allows for a clearer picture of the criteria underlying the holistic grade (Khabbazbashi and Galaczi 2020), which might facilitate comparisons between raters of the same student performance. Moreover, since the choice between using a holistic and an analytic scoring rubric depends on the assessment situation (Davis 2018), another explanation why teachers decided to use an analytical tool for a summative assessment situation could be that the templates also serve the purpose of displaying students' jagged performance profiles (Davis 2018), enabling for formative feedback to students (Jönsson and Svingby 2007) or for information to parents. This could explain construct-irrelevant concepts found in some of the STs, for instance, the CRITERION *argue* found in ST Y607. It is possible that teachers have designed analytical, task-specific rubrics

used for formative classroom-based task assessment and found these STs also useful for assessing the NEST.

Although the *assessment factors* have clearly influenced both the content and design of the teacher-generated STs, several STs indicate a need for more in-depth descriptions of criteria than the *assessment factors* offer. A case in point for this type of specification is when specific phrases are noted down as indicative of the assessment factor *different perspectives*, something that might seem peculiar. However, these phrases might be used as markers for when students *express different perspectives*. In an assessment situation where many decisions must be made more or less instantly, it might be difficult to focus on *what* students say as well as on *how* they say it. It is possible that access to notes about specific words and phrases facilitates for teachers-as-raters to assess whether different perspectives and examples are presented by the students in their treatment of the topic.

As themes emerging through content analysis of STs were heavily influenced by SNAE's *assessment factors*, no additional focal themes were added by teachers-as-raters. However, the theme *interaction* represented a strong exception. *Interaction* (or *interactive skills*) was included in the STs, but this sub-skill seemed particularly troublesome to define. It is part of the test construct yet nearly not visible at all in the templates for Y6, and in the templates for Y9, it was present in several themes as disparate as *richness and variation* and *adaptation*. There are several possible explanations for why Y6 and Y9 templates differ regarding *interaction*. One is that raters consider the level of the test-takers when focusing on criteria to assess (cf. Sato 2012) and expect Y9 students to have more elaborated interactional skills than Y6 students. Another explanation could be that Y6 and Y9 raters orient to different criteria from the assessment guidelines (cf. Frisch 2015, and a third could be the fact that our data consist of fewer STs from Y6 than Y9, which makes comparison troublesome. However, what our results show is that teachers-as-raters consider Y9 students' interactive skills as essential in the assessment of the test construct, even though their conceptualizations vary a lot about the definition (in line with Borger 2019; and May 2009). A reason for this could be that *interaction* lacks a definition in SNAE's *assessment factors*. Both May (2009) and Borger (2019) point to the importance of developing more elaborated guidelines for assessing interactional skills, since access to such guidelines could facilitate the recognition of interactional skills in test-taker performances, as well as stronger guidance on how to award individual grades for a co-constructed performance. Our study supports this view, as the results indicate that more elaborated guidelines lead to broader consensus in teachers-as-raters' conceptualizations.

Ang-Aw and Goh (2011) show that students were given credit for *effort* regardless of whether such test conduct was part of the test construct or not. Although the word *engagement* is not found in the SNAE *assessment factors*, teachers clearly base their assessment on an action-oriented approach to language teaching and testing, something SNAE instructs them to do. In light of the communicative movement to language teaching and assessment (Bachman 1990, 2007; Canale and Swain 1980), the extent to which students are engaged or show an effort to participate in communication (see *willingness to communicate*, MacIntyre et al. 1998), is conceptualized as a criterion for assessment of the test construct (cf. Sandlund and Greer 2020). However, ideas on how to spot *engagement* or *effort* in students' performances differ between teachers-as-raters in their respective STs, which suggests they might need more elaborate guidelines when it comes to whether *engagement* and/or *effort* should be part of assessment, and if so, how to assess it.

Transformations of the test construct also emerged when analysing the design of the STs, as most of them were designed to enable differentiation between the quality of test-takers' oral production and interaction. Differentiation, therefore, seemed to be an important feature in using STs for summative assessment, which is not surprising since teachers are expected to come up with a grade as the "end-product". However, as several of the *rubrics* STs included relatively extensive descriptions of what quality looked/sounded like on the three grade levels, these STs appear more beneficial for formative feedback to students than for summative grading. Further, the results showed

that, the more teachers-as-raters attempted to describe quality, the more their conceptualizations of oral sub-skills differed. A very detailed ST does not necessarily entail that a teacher-as-rater assesses differently than a teacher-as-rater who uses an ST consisting of more general formulations. The latter teacher-as-rater might very well implicitly know what an E, C, and A performance looks/sounds like, and their assessment might, therefore, be just as de-constructed as one made by a teacher-as-rater who uses an ST that, on paper, is very detailed. The more de-constructed/detailed STs in our study can, therefore, be seen as teachers-as-raters' tacit assessment knowledge verbalised. However, due to the fact that organization of content in a scoring rubric can greatly affect the outcome of grading (Brown 2012; Davis 2018; Khabbazbashi and Galaczi 2020), further studies should examine whether scoring is affected when teacher-generated analytic scoring templates are used, as well as the ways in which they are put to practical use. STs in our data showed two ways of notetaking. *Checklists* STs left room for spontaneous comments relating to pre-determined criteria, and *rubrics* STs were filled with pre-printed descriptive text relating to pre-determined criteria, leaving little or no room left for comments. A question that this study raises is how teachers use such STs in the actual assessment situation. Are STs that leave room for comments used for holistic assessment, while STs filled with pre-printed text are employed for analytic assessment? When using the latter, do teachers-as-raters mainly assess performance by "ticking off" the boxes, and do they assign different values to different criteria? Further research might shed light on how raters of L2 OP use rating scales and if these processes are similar to rater processes of L2 written proficiency (see Lumley 2005).

While this study is fairly small and examines a limited dataset of scoring templates, it nevertheless yielded results and new questions that deserve scholarly attention. In particular, our study sheds light on teachers-as-raters' *institutional relativity of knowledge* of L2 English OP when preparing and conducting assessment in practice. Beyond the scope of the present study was to examine policy documents for assessment with the same scrutiny that was applied for the STs in order for the *institutional relativity of knowledge* inherent in policy to emerge. Therefore, further research is necessary in order to compare and contrast these two different kinds of knowledge in depth. Our results revealed a broad consensus between teachers-as-raters when it comes to *what* to assess but not with regard to *how* to assess. Discrepancies in conceptualizations are particularly salient for *interactive skills*, which is in line with previous studies of rater conceptualizations of assessment of L2 English OP (Borger 2014, 2019). While assessment guidelines for the test seem to have been conceptualized in a similar way regarding what criteria to focus on, when information was missing from *assessment factors* (as is the case for *interaction*), consensus was also missing among our teachers-as-raters.

5. Conclusions

Raters of a standardized, interaction-based test of L2 English OP face at least three challenges. First, they need to consider numerous aspects of quality simultaneously. Second, each interaction is unique, context-specific, and co-constructed by participants, yet individual grades are to be awarded. Third, raters need to attend to issues of both inter- and intra-rater reliability when performing the task. These challenges are all mirrored in the STs examined. It is likely that the construction of STs helps teachers-as-raters attend to the numerous aspects of quality deemed necessary and thereby encompassing them in the holistic grade subsequently awarded. Likewise, the construction of STs might be beneficial for inter- and intra-rater reliability issues, as they might ascertain consistency in assessment. Moreover, listing how to spot and describe *interaction* and *engagement* might help when awarding individual grades for the co-constructed product. In fact, *context*—the conditions of the test situation itself—was a central feature in the teacher-generated documents, indicating that it is an important, and perhaps also particularly complex, part of assessment.

Our results indicate that equity in assessment might be at risk when teachers-as-raters conceptualize quality of learner performances differently and also when they differ regarding what method of scoring to use: analytic or holistic. However, the study sheds light on the dual role of teachers when acting as raters in a high-stakes, standardized test that might contribute to a discussion on who are best suited to assess learners' L2 English OP in this context: humans or computers? External or internal examiners? Since a student's L2 English OP is intricately interwoven with *context*, an examiner who knows the students well can take different contextual conditions into consideration when assigning the grade. Furthermore, in addition to improving reliability, constructing their own analytic scoring instrument might also enable formative feedback to one's own students despite the summative purpose and holistic approach of the test. After all, teachers' core business is to help students learn and develop skills. Being integrated in the assessment process gives teachers invaluable insight and information that can be passed on for the benefit of students' continued development of L2 English oral proficiency.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/languages6040204/s1, S1: Author translation of the assessment guidelines from the Swedish National Agency for Education, S2: Teachers-as-raters' suggestions of student productions illustrative of identified themes.

Author Contributions: Conceptualization, equal contributions by L.B.F., P.S. and E.S.; methodology, equal contributions by L.B.F., P.S. and E.S.; software, N/A; validation, L.B.F. with initial support from P.S. and later also from E.S.; formal analysis, mainly L.B.F. with support from P.S. and E.S.; investigation, mainly L.B.F. with support from P.S. and E.S.; resources, N/A; data curation, L.B.F.; writing—original draft preparation, L.B.F. with support from P.S. and E.S.; writing—review and editing, at different stages in the writing process, equal contributions of offering feedback (reviewing, editing, and supplementing new text) by P.S. and E.S., with L.B.F. always taking main responsibility for preparing revised versions of the manuscript based on the given feedback; visualization, L.B.F.; supervision, equal contributions by P.S. and E.S.; project administration, L.B.F.; funding acquisition, equal contributions by P.S. and E.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: All subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of Karlstad University, Sweden. Approval code: HS 2019/353. Approval Date: 19 March 2019.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Acknowledgments: We would like to express our gratitude to the participating teachers, to our colleague who helped us with the coding, and to the Centre for Language and Literature in Edu-cation at Karlstad University for funding the data collection for this project.

Conflicts of Interest: The authors declare no conflict of interest.

References

Achiam, Marianne, and Martha Marandino. 2014. A framework for understanding the conditions of science representation and dissemination in museums. *Museum Management and Curatorship* 29: 66–82. [CrossRef]

Alderson, J. Charles, and Lyle F. Bachman. 2004. Series editors preface to Assessing Speaking. In *Assessing Speaking*. Edited by J. Charles Alderson and Lyle F. Bachman. New York: Cambridge University Press, pp. ix–xi.

Ang-Aw, Hui Teng, and Christine Chuen Meng Goh. 2011. Understanding discrepancies in rater judgement on national-level oral examination tasks. *RELC Journal* 42: 31–52. [CrossRef]

Bachman, Lyle F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, Lyle F. 2007. What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In *Language Testing Reconsidered*. Edited by Janna Fox, Mari Wesche, Doreen Bayliss, Liying Cheng, Carolyn E. Turner and Christine Doe. Ottawa: University of Ottawa Press, pp. 41–71.

Bøhn, Henrik. 2015. Assessing spoken EFL without a common rating scale. *SAGE Open* 5: 1–12. [CrossRef]

Borger, Linda. 2014. Looking beyond Scores. A Study of Rater Orientations and Ratings of Speaking. Licentiate thesis, University of Gothenburg, Gothenburg, Sweden.
Borger, Linda. 2019. Assessing interactional skills in a paired speaking test: Raters' interpretation of the construct. *Apples–Journal of Applied Language Studies* 13: 151–74. [CrossRef]
Bosch, Marianna, and Josep Gascón. 2014. Introduction to the anthropological theory of the didactic (ATD). In *Networking of Theories as A Research Practice in Mathematics Education*. Edited by Angelika Bikner-Ahsbahs and Susanne Prediger. Cham: Springer, pp. 67–83. [CrossRef]
Brookhart, Susan M. 2018. Appropriate criteria: Key to effective rubrics. *Frontiers in Education* 3: 1–12. [CrossRef]
Brown, James Dean. 2012. *Developing, Using, and Analyzing Rubrics in Language Assessment with Case Studies in Asian and Pacific Languages*. Honolulu: National Foreign Language Resource Center, University of Hawaii.
Canale, Michael, and Merrill Swain. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* I: 1–47. [CrossRef]
Chevallard, Yves. 2007. Readjusting didactics to a changing epistemology. *European Educational Research Journal* 6: 131–34. [CrossRef]
Council of Europe. 2018. Common European Framework of Reference for Languages: Learning, Teaching, Assessment—Companion Volume with New Descriptors. Available online: https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/16807 87989 (accessed on 19 September 2021).
Davis, Larry. 2018. Analytic, holistic, and primary trait marking scales. In *The TESOL Encyclopaedia of English Language Teaching*. Edited by John I. Liontas. Hoboken: Wiley, pp. 1–6.
Ducasse, Ana Maria. 2010. *Interaction in Paired Oral Proficiency Assessment in Spanish: Rater and Candidate Input Into Evidence Based Scale Development and Construct Definition*. Language Testing and Evaluation. Frankfurt am Main: Peter Lang GmbH, Internationaler Verlag der Wissenschaften.
Ducasse, Ana Maria, and Annie Brown. 2009. Assessing paired orals: Raters' orientation to interaction. *Language Testing* 26: 423–43. [CrossRef]
East, Martin. 2015. Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform. *Language Testing* 32: 101–20. [CrossRef]
Frisch, Maria. 2015. Teachers' Understanding and Assessment of Oral Proficiency. A Qualitative Analysis of Results from Interviews with Language Teachers in Swedish Lower Secondary Schools. Licentiate degree thesis, University of Gothenburg, Gothenburg, Sweden.
Hasselgren, Angela. 2000. The assessment of the English ability of young learners in Norwegian schools: An innovative approach. *Language Testing* 17: 261–77. [CrossRef]
Housen, Alex, Folkert Kuiken, and Ineke Vedder. 2012. Complexity, accuracy and fluency. Definitions, measurement and research. In *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Edited by Alex Housen, Folkert Kuiken and Ineke Vedder. Amsterdam: John Benjamins, pp. 1–20.
Hsieh, Hsiu-Fang, and Sarah E. Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative Health Research* 15: 1277–88. [CrossRef] [PubMed]
Jönsson, Anders, and Gunilla Svingby. 2007. The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review* 2: 130–44. [CrossRef]
Kasper, Gabriele, and Steven Ross. 2013. Assessing second language pragmatics: An overview and introductions. In *Assessing Second Language Pragmatics*. Edited by Steven Ross and Gabriele Kasper. Bristol: Palgrave Macmillan, pp. 1–40.
Khabbazbashi, Nahal, and Evelina D. Galaczi. 2020. A comparison of holistic, analytic, and part marking models in speaking assessment. *Language Testing* 37: 333–60. [CrossRef]
Kramsch, Claire. 1986. From language proficiency to interactional competence. *The Modern Language Journal* 70: 366–72. [CrossRef]
Lee, Yong-Won, Claudia Gentile, and Robert Kantor. 2010. Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics* 31: 391–417. [CrossRef]
Lindberg, Viveca, and Åsa Hirsh. 2015. *Formativ Bedömning på 2000-Talet: En Översikt av Svensk och Internationell Forskning*. Stockholm: Swedish Research Council.
Lumley, Tom. 2005. *Assessing Second Language Writing: The Rater's Perspective*. Language Testing and Evaluation: 3. Frankfurt: Peter Lang Publishing Inc.
MacIntyre, Peter D., Richard Clément, Zoltán Dörnyei, and Kimberly A. Noels. 1998. Conceptualizing Willingness to Communicate in a L2: A Situational Model of L2 Confidence and Affiliation. *The Modern Language Journal* 82: 545–62. [CrossRef]
May, Lyn. 2009. Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing* 26: 397–421. [CrossRef]
May, Lyn. 2011. Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly* 8: 127–45. [CrossRef]
McNamara, Tim. 1996. *Measuring Second Language Performance*. New York: Longman.
McNamara, Tim. 2000. *Language Testing*. Oxford: Oxford University Press.
Meadows, Michelle, and Lucy Billington. 2005. *A Review of the Literature on Marking Reliability*. London: National Assessment Agency.
Orr, Mike. 2002. The FCE speaking test: Using rater reports to help interpret test scores. *System* 30: 143–54. [CrossRef]
Papajohn, Dean. 2002. Concept Mapping for Rater Training. *TESOL Quarterly* 36: 219–33. [CrossRef]

Sadler, D. Royce. 2009. Indeterminacy in the use of preset criteria for assessment and grading. *Assessment and Evaluation in Higher Education* 34: 159–79. [CrossRef]

Salaberry, M. Rafael, and Alfred Rue Burch, eds. 2021. *Assessing Speaking in Context. Expanding the Construct and Its Applications*. Bristol: Multilingual Matters.

Salaberry, M. Rafael, and Silvia Kunitz, eds. 2019. *Teaching and Testing L2 Interactional Competence: Bridging Theory and Practice*. London: Routledge.

Sandlund, Erica, and Pia Sundqvist. 2016. Equity in L2 English oral assessment: Criterion-based facts or works of fiction? *Nordic Journal of English Studies* 15: 113–31. [CrossRef]

Sandlund, Erica, and Pia Sundqvist. 2019. Doing versus assessing interactional competence. In *Teaching and Testing L2 Interactional Competence: Bridging Theory and Practice*. Edited by Rafael Salaberry and Silvia Kunitz. Oxon and New York: Routledge, pp. 357–96.

Sandlund, Erica, and Pia Sundqvist. 2021. Rating and reflecting: Displaying rater identities in collegial L2 English oral assessment. In *Assessing Speaking in Context. Expanding the Construct and Its Applications*. Edited by Rafael Salaberry and Alfred Rue Burch. Bristol: Multilingual Matters.

Sandlund, Erica, Pia Sundqvist, and Lina Nyroos. 2016. Testing L2 talk: A review of empirical studies on second-language oral proficiency testing. *Linguistics and Language Compass* 10: 14–29. [CrossRef]

Sandlund, Erica, and Tim Greer. 2020. How do raters understand rubrics for assessing L2 interactional engagement? A comparative study of CA- and non-CA-formulated performance descriptors. *Papers in Language Testing and Assessment* 9: 128–63.

Sato, Takanori. 2012. The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing* 29: 223–41. [CrossRef]

Sert, Olcay. 2019. The interplay between collaborative turn sequences and active listenership: Implications for the development of L2 interactional competence. In *Teaching and Testing L2 Interactional Competence: Bridging Theory and Practice*. Edited by M. Rafael Salaberry and Silvia Kunitz. London: Routledge, pp. 142–66.

Skehan, Peter, and Pauline Foster. 1999. The influence of task structure and processing conditions on narrative retellings. *Language Learning* 49: 93–120. [CrossRef]

Sundqvist, Pia, Peter Wikström, Erica Sandlund, and Lina Nyroos. 2018. The teacher as examiner of L2 oral tests: A challenge to standardization. *Language Testing* 35: 217–38. [CrossRef]

Swedish National Agency for Education. 2021. Nationella prov i grundskolan. Available online: https://www.skolverket.se/undervisning/grundskolan/nationella-prov-i-grundskolan (accessed on 25 May 2021).

Tsagari, Dina. 2021. Language assessment literacy: Concepts, challenges, and prospects. In *Perspectives on Language Assessment Literacy: Challenges for Improved Student Learning*. Edited by Sahbi Hidri. London and New York: Routledge, pp. 13–32.

University of Gothenburg. 2021. Nationellt prov i Engelska för Årskurs 9. Available online: https://www.gu.se/nationella-prov-frammande-sprak/prov-och-bedomningsstod-i-engelska/engelska-arskurs-7-9/nationellt-prov-i-engelska-for-arskurs-9 (accessed on 18 September 2021).

Wang, Binhong. 2010. On rater agreement and rater training. *English Language Teaching* 3: 108–12. [CrossRef]

Xi, Xiaoming. 2007. Evaluating analytic scoring for the TOEFL® academic speaking test for operational use. *Language Testing* 24: 251–86. [CrossRef]

Youn, Soo Jung, and Shi Chen. 2021. Investigating Raters' Scoring Processes and Strategies in Paired Speaking Assessment. In *Assessing Speaking in Context: Expanding the Construct and Its Application*. Edited by M. Rafael Salaberry and Alfred Rue Burch. Bristol: Multilingual Matters, pp. 107–31.

Young, Richard, and Agnes Weiyun He, eds. 1998. *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency*. Studies in Bilingualism (14). Amsterdam: John Benjamins.