Article

SAGE Open
October-December 2015: 1–12
© The Author(s) 2015
DOI: 10.1177/2158244015621956
sgo.sagepub.com
SAGE

Assessing Spoken EFL Without a Common Rating Scale: Norwegian EFL Teachers' Conceptions of Construct

Henrik Bøhn[1,2]

Abstract

This study investigated teacher cognition and behavior in a high-stakes, English as a Foreign Language (EFL) school context where no common rating scale exists. 24 EFL teachers at the upper secondary level in Norway were asked to rate the performance of a student taking her oral English exam and to give an account of what kind of performance aspects they pay attention to in the rating process. The study showed that while the raters had the same general ideas of the constructs to be assessed, there were differences in how they perceived the relative importance of these constructs, particularly as regards topical knowledge. The study has implications for language teaching and assessment practices at the intermediate to upper-intermediate levels (Common European Framework of Reference, level B1/B2), particularly with regard to the role of topical knowledge.

Keywords

language assessment, English as a Foreign Language, spoken L2, oral English exam, constructs, criteria

Introduction

The question of constructs, or *what* to be tested, is crucial in language assessment. Constructs are typically operationalized in written rating scales (Fulcher, 2012; Luoma, 2004), which are usually provided for raters in high-stakes tests. In the Norwegian educational system, however, there are no national requirements for the provision of common rating scales in the assessment of oral English at the upper secondary level. A general framework exists in the form of national legislation, general directives, and a national curriculum, but the operationalization of the constructs is left to the local level, which in many cases means the individual teachers.

A number of studies have shown that raters pay attention to different aspects of performance when rating spoken English as a Second/Foreign Language (ESL/EFL), but most of this research has focused on assessment and testing in contexts where common rating scales exist (e.g., Ang-Aw & Goh, 2011). What happens in situations where the constructs have not been operationalized is much less clear, however.

The aim of this study, therefore, is to explore how EFL teachers in Norway understand the constructs to be tested in a high-stakes, oral English exam at the upper secondary level, where no common rating scale has been provided. The major focus will be on rater cognition, but as part of this discussion, the issue of rater consistency will also be considered. Understanding which aspects of performance raters pay attention to is important for informing the design of test tasks, the selection of criteria for assessment, and the creation of rating scales (Pollitt & Murray, 1996; Taylor & Galaczi, 2011).

Literature Review

National and international studies have found variability in rater cognition and rater behavior in L2 speaking assessment. Internationally, for example, research has found that raters pay attention to a range of different aspects of performance in the rating process (Brown, 2000; Hsieh, 2011; Orr, 2002). More specifically, they may vary considerably in how they perceive the importance of the various criteria in the rating scales, such as the use of vocabulary (Ang-Aw & Goh, 2011; Brown, 1995; Eckes, 2009; Kim, 2009). There is also evidence that raters pay attention to different aspects of performance depending on level. For example, in the assessment of low-level performance, raters are more likely to heed features such as grammar and pronunciation, whereas at more advanced levels, they will

[1]University of Oslo, Oslo, Norway
[2]Østfold University College, Halden, Norway

Corresponding Author:
Henrik Bøhn, Department of Teacher Education and School Research, University of Oslo, P.O. Box 1099, Blindern, Oslo 0317, Norway.
Email: henrik.bohn@hiof.no

pay more attention to aspects such as fluency and content (Brown, Iwashita, & McNamara, 2005; Pollitt & Murray, 1996; Sato, 2012). There is also research showing that raters attend to non-relevant criteria in their assessment of performance, one example being the voice quality of the test takers (Brown, 2000; Orr, 2002; Sato, 2012). In addition, there are indications that raters give test takers credit for effort, regardless of whether it is defined as part of the construct or not (Ang-Aw & Goh, 2011; Brown, 1995; May, 2006; Pollitt & Murray, 1996).

However, with the exception of Brown et al. (2005) and Pollitt and Murray (1996), none of the above-mentioned studies have looked into rater orientations in contexts where common rating scales are absent. Moreover, it is worth noting that both Brown et al. and Pollitt and Murray studied rater cognition in high proficiency level contexts (English for Academic Purposes and the Cambridge Certificate of Proficiency in English, respectively), whereas the level under investigation in this study is upper intermediate. As raters may attend to different criteria at different levels, it is relevant to study teachers' conceptions of constructs also at the intermediate to advanced level.

In the Norwegian context, there is very little empirical evidence on how raters operationalize the construct in oral English exams. However, studies investigating assessment practices more generally, including subjects such as English and Norwegian, have found that teachers may find criterion-referenced assessment difficult, even though such assessment is required by the regulations of the Education Act (Hægeland, Kirkebøen, Raaum, & Salvanes, 2005; Prøitz & Borgen, 2010). More specifically, there are studies indicating that teachers find it difficult to describe student competence at different levels (Throndsen, Hopfenbeck, Lie, & Dale, 2009). As for the question of teacher cognition in the assessment of oral English, only a master's study, Yildiz (2011), has cursorily investigated this issue. The study indicated that Norwegian teachers heed different aspects of performance, that they weigh criteria differently, and that they employ non-relevant criterion information in the rating process. Consequently, with so little national and international research having been undertaken, the present investigation adds valuable empirical evidence to the field of spoken L2 assessment at the upper-intermediate proficiency level. In this study, the following two research questions will be addressed:

Research Question 1 (RQ1): How do EFL teachers in Norway understand the constructs and criteria to be tested in an oral English exam at the upper secondary level?

Research Question 2 (RQ2): What kind of criteria do these teachers see as salient when assessing performance?

Theoretical Considerations

Despite the view that terminology such as "construct" and "construct definition" is less useful for explaining observed behavior in assessment situations (Kane, 2006, 2012), a number of language assessment and testing specialists still find it relevant as a way of conceptualizing what should be tested (Bachman & Palmer, 2010; Fulcher, 2015; Fulcher & Davidson, 2007; Green, 2014; Hulstijn, 2011; Inbar-Lourie, 2008). According to Fulcher and Davidson (2007), a construct can be considered an unobservable concept, usually identified by an abstract noun, which needs to be defined so that it can be scientifically investigated. This means that "it can be *operationalized* so that it can be measured" (Fulcher & Davidson, 2007, pp. 369-370, emphasis in original). One of the examples Fulcher and Davidson use is "fluency." Thus, to assess "fluency," one would have to decide on its operationalization, for example, using performance features such as "pauses," "fillers," and "false starts" as indicators of this construct (Brown et al., 2005, p. 23).

In language assessments, constructs typically relate to one or more aspects of language ability. However, they may also relate to aspects of content, or topical knowledge (Bachman & Palmer, 2010). Here, it is also worth noting that constructs typically have a source or a "frame of reference," such as a course syllabus that helps assessment designers to operationalize the constructs (Bachman & Palmer, 2010, p. 211). In the Norwegian system, it is the subject curriculum that forms the basis for this operationalization.

In the language assessment and testing literature, one also frequently comes across the term "criterion" in relation to what should be assessed (e.g., Council of Europe, 2001; Cumming, 2009; Lumley, 2002; Luoma, 2004; Stoynoff, 2009; Taylor, 2006). This concept has been defined in a number of different ways (e.g., Glaser & Klaus, 1962; Popham, 1978). However, the notion of criterion that best fits with the approach taken in this study is that of Brindley (1991), who says that criteria are the "the key *aspects of performance* . . . to be assessed such as fluency, appropriacy, accuracy, pronunciation, grammar etc." (p. 140, emphasis added). Interestingly, Brindley mentions fluency as an example of a criterion. This may appear confusing as Fulcher and Davidson (above) use fluency as an example of a construct. To avoid this confusion, I will in the following reserve the term "construct" for the *broader categories of concepts* under investigation and use the terms "criteria," "sub-criteria," and "sub-sub-criteria" for the *more narrowly defined performance aspects*. An example of a construct will be "communication," whereas examples of criteria, sub-criteria, and sub-sub-criteria will be "linguistic competence," "grammar," and "subject-verb concord," respectively. In this discussion, then, there is a hierarchical relationship between the constructs and the criteria, the sub-criteria, and the sub-sub-criteria (cf., Tables 2 and 4).

The Situation in Norway

In Norway, English is a compulsory subject from the first grade onward. Consequently, by the time students enter upper secondary school, at the age of 16, they have normally reached an upper-intermediate level (Common European Framework of Reference, level B1/B2). The subject curriculum is standards based, listing a number of competence aims that specify what students are expected to master at the end of instruction at different levels. These aims are grouped together in three "main areas":

i. *Language and language learning*—involving aims such as "the pupil shall be able to . . . exploit and assess various situations, working methods and strategies for learning English."
ii. *Communication*—comprising aims such as "the pupil shall be able to . . . express him/herself in writing and orally in a varied, differentiated and precise manner, with good progression and coherence."
iii. *Culture, society, literature*—including aims such as "the pupil shall be able to . . . present and discuss international news topics and current events."[1]

At the upper secondary level, the English course involving this curriculum is obligatory for both students at the general studies program (GSP) and the various vocational studies programs (VSPs). However, the GSP students complete the course after 1 year (GSP1), whereas the VSP students complete it after 2 years (VSP2). The fact that these two groups of students are made to take the same course has caused some tension in the past, drawing criticism from stakeholders who have found the course far too academic for VSP students, who are allegedly less proficient in English (e.g., Solheim, 2009).

End-of-instruction assessment is mainly given in the form of overall achievement marks, awarded by each subject teacher at the end of the school year on the basis of various forms of classroom assessment. In addition, approximately 20% of the students are randomly selected for written English exams and 5% for the oral English exams at the GSP1 and VSP2 levels. Marks range from 1 ("fail") to 6 ("excellent").

In contrast to the written exam, which is administered nationally by the Directorate for Education and Training, the administration of the oral exam is left to the local educational authorities through the county governors. A direct consequence of this policy is that while the written exam is standardized in terms of a common exam format, common exam tasks, and a common written rating scale, there is no such standardization for the oral exam. Actually, in many cases, the local educational authorities leave it to the individual schools to decide for themselves, particularly with regard to exam tasks and rating scales.[2]

Method

Research Design

This study is primarily qualitative. As the focus is predominantly on rater cognition, it was decided to use semi-structured interviews as a means to tap into the "life-world" of the informants (Kvale, 2007; Kvale & Brinkmann, 2009). To obtain relevant interview data, it was also decided to use a prompt in the form of a videotaped oral exam performance. This prompt was then distributed to a group of teachers who were asked to watch the video-clip, score the performance, and write down their comments explaining what kind of criteria they applied in the rating process. The informants were then interviewed individually, and in the interviews they were asked to answer both the question on criteria related *specifically* to the performance in the video-clip and the question on criteria to be applied more *generally*. In addition, the informants were asked to score the performance, to obtain a consistency measure as well as an indication of rater behavior, which could then be used to validate the rater orientation analysis (Krippendorf, 2013).

Content analysis was used in the exploration of the data. This method can be used both quantitatively and qualitatively, and in the present study, I have used both approaches (Galaczi, 2014; Hsieh & Shannon, 2005; Krippendorf, 2013). According to Hsieh and Shannon (2005), the qualitative approach is particularly relevant when existing theory or research literature on a phenomenon is limited, which is the case in the Norwegian context. Due to this lack of prior conceptualizations, I carried out the analysis inductively, letting the construct and criterion categories emerge from the data (Galaczi, 2014). As for the quantitative aspects of the investigation, the frequencies of the categories that emerged may serve as an index of the salience of these categories (cf., Krippendorf, 2013).

Participants

As for the filming of student performance, a VSP student who volunteered to participate was videotaped as she was taking her oral exam. The exam format consisted of three tasks: (a) a pre-planned monologue task in the form of a presentation, followed by a discussion of the presentation; (b) an oral interview task based on a short story from the syllabus; and (c) an oral interview task based on a listening comprehension exercise. When it comes to the recruitment of teacher interviewees, purposeful sampling (Creswell, 2013) was employed to ensure variation with regard to age, gender, geographical location, experience, and study program affiliation. In total, 24 teachers from the three counties of Finnmark, Oslo, and Østfold were recruited by contacting schools directly.[3] All the informants had English teaching experience from the upper secondary level, and all but one of them (No. 23) had examined students at the oral English GSP1/VSP2 level. Some only taught students in the VSPs, some were exclusively involved in the GSP, and some were involved in both types of programs. The background information on the informants as well as the score they awarded to the student in the video-clip are summarized in Appendix A.

Interview Procedure

A semi-structured interview format was chosen, and an interview guide was piloted and revised (cf. Appendix B). Seven teachers in Østfold and one in Oslo were interviewed face-to-face, whereas the rest were interviewed via telephone. The informants were asked to watch the video-clip immediately before the interview was scheduled to keep the event as vividly in their minds as possible. All interviews were recorded. No specific analysis of interviewer effects, that is, interviewer influence on informants' answers, has been carried out (e.g., Kreuter, 2008). However, it appears that typical variables known to make a difference in this respect, such as

sensitivity of topic, marginalized respondents, and older or hearing-impaired respondents, have not affected the responses negatively (e.g., Shuy, 2003).

Data Analysis

The data were analyzed using the computer software package QSR NVivo10. The analysis was carried out in several stages. First, the interviews were transcribed and checked, and all the transcripts were read through to get an overall impression of the material. Second, the transcripts were divided into three sections, each corresponding to the three interview questions: (a) criteria used for scoring performance in the video-clip, (b) criteria generally, and (c) most important criteria. Within these three sections, teacher statements were divided into "ideas units" based on the nature of the research questions. An ideas unit can be described as "a single or several utterances with a single aspect of the event as the focus," that is, a unit that is "concerned with a distinct aspect of performance" (Brown et al., 2005, p. 13). The following excerpt serves as an illustration, in which the ideas units boundaries have been marked by a "/":

> /She is fairly fluent,/and there are no serious errors hampering communication, right?/And she tackles that quite well, even if she has to stop and switch into Norwegian a couple of times./But there are no errors hampering communication./

The ideas units in the above excerpt were coded as "Fluency," "Disruptive features," and "Compensatory strategies."

In the next stage, a coding scheme was developed. This entailed the comparison of codes with statements and codes with other codes in a cyclical process (Galaczi, 2014). Having developed the coding scheme, I coded all the transcripts. This process also involved the quantification of statements by making category counts. An ideas unit that was mentioned in one section, such as "and there are no serious errors hampering communication, right?" (cf. the extract above), was counted once. If the same ideas unit appeared within the same section, like, for example, "But there are no errors hampering communication" (cf. above), it was not counted again. However, if it appeared in two or three sections, it was counted 2 or 3 times. It should be pointed out that this type of quantification does not automatically reveal strength of correlation between statements and the prominence of a category. However, it can be validated against "behavioral effects" (Krippendorf, 2013, p. 31) such as the scores awarded by the teachers. In addition, it can be corroborated by the qualitative analysis, which may support findings through the in-depth scrutiny of statements. Both of these validation procedures were employed here.

To validate the analysis, two colleagues, who had previously worked as English teachers at the upper secondary level in Norway, were asked to code four transcripts (a total of 16% of the transcripts). The match between their coding and my own resulted in a Cohen's Kappa reliability estimate of .69, which is regarded as moderate agreement (Landis & Koch, 1977). The mismatched codes were then discussed, and the coding scheme was revised. I then re-analyzed all the transcripts, and one of the colleagues agreed to code two new transcripts. The intercoder reliability analysis in this phase resulted in a Kappa estimate of .89, which can be labeled very good (Landis & Koch, 1977).

Findings

As the investigation of rater behavior was used to validate the rater orientations analysis (cf., "Method" section), it is relevant to briefly look at the interviewees' scoring of the performance in the video-clip. Table 1 gives an overview of the frequencies and percentages of the scores, as well as the mean score and the standard deviation. (For the assignment of individual scores, see Appendix A.)

Table 1. Frequencies and Percentages of Scores, Mean Score, and Standard Deviation.

Grade	Frequency	Percentage	*M*	*SD*
2	3	12.5		
3	15	62.5		
4	6	25.0		
N	24	100	3.13	.612

As Table 1 shows, most of the teachers awarded the performance a 3. The standard deviation of .612 further indicates moderate spread in the scoring. This means that the teachers largely agreed that it was an average performance.

RQ1: Teachers' Notions of Constructs and Criteria

The results for RQ1 are based on the informants' answers to the questions of which performance aspects they pay attention to in the rating process, both in terms of the *specific* case of the student performance in the video-clip, as well as in the assessment of oral performance at this level more *generally*. It should be observed that all the teachers reported that they score performance holistically.

The coding of statements yielded a total of 56 categories. These were then ordered into "construct," "criterion," "sub-criterion," and "sub-sub-criterion" categories (see Tables 2-5). In total, 38 of these categories related to student performance irrespective of task, 17 were relevant for the presentation task only, and one related solely to the short story discussion task. However, for reasons of space, I will restrict my presentation and discussion here to the 38 categories that relate to performance irrespective of task. An overview of the construct and criterion categories, with one example statement for each criterion, is presented in Table 2. Note that many of these criteria, such as "Linguistic competence," have sub-criteria and sub-sub-criteria that are not displayed in Table 2 but will be displayed in Table 4:

Table 2. Constructs and Criteria Developed From Teacher Statements: Categories and Examples.

Constructs	Criteria	Examples
Communication	(General reference)[a]	"I think in terms of communication . . . she was able to communicate" (No. 21)
	Linguistic competence	". . . the vocabulary was reasonably limited . . . simple sentences with quite a few grammar errors" (No. 9)
	Compensatory strategies	"And if they can't [find the word], they should try to circumvent it, rather than switching into Norwegian" (No. 4)
	Listening comprehension	". . . she's got good listening skills. When she was asked a question, there was no problem understanding" (No. 9)
	Take initiative	"the student needs to . . . contribute to keep the conversation going" (No. 5)
	Adapt communication to situation and audience	". . . she adapts her language to the situation" (No. 2)
	Cohesion	". . . the importance of using paragraph-connectors . . . 'firstly,' 'secondly'" (No. 4)
	Ability to repair	". . . I see students who are able to self-correct orally . . . " (No. 24)
	Social competence	". . . to me, that communication thing is to some extent a social issue; you are supposed to put yourself into it" (No. 22)
Content	(General reference)[b]	"She probably says too little about the topic, actually" (No. 13)
	Application, analysis, reflection	"She totally missed out on the second . . . the analysis part of the question" (No. 9)
	Comprehension (explain using own words)	"you're testing their understanding" (No. 23)
	Knowledge (reproduction)	". . . she is able to recount the content of the short story" (No. 2)
	Addressing task or problem statement	". . . she doesn't really answer the task question" (No. 11)
	Elaborated response	". . . she didn't respond well to the questions . . . she answered in three words and ended with a 'yes'" (No. 2)
	Content structure	"I think she structured the retelling of the story well" (No. 24)
(Other)	Disruptive features	"[these aspects] are not really hampering communication" (No. 24)
	Preparation	"I think she has prepared well, according to her level, that is" (No. 21)
	Effort	"But trying isn't in the competency criteria. But I think it should be." (No. 14)

[a]"General reference" is not a criterion in itself, but a category that summarizes all the instances where the informants mentioned "communication" or "to communicate" (as in the example provided in Table 2).
[b]cf. Note 5.

In this analysis, two constructs emerged from the coded statements, namely, "Communication" and "Content." With the exception of the criteria "Disruptive features," "Preparation," and "Effort," which were put in an "Other" category, all the criteria, sub-criteria, and sub-sub-criteria relate to these two constructs. This is not surprising, given that the subject curriculum warrants the identification of the same two constructs. The three main areas in the curriculum, that is, "Language learning," "Communication," and "Culture, society and literature" (cf., "The Situation in Norway" section), can, in my interpretation, be subsumed under the headings "Communication" and "Content." On one hand, students are expected to be able to communicate, and on the other, they are expected to know something about language, language learning, and cultural issues.

When it comes to the "Communication" construct, it should be noted that 11 informants spoke of "language" as an "overall category." For example, Informant No. 9 said, "Usually I identify three areas—content, organization and language—and then, in the descriptors for each grade, I write exactly what I expect students to perform."[4] However, I would argue that "language," which I have here termed "Linguistic competence," is in fact a *sub-category*, or a criterion that belongs to "Communication." Support for this claim is found in the many statements concerning other criteria that are not linguistic, but which are closely connected to "Linguistic competence," and which, taken together, logically make up what can be labeled "Communication" (cf., Tables 2 and 4). This also fits theoretically with a communicative approach to language teaching, which the Norwegian educational system draws on through the Common European Framework of Reference (North, 2004; Simensen, 2010).

As for the "Content" construct, the informants' statements pertaining to this category largely turned out to involve classification. Consequently, it was deemed relevant to use a taxonomy in the coding of some of these references. Adapting Bloom, Englehart, Furst, Hill, and Kratwohl's (1956) taxonomy, I therefore found it pertinent to apply the criterion categories "Knowledge," "Comprehension," and "Application, analysis, reflection." A statement from Informant No. 2 supports this decision:

> And it's often these three levels one relates to: Are they just reproducing facts, are they on a level where they understand some

> more, and are able to use it to some extent, or have they reached a level where they are able to analyse, reflect and compare?

It may be objected here that a criterion should not involve classification. However, keeping in mind the definition of *criteria* used in this study, that is, "aspects of performance . . . to be assessed," I would argue that such criterion categories are relevant in the present analysis. Not only do some teachers report that they assess performance according to these three categories, such a classification is also internally consistent in the sense that all the criteria reflect aspects of performance that can be linked to level indicators, like, for instance, "poor," "average," or "good." In other words, just as the teachers may find a student's linguistic competence to be good, they may also judge her ability to reflect upon topical knowledge to be good.

When it comes to the categories "Disruptive features," "Preparation," and "Effort," which did not clearly relate to the two overall constructs, they were referred to less frequently (see Table 3). As for "Effort," there are indications that some of the teachers rate VSP students more leniently than they do GSP students, especially the weaker students who risk failing. This means that the teachers may give credit to students who "try their best" in order to compensate for lack of language or content knowledge. Informant No. 14 reflects this sentiment:

> We've had a lot of non-native Norwegians, who are in a [vocational] programme. They're going to become hairdressers and they're going to work at [the local supermarkets], and oftentimes we have students that understand very little English. They can't even have an ongoing, real discussion with you in the classroom. . . . We see how broken these kids are [and] passing English is the difference between getting a job and not getting a job. . . . I say to a lot of kids: "If you come and you try, I will do my best to give you a two" . . . But trying isn't in the competency criteria. But I think it should be.

However, this picture is balanced by some of the other VSP teachers who take the opposite stance. Nos. 22 and 23, for example, categorically deny that they would give extra credit for effort. "I am not allowed to do that," No. 22 says. This is aptly remarked as the national educational authorities have stipulated that effort is not to be assessed (Norwegian Directorate for Education and Training [UDIR], 2010).

RQ2: Teachers' Notions of Salient Criteria

As for RQ2, the answer to this question is based on two types of evidence. First, it is based on the reference counts for each of the categories that emerged in the general quantitative analysis (cf., Tables 3 and 4). Second, it is evidenced by the answers given to the interview question concerning which performance aspects the teachers considered as salient (the "most important criterion" question; cf. Table 5).

Table 3 presents the total number of counts that were made for the constructs and the criteria in the general quantitative analysis. (Note that the figures in Table 3 include the counts for the sub-criteria and the sub-sub-criteria, although these have not been specified here; cf., Table 4.)

Table 3. Number of Reference Counts for the Different Statements Pertaining to Constructs and Criteria.

Constructs	Criteria	Reference counts
Communication	(General reference to communication)	28
	Linguistic competence	240
	Compensatory strategies	24
	Listening comprehension	21
	Take initiative	15
	Adapt communication to situation and audience	6
	Cohesion	2
	Ability to repair	2
	Social competence	2
	Sum Communication	**340**
Content	(General reference to content)	43
	Application, analysis, reflection	44
	Comprehension (explain using own words)	30
	Knowledge (reproduction)	27
	Addressing task or problem statement	26
	Elaborated response	15
	Content structure	4
	Sum Content	**189**
(Other)	Disruptive features	17
	Preparation	14
	Effort	7
	Sum Other	**38**

As Table 3 shows, "Linguistic competence" was the criterion category that received by far the most counts in the general quantitative analysis (240 counts). In fact, it is more than 5 times larger than the second largest category, "Application, analysis, reflection" (44 counts), and 8 times larger than the third category, "Comprehension" (30 counts). This does not necessarily mean, however, that the teachers see language ability as 5 to 8 times more important than the ability to understand or analyze content, but it reflects the fact that they mention a larger number of different aspects of language when they are asked to discuss criteria. This can be seen in Table 4, which lists the seven sub-criteria and nine sub-sub-criteria that were developed from teacher statements relating to "Linguistic competence." In comparison, no teacher statements produced sub-criteria or sub-sub-criteria within the "Application, analysis, reflection" or "Comprehension" categories. In passing, it should be mentioned that a number of references to "structure" have been left out, because it was difficult to decide whether the respondents referred to "Cohesion" or "Content structure" (cf.,Table 2).

Table 4. "Linguistic Competence": Sub-Criteria, Sub-Sub-Criteria, and Reference Counts.

Criterion	Sub-criteria	Sub-sub-criteria	Reference counts
Linguistic competence	(General reference)		31
	Grammar	(General reference)	25
		Syntax	12
		Subject-verb concord	9
		Tense	2
		Adjective/adverb	1
		Sum Grammar	**49**
	Vocabulary	(General reference)	40
		Technical	8
		Advanced/nuanced	6
		Sum Vocabulary	**54**
	Phonology	(General reference)	0
		Pronunciation	48
		Intonation	15
		Stress, rhythm, pauses	4
		Sum Phonology	**54**
	Fluency		25
	Idioms, metaphors		9
	Independence/originality		3
	Accuracy		2
Sum linguistic competence			**240**

Table 5. Most Important Criteria and Sub-Criteria Mentioned by 19 Out of 24 Informants.

Constructs	Criteria	Sub-criteria	Counts
Communication	(General reference)		
	Linguistic competence	(General reference)	2
		Pronunciation	4
		Vocabulary	4
		Grammar	2
		Sum Linguistic competence	12
	Compensatory strategies		3
	Listening comprehension		1
	Sum Communication		**16**
Content	(General reference)		3
	Application, analysis, reflection		9
	Addressing task or problem statement		5
	Sum Content		**17**

As the "Linguistic competence" category turned out to be so comprehensive, including sub-criteria that received a substantial number of counts, it is relevant to briefly consider some of these subcategories. As can be seen in Table 4, the two most prominent of these, "Vocabulary" and "Phonology," both received 54 counts. The third largest of the subcategories, "Grammar," received 49 counts. In other words, all these sub-criterion categories received more counts than the biggest "Content" category, that is, "Application, analysis, reflection."

Moving on to the analysis of the most important criterion question, one gets a fuller account of what the teachers see as salient criteria. This analysis is based on statements like the following (from No. 6): "And then there is the fact that she hasn't answered the whole task. That's what marks it down the most." Due to the emergent nature of this research design (Creswell, 2013), not all the informants were systematically asked about which criteria they see as most important. Only 19 out of 24 teachers gave answers to this question. Consequently, the findings reported in Table 5 may give an incomplete picture of the entire teacher sample's response to this question. Nevertheless, when comparing the results in Table 5 with the number of counts in Tables 3 and 4, one gets a more complete picture of the salient criteria.

The answers reported in Table 5 largely supported the findings summarized in Tables 3 and 4, although with some modifications. Again, "Linguistic competence" turned out to be the most prominent criterion category, followed by "Application, analysis, reflection." However, the difference between these two categories was not in any way as substantial as was the difference in the general quantitative analysis. As can be seen in Table 5, 12 counts were made for "Linguistic competence" and nine counts were made for "Application, analysis, reflection." Interestingly, the third largest category identified in the general analysis, that is, "Comprehension," was not mentioned at all in the "most important criterion" discussion. Instead, "Addressing task or problem statement" was cited as a salient criterion (five counts). As this criterion also received a number of counts in the general quantitative analysis (26 counts), it seems clear that the teachers consider it to be important. Finally, it is worth noting that the three subcriteria emerging as important in the general analysis, that is, "Vocabulary," "Phonology," and "Grammar," were all pointed to as salient criteria. "Vocabulary" and "Phonology" received four counts each, and "Grammar" received two counts.

As for variation in the teacher responses, I found a clear distinction in the data pertaining to the criterion category "Addressing task or problem statement." The informants only involved in the GSP seemed to be particularly concerned with this criterion, whereas the teachers only involved in the VSPs did not mention it at all. For example, the three informants who awarded the student in the video-clip a 2 (Nos. 11, 12, and 17) mentioned lack of task response as a dire weakness in the candidate's performance. All of these are GSP teachers. Informant No. 11 put it this way:

> So I would have put her at a two, apart from the listening task, since she doesn't quite answer the task, and since the assessors have to "pull" so much information out of her.

Conversely, none of the teachers who awarded the candidate a 4, and most of whom mainly or only teach VSP students, mentioned the criterion "Addressing task or problem statement" at all. One reason for this may be that they put more emphasis on language features in their assessment. A quote from Informant No. 24 supports this interpretation:

> So I'm not so concerned with whether they have necessarily acquired so much factual knowledge and societal aspects. I consider myself more of a language teacher in my English lessons, rather than a teacher of cultural studies.

As I will return to below, this suggests that there is a difference between the teachers in how they regard the importance of content knowledge.

Discussion

In response to the two research questions, then, this investigation has found variability in the way teachers understand the constructs and criteria to be tested and what kind of criteria they see as salient. In addition, it has found variability in scoring behavior. As for the teachers' notions of constructs and criteria, all the informants reflect an understanding of the two constructs that can be identified in the subject curriculum, namely, "Communication" and "Content." However, they view the relative importance of these two constructs somewhat differently. The VSP teachers have a tendency to put more emphasis on "Communication," and particularly "Linguistic competence," whereas the GSP teachers see "Communication" and "Content" as being more juxtaposed. For example, it was highly conspicuous the way the GSP teachers penalized the student in the video-clip for not answering the topic question and not reflecting sufficiently on the issues under discussion. Assuming that GSP students are on average more proficient in English than VSP students, one may infer that the GSP teachers are used to focusing more on "Content" because of the higher level of proficiency of their regular students. Such a conclusion supports the research results mentioned above, which have indicated that raters focus more on linguistic features at lower levels and pay more attention to content at the higher levels of proficiency (Brown et al., 2005; Pollitt & Murray, 1996; Sato, 2012)

Beyond this, the present study confirms the findings reported by Brown et al. (2005) in that the teachers largely focus on the same overall features of performance, but that there is some variation in the way that they attend to the more narrow features. For instance, all the teachers mention phonology as a criterion that should be heeded, whereas only two mention the ability to repair mistakes. Of course, this does not necessarily mean that only these two informants pay attention to a student's ability to repair, but it suggests that it is seen as a less salient performance criterion.

When it comes to the three categories "Disruptive features," "Effort," and "Preparation," they did not correlate well with the overall constructs "Communication" and "Content." Actually, one may question their status as criteria to be tested. The first one of these, "Disruptive features" is not unambiguously a criterion in the sense of "aspect of performance" as defined above. Rather, it is an *effect* of the failure of a student to perform well on other criteria, like, for example, "Linguistic competence." In other words, if a student cannot pronounce a word correctly, this may disrupt communication. Still, several informants appeared to treat it as a criterion, and it is actually included in the written rating scale for Østfold county.

As for "Effort," it is not uncommon that raters pay attention to such a feature, even in contexts where it is not included in the construct to be tested (Brown, 1995; May, 2006). In the present study, this aspect may further be linked to differences in rater severity, an aspect that is also commonly found in the research literature (Bonk & Ockey, 2003; Hsieh, 2011; Iwashita, McNamara, & Elder, 2001; Lumley & McNamara, 1995). As the results showed, some teachers were rating the VSP students more leniently than the GSP students, especially the weaker students who might fail. Such a practice may be attributed to the already mentioned belief held by some teachers that it is unfair

to make VSP students take the same course, and the same exam, as the GSP students, who are supposedly theoretically stronger.

Finally, it can be argued that the category "Preparation" is not an aspect of the performance. Rather, it is an assumed *cause* of one or more aspects of the performance. What is more, just like "Effort," it is not criterion-relevant according to the national educational authorities (UDIR, 2010, pp. 13, 44). Overall, then, these findings corroborate results from earlier studies that have found that raters apply non-criterion relevant information when scoring performance (May, 2006; Pollitt & Murray, 1996; Sato, 2012; Yildiz, 2011).

Conclusion and Implications

This study has investigated what kind of performance aspects teachers pay attention to in an EFL oral exam at the upper-intermediate level where no common rating scale exists. The study found that the teachers generally have the same broad understanding of the constructs and criteria to be tested, but indicated that there is some variation when it comes to how they value the relative importance of these constructs and criteria. In particular, there is variation as regards how the teachers' view the significance of content knowledge. In addition, the study found variability in scoring outcomes.

Three important limitations of this study must be kept in mind. First, it is based on a purposeful sample comprising only 24 informants. The generalization of these results to raters in Norway, or raters generally, is therefore, of course, problematic. Second, there is the possibility that the teachers' accounts of general criteria may have been influenced by the particular student performance shown to them in the video-clip. Had there been a different performance, the teachers may have mentioned different aspects in the discussion of general criteria. Third, it may be difficult for teachers to describe the salience of individual criteria because performance is assessed holistically. Considering these limitations, it would be relevant to undertake a larger study involving a number of student performances at different levels, as well as a more sizable teacher sample, to see if the conclusions in this study could be supported. Despite these limitations, however, the findings provide important empirical evidence about teacher cognition and behavior, which may help inform the development of rating scales, test tasks, and classroom assessment practices.

The study has three major implications. First of all, it points to the problem of not having a common rating scale in a high-stakes oral L2 testing situation. As there is evidence that a common rating scale may lead to "sounder, if imperfect, inferences . . . in the process of decision making" (Fulcher, 2012, p. 379), it is likely that the introduction of a common rating scale would strengthen the validity of the score interpretations.

Second, this investigation highlights the problem of introducing a comprehensive content construct at the intermediate to upper-intermediate proficiency levels. There are indications that teachers working with lower proficiency level students downplay the content construct, despite curricular requirements, as many of their students find it difficult enough to come to grips with basic linguistic features. However, as the findings here do not warrant firm conclusions, and the assessment of content in language learning contexts is an underexplored area (Snow & Katz, 2014), it is recommended that more research be undertaken.

Third, given that many examiners in the oral exam seem to be quite concerned with students' abilities to reflect on content, it is important that classroom practices at this level involve tasks that give students the opportunity to reflect on topical knowledge. Restricting work in class to language-related exercises or the simple recounting of content will not prepare them sufficiently for an oral exam such as the one investigated here.

Appendix A

Rater Background and Scores Awarded

Rater Background Information and Scores Awarded.

No.	Age	Gender	L1	Education	Teaches at study program	Score given
1	39	Male	English	Master	Both GSP and VSP	3
2	57	Female	Norwegian	Bachelor	VSP only	4
3	57	Male	Norwegian	Bachelor	Both GSP and VSP	3
5	48	Male	Norwegian	Bachelor	Both GSP and VSP	3
6	35	Female	Norwegian	Bachelor	Mainly GSP	3
7	29	Female	Norwegian	Master	Both GSP and VSP	3
8	42	Male	Norwegian	Bachelor	Mainly GSP	4
9	41	Female	Russian	Master	GSP only	3
10	59	Male	Norwegian	Master	Mainly GSP	3
11	55	Female	Swedish	Bachelor	GSP only	2
12	28	Female	Norwegian	Master	GSP only	2
13	39	Female	Norwegian	Master	GSP only	3
14	55	Male	English	Master	VSP only	3
15	36	Female	Finnish	Master	GSP only	3

(continued)

Appendix A (continued)

No.	Age	Gender	L1	Education	Teaches at study program	Score given
16	58	Female	Norwegian	Bachelor	VSP only	4
17	38	Female	Norwegian	Bachelor	GSP only	2
18	36	Male	Norwegian	Master	Both GSP and VSP	3
19	41	Male	Norwegian	Bachelor	VSP only	3
20	54	Male	Norwegian	Master	Both GSP and VSP	4
21	35	Female	Mandarin	Master	Mainly VSP	4
22	35	Male	English	Doctor	Mainly VSP	3
23	34	Female	Romanian	Master	Mainly VSP	3
24	47	Male	Norwegian	Bachelor	VSP only	4
N = 24						

Note. GSP = general studies program; VSP = vocational studies program.

Appendix B

Interview Guide—Assessing the GSP1 (General Studies Program – 1st Year)/VSP2 (Vocational Studies Program – 2nd Year) Oral English Exam

1. Background:
   - 1.1. Age:
   - 1.2. First language:
   - 1.3. Education (English):
   - 1.4. Number of years as a teacher (upper secondary level):
   - 1.5. Experience as examiner (at the GSP1/VSP2 level):
   - 1.6. Has been teaching: GSP ___ VSP ____ Health/social ___
   - 1.7. Worked as a teacher outside your county?
   - 1.8. Attended rater training courses?
   - 1.9. Do you use a written rating scale while rating? If yes, who has developed this scale?
2. How would you assess the performance you have just seen? Which grade would you have given and why? In other words, which criteria would you have applied in the assessment process?
3. Are there any other criteria, which you haven't applied here, that would be relevant in the general scoring of performance in this exam?
4. Do you score analytically or holistically?
5. Do you compare students when grading?
6. What, in your opinion, does the grade reflect? General English competence, competence relating to vocational English, academic English, or what?
7. How do you understand the concept of "communication"?
8. What would it take to get a top score? What criteria are the most important?
9. Conversely, when will a student fail?
10. What about phonology? Some teachers say that a near-native speaker accent is important to get a top score? What is your comment on that?
11. Would you give credit for effort?

Acknowledgment

I would like to thank the student and the 24 teachers who agreed to participate in this study.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

Notes

1. Due to curriculum revisions in 2014, the main area "Communication" has now been divided into "Oral Communication" and "Written Communication," and minor revisions of some of the aims have been undertaken (cf. www.udir.no).
2. In the county of Østfold, in which eight of the 24 teacher informants in this study were employed, the county governor has developed a common, written rating scale to be used by all English teachers in the county.
3. In Norway, there is a total of 19 counties.
4. With the exception of the quotes from Informants Nos. 9, 14, and 23, which are verbatim, all the quotes in this article have been translated from Norwegian.

References

Ang-Aw, H. T., & Goh, C. C. M. (2011). Understanding discrepancies in rater judgement on national-level oral examination tasks. *RELC Journal*, *42*, 31-51. doi:10.1177/0033688210390226

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, UK: Oxford University Press.

Bloom, B. S., Englehart, M., Furst, E., Hill, W., & Kratwohl, D. (Eds.). (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: The cognitive domain*. London, England: Longman.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*, 89-110. doi:10.1191/0265532203lt245oa

Brindley, G. (1991). Defining language ability: The criteria for criteria. In S. Anivan (Ed.), *Current developments in language testing. Anthology series 25*. Singapore: Southeast Asian Ministers of Education Organization (SEAMEO) Regional Language Centre.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*, 1-15. doi:10.1177/026553229501200101

Brown, A. (2000). *An investigation of the rating process in the IELTS oral interview* (IELTS research reports, Vol. 3). Retrieved from https://www.ielts.org/pdf/Vol3Report3.pdf

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks* (TOEFL Monograph Series, MS-29). Princeton, NJ: Educational Testing Service.

Council of Europe. (2001). *The common European framework of reference for languages: Learning, teaching, assessment*. Strasbourg, France: Council of Europe, Language Policy Unit.

Creswell, J. W. (2013). *Qualitative inquiry & research design: Choosing among five approaches*. Los Angeles, CA: SAGE.

Cumming, A. (2009). Language assessment in education: Tests, curricula, and teaching. *Annual Review of Applied Linguistics*, *29*, 90-100. doi:10.1017/S0267190509090084

Eckes, T. (2009). On common ground? How raters perceive scoring criteria in oral proficiency testing. In A. Brown & K. Hill (Eds.), *Tasks and criteria in performance assessment: Proceedings of the 28th language testing research colloquium* (Vol. 13, pp. 43-73). Frankfurt, Germany: Peter Lang.

Fulcher, G. (2012). Scoring performance tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 378-392). Oxford, UK: Routledge.

Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. Oxon, UK: Routledge.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. Oxford, UK: Routledge.

Galaczi, E. (2014). Content analysis. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 3). Chichester, UK: Wiley-Blackwell.

Glaser, R., & Klaus, D. J. (1962). Proficiency measurement: Assessing human performance. In R. M. Gagné (Ed.), *Psychological principles in systems development* (pp. 419-474). New York, NY: Holt, Rinehart & Winston.

Green, A. (2014). *Exploring language assessment and testing: Language in action*. Oxon, UK: Routledge.

Hægeland, T., Kirkebøen, L. J., Raaum, O., & Salvanes, K. G. (2005). *Familiebakgrunn, skoleressurser og avgangskarakterer i norsk grunnskole* [Family background, school resources, and final grades in Norwegian primary and lower secondary schools]. Retrieved from http://www.ssb.no/a/publikasjoner/pdf/sa74/kap-2.pdf

Hsieh, C.-N. (2011). Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency. In *Spaan Fellow Working Papers in Second or Foreign Language Assessment* (Vol. 9, pp. 47-74). Retrieved from http://www.cambridgemichigan.org/wp-content/uploads/2014/12/Spaan_V9_FULL.pdf

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277-1288. doi:10.1177/1049732305276687

Hulstijn, J. H. (2011). Language proficiency in native and non-native speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, *8*, 229-249. doi:10.1080/15434303.2011.565844

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, *25*, 385-402. doi:10.1177/0265532208090158

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning*, *51*, 401-436. doi:10.1111/0023-8333.00160

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: Praeger.

Kane, M. (2012). All validity is construct validity. Or is it? *Measurement: Interdisciplinary Research and Perspectives*, *10*, 66-70. doi:10.1080/15366367.2012.681977

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*, 187-217. doi:10.1177/0265532208101010

Kreuter, F. (2008). Interviewer effects. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 369-371). London, England: SAGE.

Krippendorf, K. (2013). *Content analysis* (3rd ed.). Thousand Oaks, CA: SAGE.

Kvale, S. (2007). *Doing interviews* (Vol. 2). London, England: SAGE.

Kvale, S., & Brinkmann, S. (2009). *Interviews: Learning the craft of qualitative research interviewing*. Los Angeles, CA: SAGE.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174. doi:10.2307/2529310

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*, 246-276. doi:10.1191/0265532202lt230oa

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

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

May, L. A. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing*, *11*, 29-51. Retrieved from http://eprints.qut.edu.au/15747/1/15747.pdf?ev=pub_ext_prw_xdl

North, B. (2004, April 15). Europe's framework promotes language discussion, not directives [Online edition]. *Guardian Weekly*. Retrieved from http://www.theguardian.com/education/2004/apr/15/tefl6

Norwegian Directorate for Education and Training. (2010). *Rundskriv Udir-1-2010: Individuell vurdering i grunnskolen og videregående opplæring etter forskrift til opplæringsloven kapittel 3* [Circular Udir-1-2010: Individual assessment in

primary and secondary school according to the regulations of the Education Act]. Oslo, Norway: Directorate for Education and Research. Retrieved from http://www.udir.no/Regelverk/Finn-regelverk-for-opplaring/Finn-regelverk-etter-tema/Vurdering/Udir-1-2010-Individuell-vurdering/

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, *30*, 143-154.

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th language research testing colloquium*. Cambridge, UK: Cambridge University Press.

Popham, W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice Hall.

Prøitz, T. S., & Borgen, J. S. (2010). *Rettferdig standpuktvurdering - det (u)muliges kunst?* [Fair overall achievement marks—A(n) (im) possible goal?] (Report 16/2010). Oslo, Norway: NIFU (Nordisk Institutt for Studier av Innovasjon, Forskning og Utdanning) STEP (Studies in Technology, Innovation and Economic Policy).

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

Shuy, R. W. (2003). In-person versus telephone interviewing. In J. A. Holstein & J. F. Gubrium (Eds.), *Inside interviewing* (pp. 175-193). Thousand Oaks, CA: SAGE.

Simensen, A. M. (2010). Fluency: An aim in teaching and a criterion in assessment. *Acta Didactica Norge, 4*(1), 1-13.

Snow, M. A., & Katz, A. M. (2014). Assessing language and content. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 1, pp. 230-247). Chichester, UK: Wiley-Blackwell.

Solheim, T. (2009). *Opplæring i yrkesfag: Teori-praksis* [Education in the vocational subjects: Theory-practice] (Bedre skole, 4/2009). Retrieved from https://www.utdanningsforbundet.no/upload/Tidsskrifter/Bedre%20Skole/BS_nr_4-09/4328-04-09-BedreSkole-Solheim.pdf

Stoynoff, S. (2009). Recent developments in language assessment and the case of four large-scale tests of ESOL ability. *Language Teaching*, *42*, 1-40. doi:10.1017/S0261444808005399

Taylor, L. (2006). The changing landscape of English: Implications for language assessment. *ELT Journal*, *60*, 51-60. doi:10.1093/elt/cci081

Taylor, L., & Galaczi, E. (2011). Scoring validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (Vol. 30, pp. 171-233). Cambridge, UK: Cambridge University Press.

Throndsen, I., Hopfenbeck, T. N., Lie, S., & Dale, E. L. (2009). *Bedre vurdering for læring: Rapport fra "Evaluering av modeller for kjennetegn på måloppnåelse i fag"* [Better Assessment for Learning: Report from "The Evaluation of Models for Assessment Criteria for Goal Achievements in Subjects"]. Retrieved from http://www.udir.no/Upload/Forskning/5/Bedre_vurderingspraksis_ILS_rapport.pdf?epslanguage=no

Yildiz, L. M. (2011). *English VG1 level oral examinations: How are they designed, conducted and assessed?* (Unpublished master's thesis). University of Oslo, Norway.

Author Biography

Henrik Bøhn is a lecturer in English at Østfold University College, where he teaches language proficiency and English education. He is currently undertaking a PhD in language assessment.