ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 9, No. 1, pp. 16-26, January 2018
DOI: http://dx.doi.org/10.17507/jltr.0901.03

Assessing Content in a Curriculum-based EFL Oral Exam: The Importance of Higher-order Thinking Skills

Henrik Bøhn
Department of Languages and Literature Studies, University College of Southeast Norway, Borre, Norway

***Abstract*—In this study data from verbal protocols and semi-structured interviews was analysed to explore Norwegian EFL teacher raters' (*n*=10) orientations towards content in an oral English exam at the upper secondary school level, a context characterized by the absence of a common rating scale for the teacher raters. The content construct was mainly analysed in terms a subject matter dimension and a skills and processes dimension. The results indicated that the teachers were more concerned with the skills and processes dimension (e.g. analysis, reflection) than with the subject matter dimension (e.g. cultural knowledge). Moreover, their understanding of subject matter compared fairly well with the subject curriculum, despite instances of construct underrepresentation. The study points to the prominence of guidance for teacher raters in the assessment of content and to the significance of alerting students to the importance of higher-order thinking skills in language education at this level.**

***Index Terms*—EFL, ESL oral L2 assessment, content, subject matter, higher-order thinking skills**

I. Introduction

Aspects of content may be said to be involved in all language use (Bachman & Palmer, 2010, p. 41). Despite this, the role of content in second and foreign language instruction and assessment varies substantially from context to context. Historically, it has been considerably downplayed in language assessment, as the primary focus has been on the evaluation of language features. In fact, in some cases it has even been treated as a potential source of language bias (Douglas, 2000, p. 2). More recently, however, the assessment of content has been emphasized in a number of settings, for example in content-based instruction and specific purposes courses (Byrnes, 2008; Snow & Katz, 2014). Overall, language instruction and assessment may thus be regarded as a continuum from language-driven approaches to content-driven approaches (Met, 1998).

The concept of content is somewhat elusive, however. In the language assessment literature it has been related to, or used synonymously with, as diverse terms as 'subject matter', 'cultural knowledge', 'ideas', and 'framing', to mention a few (Bachman & Palmer, 2010 p. 41; Brown, Iwashita, & McNamara, 2005, p. 27; Kratwohl, 2002, p. 213). As for expected test taker response in the content area, the concept has been linked to performance features such as '[task] fulfilment', 'description', 'explanation', 'accuracy', 'elaboration', and 'development' (Bachman & Palmer, 2010, p. 218; Brown, 2000, p. 68; Douglas, 2010, p. 117; Eckes, 2009, p. 48; Frost, Elder, & Wigglesworth, 2012, p. 349). In other words, the concept is multifaceted and complex, and there is evidence that it is not well understood in all contexts (Frost et al., 2012). More research has therefore been called for (Snow & Katz, 2014).

The context of the present study is curriculum-related English as a Foreign Language (EFL) education at the upper-intermediate proficiency level (Common European Framework of Reference, level B1/B2) in Norway. This may be said to belong to the middle of the language-content continuum. A defining feature of this context is the lack of a common rating scale, or scoring rubric, to guide teachers in their assessment of oral performance. Little empirical evidence exists to describe how content is assessed in such settings. As knowledge of which performance aspects teacher raters attend to is important for the validity of the scoring outcomes (Bejar, 2012), as well as for understanding potential washback effects relating to what teachers may prioritize in the language classroom, this study looks more closely at *Norwegian EFL teachers' unguided orientations towards content* in an oral English exam at the upper-intermediate proficiency level.

II. Literature Review

The research literature on EFL/ESL raters' unguided orientations towards aspects of content in spoken performance is very scarce. This may come as no surprise as language tests are rarely, if ever, used without rating scales. Only four such studies were identified in the literature search for the present article. The first one is Pollitt and Murray (1996), which investigated five trained examiners' general assessment of different speech samples (*n*=5) taken from the Cambridge Certificate of Proficiency in English test. This is a non-curriculum based, high proficiency level oral examination. The results did not provide very elaborate descriptions of the examiners' assessment of content, but

indicated that the raters paid more attention to content, or 'what' was being said, at the higher levels of performance, whereas they attended more to linguistic features and associated notions of 'correctness' at the lower levels of performance.

The second study which was identified, Brown et al. (2005), examined the rater focus of 10 English for Academic Purposes (EAP) specialists' in a pilot Test of English as a Foreign Language (TOEFL). The TOEFL test is an advanced level proficiency test used as an entry requirement to a number of English-medium universities. The authors used verbal-report methodology to uncover which performance features the raters paid attention to in 40 audio-recorded spoken student performances. The results showed that content was a major focus in the raters' reports (together with linguistic resources, phonology and fluency). More specifically, content was associated with the following three performance aspects:

(i) *Task fulfillment*, i.e. the degree to which the test takers were 'on topic' or 'addressed the question'.
(ii) *Ideas*, e.g. in terms of 'relevance' (cf. Table I, below)
(iii) *Framing*, i.e. 'the generic structure of test-takers responses' in terms of an introduction and a conclusion'
(Brown et al. 2005, pp. 27-30).

The third study, Yildiz (2011), used interviews and open survey questions to examine both the exam format of an oral English examination at the upper secondary school level in Norway, as well as the general rater orientations of 16 teachers involved as judges in this exam. In other words, the context of the investigation was the same as for the present study. Overall, the study found that the teachers were concerned with five general criteria: (i) 'Language competence'; (ii) 'Communicative competence', (iii) 'Subject competence'; (iv) 'Ability to reflect and discuss independently'; and (v) 'Ability to speak freely and independent of manuscript'. In terms of the content-related criteria here, i.e. 'Subject competence' and 'Ability to reflect and discuss independently', the study found that 'understanding' and '[ability to] use the knowledge that [the students] have in a relevant manner' was seen as important by the teachers (p. 55). As an example of ability to use knowledge, the teachers mentioned ability to discuss an unknown text with a topic similar to a syllabus text.

The fourth and final study identified in the present review, Bøhn (2015), also investigated unguided rater orientations in an oral English exam at the upper secondary school level in Norway. In the study semi-structured interviews were used to explore the general assessment focus of 24 teacher raters. The results showed that content was one of the aspects which caused the most variability among the teachers. Overall, the study showed that the majority of the teachers saw content as consisting of the following four aspects:

(i) addressing task or problem statement;
(ii) elaborated response;
(iii) content structure;
(iv) a Bloom-like taxonomy of reproduction, comprehension, application, analysis and reflection.

A comparison of the four studies shows several similarities. In Table I the major features of three of them have been listed. (Since Pollitt and Murray's (1996) study provided little conceptual information on the content construct, it has been left out of the table.)

TABLE I.
RATERS' UNDERSTANDING OF CONTENT IN BROWN ET AL. (2005), YILDIZ (2011) AND BØHN (2015)

Brown et al. (2005)	**Yildiz (2011)**	**Bøhn (2015)**
- *Task fulfilment*		- *Addressing task or problem statement*
- *Ideas* - amount of speech produced - response to functional demands of task - sophistication (independent) - relevance	- *Understanding* - *Ability to apply knowledge*	- *Elaborated response* - *Bloom-like taxonomy of* - reproduction - comprehension - application - analysis - reflection
- *Framing*		- *Content structure*

The comparison of findings illustrated in Table I shows that Task fulfillment in Brown et al.'s study parallels Addressing task or problem statement in Bøhn. Moreover, Brown et al.'s notion of Ideas, understood as 'amount of speech produced' has affinities with Bøhn's concept of Elaborated response. Similarly, Brown et al.'s concept of Ideas, understood as 'sophistication', intersects with Yildiz' categories Understanding and Ability to apply knowledge, and Bøhn's notion of the Bloom-like taxonomy. The logic behind this argument is that the raters' application of the Bloom-like taxonomy, which is apparent in both Yildiz and Bøhn, may be said to reflect a continuum of sophistication going from 'reproduction' at the lowest level to 'analysis' and 'reflection' at the highest level. Finally, the notion of Content structure in Bøhn has similarities with Framing in Brown et al.

Another interesting similarity between these studies is the fact that the categories developed mainly describe the skills or abilities involved in the handling of content. Except for the general references to 'task', 'ideas' and 'problem statement', very little is said about *what* should be tested. In Pollitt & Murray's (1996) and Brown et al.'s (2005) studies this may come as no surprise. As the test takers were responding to proficiency tests (Douglas, 2010, pp. 1-2), they were

not judged on their knowledge of specific EAP topics, but rather on their language abilities and their capacity to handle whatever topic was under discussion. However, in Yildiz (2011) and Bøhn (2015) the oral exam used as a basis for the inquiries may be classified as an achievement test, as it is based on a subject curriculum (Bachman & Palmer, 2010, p. 213). Consequently, one may expect raters in such settings to comment on specific subject matter. No such comments were reported, however. Given that content is such an elusive concept, it is worth looking more closely into EFL teachers' understanding of this construct.

Against this background the present study investigates Norwegian EFL teachers' orientations towards subject matter content in an oral English exam at the upper-intermediate proficiency level. As part of the investigation the teachers' orientations will be compared with the aspects of content identified in the English subject curriculum.

III. Analytical Framework

As was touched upon in section one, the concept of content is used in a number of different ways in the assessment literature. Traditional language assessment theory gives the concept scant treatment, but where mentioned, it is often related to Bachman and Palmer's (1996) model of communicative language ability (e.g. Douglas, 2000; 2010; Fulcher & Davidson, 2007; Green, 2014; Luoma, 2004; McNamara, 1996). In this model content is referred to as *topical knowledge* or *real-world knowledge* and defined loosely as 'knowledge structures in long-term memory' (Bachman & Palmer, 1996, p. 65). Although Bachman and Palmer offer interesting perspectives on the testing of content, their focus is nevertheless predominantly on language aspects. Therefore, in order to broaden the analytical framework, I will briefly turn to the field of curriculum-related, content-based L2 language instruction, which provides richer theoretical support for the subsequent analysis.

In content-based instruction (CBI) the overall objective is typically to teach 'subject-specific curricular content', such as mathematics, social science or language arts, alongside a second language, in order to help students develop L2 for communicative purposes and to 'access academic content' in regular subject classes (Snow & Katz, 2014, p. 230). In this sense CBI belongs to the content side of the above mentioned language-content continuum. Exactly what the subject-specific or academic content is will depend upon the nature of the subject field, and it is of course impossible to make an inventory of all the different subject matter issues that may be treated. Overall, however, one may find content described in terms of words such as *facts*, *concepts*, *laws*, *principles* and *theories* (Chamot, 2009, p. 239). The reference to 'concepts' here is particularly noteworthy, as Chamot claims that '[c]ontent subject concepts and relationships are the *foundation* of academic knowledge' (p. 20, emphasis added). In passing, it is also worth observing that Chamot lists a number of skills and abilities needed to process subject matter content. Stressing the importance of teaching higher-order thinking skills, she claims that students should be encouraged to speculate, predict, synthesize and make judgements about the material they are learning, 'rather than merely recall facts' (p. 30).

Another relevant feature of curriculum-related CBI in this discussion is its focus on content standards (Chamot, 2009; Echevarría, Vogt, & Short, 2008; Snow & Katz, 2014). Content standards specify learning outcome objectives, which state what students should know and be able to do in relation to some defined subject matter content (Chamot, 2009, p. 16). As Kratwohl (2002) explains,

> statements of objectives typically consist of a noun or noun phrase – the *subject matter content* – and a verb or a verb phrase – the *cognitive process(es)*. Consider, for example, the following objective: The student shall be able to remember the law of supply and demand in economics. (p. 213, italics added)

Learning objectives like the one Kratwohl mentions here offer a suitable framework for the assessment of content because they provide tools for identifying both *what* to assess (i.e. subject matter content) and *how* to assess it (i.e. the range of cognitive processes, or 'skills' or 'abilities' involved). In order to further specify this, it is relevant to briefly consider Bloom's taxonomy, which has been used as a basis for the development of content standards in many contexts, and which is frequently drawn upon in CBI (Chamot, 2009; Echevarría et al., 2008; Kratwohl, 2002). Here I will concentrate on the revised version of the taxonomy.

Bloom's revised taxonomy arranges learning outcome objectives in a two-dimensional grid (Anderson & Kratwohl, 2001). One dimension represents different types of knowledge (the *what*-aspects), and the other represents various types of cognitive processes (the *how*-aspects). The organization of knowledge and cognitive processes along dimensions is meant to demonstrate their hierarchical nature, from the simpler forms of knowledge and processes to the more complex. According to Kratwohl (2002), knowledge is related to subject matter content and can be divided into four types: factual, conceptual, procedural and metacognitive. Cognitive processes fall into the following six categories listed from the simple to the complex: remember, understand, apply, analyse, evaluate and create. Figure 1 illustrates how the two dimensions interrelate and places Kratwohl's example in this grid.

The cognitive process dimension

The knowledge dimension	*Remember*	*Understand*	*Apply*	*Analyse*	*Evaluate*	*Create*
Factual knowledge						
Conceptual knowledge	Remember the law of supply and demand					
Procedural knowledge						
Metacognitive knowledge						

Figure 1. Example of a learning objective represented in Bloom's taxonomy table, as outlined by Kratwohl (2002).

In summary, the theoretical frameworks discussed in this section all point to important ways in which subject matter content can be understood. Both Chamot's framework and Bloom's revised taxonomy seem to be highly relevant, as they relate to curriculum-based contexts. More specifically, Kratwohl's description of subject matter in terms of nouns and noun phrases provides a particularly useful tool for analysing content in the present study.

IV. Research Question

Against the empirical findings and theoretical frameworks presented in the preceding sections the present study addresses the following research question: *What do EFL teachers at the upper secondary school level in Norway perceive as relevant subject matter content to be assessed in the GSP1/VSP2 oral English exam?*[1] As part of this investigation, the teachers' orientations will be compared against the aspects of subject matter identified in the subject curriculum.

V. The Context of the Study

In Norway, English is a compulsory subject for all students from the first grade onwards (age six). By the time the students start upper secondary school at the age of 16, they have on average reached an upper-intermediate proficiency level (CEFR B1/B2). The subject curriculum is centered on competence aims, which define what students are expected to master at the end of the different levels of instruction. Grades are mainly given in the form of overall achievement marks, awarded by each individual subject teacher on the basis of various forms of classroom assessment. In addition, around 20 per cent of the students are randomly selected to sit for a written exam, and five per cent are selected to take an oral exam. Whenever a group of students are assigned for the oral exam, their English subject teacher is required to act as an examiner. In addition, an English teacher external to the students' school is assigned the role as assessor. Grades range from 1 ('fail') to 6 ('excellent'), and performance is scored holistically.

The English subject curriculum, which works as a framework for the operationalization of the constructs (Bachman & Palmer, 2010, p. 211), stipulates a number of competence aims relating to subject matter content. At the level under investigation here, the GSP1/VSP2 level, 10 aims explicitly address content. These aims have been listed in Table II, with a corresponding description of the subject matter content as defined above.

[1] This upper secondary school exam is taken by students in their first year in the general subjects programme (GSP1) or by students in their second year in the vocational subjects programmes (VSP2).

TABLE II.
SUBJECT MATTER CONTENT SPECIFIED IN THE ENGLISH SUBJECT CURRICULUM (GSP1/VSP2 LEVEL)[2]

Competence aim	**Subject matter content**
Assess and use different situations, working methods and learning strategies for developing one's English skills	Metacognitive strategies
Assess different digital resources and other aids critically and independently and use them in one's own language learning	Resources
Understand the main content and details of different types of oral texts about general topics and subject-specific topics related to one's own study programme	General topics; subject-specific topics related to study programme
Discuss cultural and societal conditions in a number of English-speaking countries	Cultural and societal conditions in English-speaking countries
Present and discuss current news from English-speaking sources	Current news topics from English-speaking sources
Discuss the development of English as a world language	English as a world language
Discuss different types of English-speaking texts from different parts of the world	English-speaking texts
Discuss English-speaking films and other cultural forms of expression from different parts of the world	English-speaking cultural forms of expression
Discuss texts by and about indigenous peoples in English-speaking countries	Texts by and about indigenous peoples
Select an in-depth study topic of one's own study programme and present this	Subject-specific topic related to study programme

As can be seen in Table II, there is a very broad range of subject matter aspects. Not only are the students expected to handle a number of topics related to the English-speaking world, such as literary texts, cultural conditions and indigenous peoples, they are also expected to be able to understand the content of subject-general texts and subject-specific topics related to their own study programme. In addition, they are also required to know and to assess both metacognitive strategies and (re)sources.

As for the operationalization of the constructs to be assessed, there is a notable difference between the oral and the written exam. Whereas the written exam is administered nationally by the Norwegian Directorate for Education and Training, the oral exam is controlled by the local education authorities in each of the 19 counties. This means that, for the written exam, there is a nationally developed rating scale and nationally designed test tasks, whereas for the oral exam, different types of locally developed scales and tasks exist. As a rating scale can be understood as an operationalization of the constructs to be assessed (Fulcher, 2012, p. 378; Luoma, 2004, p. 59), this means that the constructs are operationalized differently in the various counties (see e.g. Bøhn, 2015).

VI. METHOD

A. Research Design

The study used data from two sources of evidence: verbal protocols and semi-structured interviews (Brinkmann & Kvale, 2015; Green, 1998), involving 10 EFL teachers at the upper secondary school level in Norway. A prompt in the form of a video-taped performance of a student taking the oral exam was used as a stimulus for the generation of verbal protocols. On the basis of this video-clip the teachers were asked to comment on the performance in real time (concurrent verbal reporting) and to give it a score. Directly after the protocols had been recorded, the teachers were interviewed by the researcher on their conceptions of content in the oral English exam. Both data sets were analysed using provisional coding (Miles, Huberman, & Saldaña, 2014)

B. Participants

The teachers were recruited for the study through purposeful sampling (Creswell, 2013), in order to obtain variation in the sample with regard to school and county background, teaching and rater experience and study programme affiliation. The teachers were contacted directly by telephone, and all who agreed to participate did so on a voluntary basis, with no financial compensation. The participants were between 32 and 51 years of age (M=40), and their teaching experience ranged from one and a half to 26 years. They represented six different schools in three different counties. Three of them worked only in the vocational studies programmes (VSP), three worked only in the general studies programme (GSP) and four worked in both programmes. All of them were fully qualified teachers and had previously been involved as examiners.[3]

As for the video-taped prompt, a VSP student agreed to be filmed as she was taking her oral exam. The exam format consisted of three tasks: (i) a pre-planned monologue task in the form of a presentation, followed by a discussion of the presentation; (ii) an oral interview task based on a short story from the syllabus; and (iii) an oral interview task based on a listening comprehension sequence. For the pre-planned monologue task the student had been given 48 hours in advance to respond to the following prompt:

[2] The English subject curriculum can be accessed at http://www.udir.no/kl06/ENG1-03?lplang=eng.
[3] Further information about teacher background can be retrieved from http://www.fag.hiof.no/~heb/PhDArt3AppendixA.pdf.

Choose a common health issue in today's society and make a presentation of the problems it causes the individual and in society. Use examples from fictional and factual texts as well as films from your reading list to illustrate your examples.

The student had chosen to give a presentation about HIV/AIDS in South Africa. As regards the topics of the other two tasks, the short story focused on obesity and eating disorders, and the listening comprehension sequence involved a discussion about English a world language.

C. Procedure

An interview guide was piloted and revised. The questions in the interview guide were formulated on the basis of the findings in Bøhn (2015), the analytical framework presented above and the content-related statements identified in the English subject curriculum (cf. Table II, above).[4] The verbal protocols were generated by the teachers in individual think-aloud sessions (Green, 1998). The video-clip was shown to the participants on a lap-top computer, and a headset was provided in order to ensure good sound quality. Before the recording started, the teachers were instructed to verbalize their thoughts on the *general* aspects of the performance and then to give it a grade. They were also given five minutes to familiarize themselves with the equipment and the procedure. All the teacher comments were recorded on an Olympus DM-450 digital voice recorder.

Immediately after the think-aloud sequence, the teachers were interviewed on their judgments of the performance they had just seen, as well as on their assessment orientations more broadly. In the first half of each interview only open, 'nondirective' questions (Yin, 2016, p. 144) concerning general assessment criteria were asked, in order not to impose researcher-generated conceptions of content on the participants (see questions B1-3 in the interview guide, footnote 4). Thus, it was hoped that 'unsolicited' answers regarding content would emerge. Subsequently, the teachers were questioned specifically on whether and to what extent they considered content while rating. This included questions concerning what they regarded as content, how they thought it should be evaluated, and to what extent they found the subject matter identified in the curriculum to be relevant in the assessment of oral exam performance (see question B4-10 in the interview guide, footnote 4).

D. Data Analyses

After the verbal protocols had been recorded I transcribed, checked and segmented them (cf. Green, 1998). In the segmentation process the transcripts were divided into ideas units. An ideas unit can be defined as 'a single or several utterances with a single aspect of the event as the focus', i.e. a unit which is 'concerned with a distinct aspect of performance' (Brown et al., 2005, p. 13). The following excerpt, divided into five units (separated by '/'), serves as an illustration:

/Good vocabulary / She corrected herself. There was an error there / There was a Norwegian word there / She is doing well in terms of content / Here her pronunciation is not that good […] There were some long words… loan words/

All the segments were then coded into categories, using the computer software package QSR NVivo 10. The transcripts were coded in two cycles (Saldaña, 2013). In the first cycle all the segments were assigned codes, using provisional coding based on the categories developed in Bøhn (2015) and the conceptual framework presented in the analytical framework section, above (cf. Miles et al., 2014; Saldaña, 2013). For example, the segments in the above quoted excerpt were coded as Vocabulary, Ability to repair, Compensatory strategies, Content and Pronunciation, respectively. After all the statements had been coded, the codes relating to content were sifted out and re-analysed in a second cycle in order to validate these categories. In this cycle ideas units were specifically checked for nouns and noun phrases relating to content, as specified by Kratwohl (2002) (cf. Analytical framework section). For example, in one statement, the following noun phrase occurred: 'She remembers the syllabus, so she has studied' Here, the noun phrase the syllabus was categorized as a subject matter item. More specifically, the unit was coded in the category 'Syllabus texts' (cf. Table III, below). In order to further validate the analysis, a colleague with prior experience as an EFL teacher at the upper secondary level was asked to code two transcripts, using provisional coding. The inter-coder reliability analysis yielded a Kappa estimate of .83, which may be regarded as very good (Landis & Koch, 1977).

The interviews were also analysed using provisional coding. First, they were transcribed and checked and then divided into two sections corresponding to the unsolicited and solicited answers to the open and specific questions that had been asked (cf. Procedure section). Next, these two sections were divided into ideas units, in a process similar to the verbal protocol analysis (VPA), and again the transcripts were coded in two cycles. In the first cycle the idea units were compared against the analytical framework developed in Bøhn (2015) and in the analytical framework section, whereas in the second cycle the content segments were separated out and analysed with a particular focus on nouns and noun phrases. The following extract gives an illustration (the segments have been separated by '/'):

Researcher: *How would you define content?*

Informant: / *First of all, that she answers the task, and that it is an answer which is relevant to the task / that it is an answer which shows that she has knowledge of English-speaking countries and English-speaking literature, something which this student doesn't have at all* /

[4] The interview guide can be accessed at http://www.fag.hiof.no/~heb/PhDArt3AppendixB.pdf.

Here the first ideas unit contained the content-related noun phrases *the task*. It was coded as 'Task / Topic statement' (cf. Table IV, below). The second ideas unit comprised the noun phrase *knowledge of English-speaking countries and English-speaking literature.* However, as 'knowledge' in Kratwohl's (Bloom's) framework relates to the process dimension of learning objectives rather than to the subject matter dimension, this noun phrase head may be excluded. We are then left with the two noun phrases *English speaking countries* and *English-speaking literature*. These two phrases were coded as 'Knowledge of culture and literature in the English-speaking world' (cf. Table IV). In order to validate the coding, the above mentioned colleague agreed to analyse another two transcripts. The inter-coder consistency between my own coding and hers resulted in a Kappa estimate of .78, which may be regarded as substantial (Landis & Koch, 1977).

VII. Results

A. Results from the Verbal Protocol Analysis

The VPA regarding relevant subject matter content to be assessed produced five specific subject matter categories, in addition to a general one (cf. Table III, below). The first one, which comprised a number of statements from all the participants, was labelled *Task / Topic statement.* This category reflects the fact that the teachers mainly commented on subject matter in relation to the three exam tasks: the presentation about HIV/AIDS in South Africa, the discussion of the text from the syllabus about eating disorders and the listening comprehension task about English as a world language. Hence, a large proportion of the statements were simply references to 'HIV/AIDS', "South Africa', 'symptoms', 'obesity", 'English around the world', 'accents' and the like. Similarly, the teachers used a number of general descriptions such as 'topic', 'theme', 'problem statement' and 'concepts' to refer those task-related issues. Three statements illustrate this:

She is reflecting a bit on <u>the consequences of HIV and AIDS</u> and the fact that there is no proper cure. (Informant no. 10)

She shows understanding of <u>the complexities of eating disorders</u>. (Informant no. 3)

She's managed to demonstrate that she understood <u>some of what she has listened to</u>. (Informant no. 5)

None of the other categories comprised nearly as many statements as Task / Topic statement. The second one, termed *Sources*, was commented on by six teachers. This category was found to be related to Task / Topic statement, but it was singled out as a separate category. The reason for this was that the teachers seemed to expect the student in the video-clip to reflect, or at least to comment on, the sources of her presentation. Hence, this analysis is consistent with the other categorizations made here, considering that Sources is realized by a noun and would fit neatly into Bloom's taxonomy table as presented in Figure 1. A quote from informant no. 6 illustrates this point: 'She doesn't say much about her <u>sources</u>'. Comments like this one were only made in relation to the presentation task.

The third category, labelled *Personalized knowledge*, was developed from three teacher statements which pointed to the fact that the student in the video-clip related the topic of task three to personal experiences:

[*She was asked*] *a question about whether she speaks English outside of Norway… She communicates o.k. when she speaks freely. That's quite common… when they are allowed to speak about <u>what they want</u>, they usually do o.k.* (Informant no. 8)

As may be observed, the underlined content aspect in this extract is not represented by a noun phrase, but rather by a nominal relative clause (Hasselgård, Lysvåg, & Johansson, 2012). However, such clauses have syntactic functions similar to noun phrases, and in this case it is seems clear that it denotes subject matter. The statement suggests that speaking about personal experiences is a type of subject matter which is sometimes seen as relevant by Norwegian teachers. Four informants made comments of this kind.

The fourth category, *Syllabus texts*, indicates that some teachers seem to expect students to remember, and possibly to reflect on, texts from the syllabus. As the informant quoted in the methods section put it, 'She remembers the syllabus' (cf. above). Again, this is an example which fits in Bloom's taxonomy table presented in Figure 1.

The fifth category, termed *Knowledge of culture and literature in the English-speaking world*, suggests that some teachers expect students to be able to refer to culture-specific issues in their responses to the exam tasks. Commenting on the student's response to exam task number two regarding English as a world language, one informant said:

[*A*] *really good student would jump on that question and talk about <u>different values</u> and… how some people look up to posh accents, or might look down at another. But she's not at all in that category of students.* (Informant no. 5)

Only informant no. 5 made comments which were coded in this category, however. Finally, a general content category emerged, comprising statements where the teachers merely referred to 'content'. In Table III, all the categories from the verbal protocol analysis have been listed, together with an example for each category.

TABLE III.
SUBJECT MATTER CATEGORIES DEVELOPED FROM THE VERBAL PROTOCOL DATA

Content category	**Example**
Task / Topic statement	*It is good that she is able to reflect, at least a little bit, on the task.* (Informant no. 7)
Sources	[*There are*] *sources* [*in the last PowerPoint slide*] *... which are only URLs, nothing more.* (Informant no. 4)
Personalized knowledge	*But she is telling an interesting story here about a friend… with an eating disorder*. (Informant no. 3)
Syllabus texts	*She remembers the syllabus. So she has studied.* (Informant no. 3)
Knowledge of culture and literature in the English-speaking world	[*A*] *really good student would jump on that question and talk about different values and… how some people look up to posh accents.* (Informant no. 5)
Content – general	*It seems that she doesn't know the content very well.* (Informant no. 8)

B. Results from the Interview Analysis

The analysis of the answers to the unsolicited questions in the first part of the interviews yielded no additional subject matter content categories. As the teachers had only been asked to explain their general assessment orientations in this sequence, they mainly reiterated the aspects of content which they had reported in the VPA. However, when they were asked specific questions regarding content in the second part of the interview, other and more nuanced aspects emerged. The following exchange between the researcher and informant no. 3 serves as an example:

Researcher: *How do you understand content?*

Informant: *That you have understood some concepts and relationships and are able to show me that you understand by explaining on the basis of texts and examples. And in order to be able to do that, you obviously have to remember some texts and be able to remember some facts and stuff, but it's not particularly important that you have remembered the exact year* [*of an event*] *or the full name and title of a king, or an author or the like* […] *And I would like you to know that it is Hemingway who has written the short story, but I would much rather that you really understand… why Nick Adams acts like he does, and why he has that relationship to his father… and if possible compare with another short story*. (Informant no. 3)

Here the informant relates relevant subject matter content to *concepts*, *relationships*, *texts*, *facts* and 'literary topics' and exemplifies with reference to a literary text from the syllabus. In the analysis, 'concepts' was coded as a separate category labelled *Concepts*, whereas 'texts' was classified as *Syllabus texts.* 'Literary topics' was coded in the category *Knowledge of culture and literature in the English-speaking world* (cf. Table IV). Interestingly, the emphasis placed on 'concepts' and 'relationships' is an echo of Chamot's (2009) claim that the ability to understand concepts and to see the relationship between them is the foundation of academic knowledge (cf. Analytical framework section). However, as the idea of 'seeing relationships [between concepts]' may have more in common with the process aspects (understand, analyse, evaluate etc.) than with the subject matter aspects of content, it was decided not to place "relationships" in a separate category in the analysis.

Similarly, 'facts' was not coded as a separate category, as the data also contained answers to a question concerning how the teachers perceived the notion of 'facts' (cf. the interview guide, Appendix B). Thus, it was hoped that more explicit features of subject matter could be discerned. In response to this question, one informant answered: 'Well, I think of general knowledge' (informant no. 2). This view that general knowledge is part of the content construct was supported by seven other informants. Hence, a separate category labelled *General world knowledge* was included in the analysis (cf. Table IV). In addition, on the question of facts, another informant explained:

Well, if you look at the English subject curriculum, there is no list of facts that you have to remember; absolutely not. You don't have to know that Sydney is the capital of Australia (sic) *in order to pass in English* […]. *But if you get that task, you are expected to find some information about Australia.* (Informant no. 5)

As informant no. 5 quite correctly points out, the curriculum does not list any facts that students must remember. For him, this seems to mean that subject matter largely relates to the information that that the student has collected in preparation for the presentation task (task number one). Accordingly, it appears that this task is seen as central for the *what* to be tested. Another informant also mentioned the fact that detailed subject matter aspects are absent from the curriculum:

I had a student in an oral exam once who didn't know anything about the Tea Party [*Movement*]*… and there is nothing* [*in the curriculum*] *about the Tea Party in the U.S. But he had to know something. Exactly what that 'something' is* […] *isn't so important. But it has to be something. And what he or she shows... has to be thoroughly done… and be at a certain level… not just surface level knowledge.* (Informant no. 4)

In this extract informant no. 4 does not mention the centrality of task one, but rather alludes to the very general and wide-ranging content aspects of the curriculum. A consequence of this appears to be that the *what*-aspects to be presented are seen as less important. What matters is *how* subject matter is presented. Elaborating on this point, she explained:

But then I also think that… every now and then… we are assessing general maturity. […] *… content… how much do they actually understand of the world around them? And what is kind of… more a type of general intelligence or general knowledge, which is perhaps not always linked to the English subject.* (Informant no. 4)

Here, the formulation 'general knowledge, which is perhaps not always linked to the English subject' suggests that any topic is potentially relevant for discussion. Moreover, the use of the phrase 'actually understand' again points to an emphasis on skills and processes. This view was, in fact, supported by several other teachers. For example, in a response to my question on whether good general knowledge could help improving a student's score, informant no. 8 replied: 'Yes… In fact, I would put it in the category "Having the ability to reflect"'.

Finally, as regards the relevance of the specific content issues identified in the curriculum (cf. Table II), all the teachers confirmed that the aspects listed there, except for Metacognitive strategies, were relevant features to be tested. However, the issue of Metacognitive strategies left most teachers hesitant. One teacher even categorically denied they were to be tested in the exam, calling such strategies 'a meta-science'. Only one teacher clearly affirmed that these strategies were a relevant part of the content construct. That being said, it should be emphasized that the teachers did not appear to expect students to respond impromptu to detailed questions concerning all of these issues. Rather, they emphasized the importance of being able to analyse, reflect on and evaluate whatever subject matter that the task or question addressed.

In Table IV all the subject-matter categories which emerged from the interview analysis have been listed. As has been mentioned, the first one of these, Task / Topic statement was by far the largest one. Some, such as News topics, Syllabus texts and particularly Metacognitive strategies, were rather marginal.

TABLE IV.
SUBJECT MATTER CONTENT SPECIFIED IN THE ENGLISH SUBJECT CURRICULUM (GSP1/VSP2 LEVEL)

Content category	**Example**
Task / Topic statement	*Well, I* [*think of content as*] *her focusing on the theme that she has been given and actually talks about this topic.* (Informant no. 7)
Sources	*Content* [*relates to the student's ability to*] *use some sources… Because she doesn't say anything about that either.* (Informant no. 2)
Personalized knowledge	*But she is telling an interesting story here about a friend… with an eating disorder*. (Informant no. 3)
Knowledge of culture and literature in the English-speaking world	[*Content relates to the fact that*] *she has knowledge of the English-speaking world and of English literature.* (Informant no. 6)
English as a world language	*She didn't get the chance to sort of talk about the English language, as a world language and international language.* (Informant no. 4)
General world knowledge	*You will get a better grade if you have good general knowledge of the world.* (Informant no. 6)
Concepts	[*Content means*] *that you have understood some concepts* […] (Informant no.3)
Indigenous peoples	*We could have asked a question like: "Have you learnt anything about indigenous peoples?"* (Informant no. 5)
News topics	[*Students should have*] *the ability to reflect on* [news topics from] *Fox news, right… those kinds of things.* (Informant no. 4)
Syllabus texts	*If I have taken some texts* […] *from the syllabus* […] *and they don't know anything about them* […] *then they are in trouble, I'd say* (Informant no. 9)
Metacognitive strategies	[Interviewer:] *Does this mean that learning strategies could be tested?* [Informant:] *Yes, it's there* [*in the curriculum*], *isn't it?* (Informant no. 1)
Content – general	*I think part 1* [*task 1*] *is good in the sense that she shows good knowledge* (Informant no. 4)
Miscellaneous	*Content in English* […] *that's an inexhaustible field.* (Informant no. 4)

One final comment is worth making. It is interesting to observe the apparent discrepancy between the general agreement on certain assessment criteria – e.g. that answering the task is an important criterion – and the occasional disagreement on what kind of performance that is characteristic of a given level in relation to a criterion. For example, informant no. 2 reported in her verbal protocol: 'She doesn't mention film or literature at all. She is not answering the whole task'. In a response to this remark, which was presented to her in the interview, informant no. 10 replied: 'Not answering the task? Well of course she does!' In other words, the informants do agree that answering the task is important, but they do not agree on what kind of performance is indicative of task fulfillment.

VIII. DISCUSSION

Overall, in response to the research question *What do EFL teachers at the upper secondary school level in Norway perceive as relevant subject matter content to be assessed in the GSP1/VSP2 oral English exam?* the analyses showed that the teachers understand subject matter in very general terms. They confirmed that the aspects listed in the subject curriculum, apart from meta-cognitive strategies, are relevant features to be tested, but as a number of these aspects are very wide-ranging, the teachers pointed out that it is unrealistic to expect students to remember details from all kinds of potential topics. Consequently, they appear to adopt an assessment strategy where the testing of skills and processes (describing, analysing, evaluating etc.) becomes more important than the assessment of clearly defined subject matter. Simply put, the specifics of the subject matter are notably downplayed. This supports the finding in Bøhn (2015) which showed that the teachers' operationalization of the content construct to a large extent involved skills and abilities.

On closer inspection, the teachers seemed to focus most of their attention on subject matter in the pre-planned presentation task. However, as some of them pointed out, the material that the students have prepared beforehand, for example PowerPoint slides, is not to be tested. Rather, it is their ability to present and discuss this material which should be the focus of the assessment. Such a position is consistent with stipulations made by the national educational

authorities, which specify that whatever the students have prepared beforehand is not to be assessed (Norwegian Directorate for Education and Training [UDIR], 2014). This fact may further explain why the teachers seemed more concerned with the students' ability to present, reflect and analyse than with the specifics of what was being presented. In particular, they gave the impression of being preoccupied with the higher-order thinking skills of analysis, reflection and evaluation, as mentioned by Chamot (2009). A potential consequence of this line of thinking is that the subject matter construct is largely understood as *general world knowledge* and that a student with good general knowledge might obtain a high score in the exam as long as he or she possesses well-developed *higher-order thinking skills*.

As for the validity of the scoring process, the general agreement between the teachers' orientations towards subject matter and the aspects of content identified in the subject curriculum attests to fairly good correspondence between the teachers' cognitive processes and the intended construct to be measured (Bejar, 2012). One potential threat to validity, however, is the reluctance towards metacognitive strategies as an assessment criterion. Another is the variation in teacher perceptions regarding what kind of performance indicates proficiency at the different levels.

IX. Conclusion and Implications

This study has investigated Norwegian EFL teachers' conceptions of subject matter content in an oral exam at the upper secondary level and compared these conceptions with aspects of content specified in the English subject curriculum. The results show that the teachers have very general conceptions of content, something which corresponds well with the content construct as defined in the curriculum. Moreover, the findings indicate that the teachers are generally more concerned with the skills and process aspects of content than with specific subject matter. In particular, they seem oriented towards higher-order thinking skills, such as the ability to reflect on a given topic.

Three limitations of this study must be kept in mind. First of all, the teacher sample was small, something which makes generalizations to other contexts problematic. Secondly, although the teachers were interviewed on their orientations towards the content construct generally, they were probably influenced by the performance of the student in the video-clip. Therefore, had there been another student giving a different performance, the teacher responses may also have been somewhat different. Thirdly, introspective methods such as interviews and VPA, which were used in this study, do not automatically predict genuine teacher behaviour in authentic assessment situations.

The study has two major implications. The first one relates to the importance of higher-order thinking skills in upper-intermediate level L2 language education. As raters appear to be concerned with such skills in this context, teachers should take care to provide classroom tasks and material which let students develop their ability to analyse, reflect on and evaluate subject matter content. Secondly, the validity problems related to the role of metacognitive strategies in the test construct, as well as the differences in perceptions concerning what kind of behaviour that is indicative of performance at the different levels, need to be addressed. One feasible solution to these problems is the introduction of a common rating scale, which may better guide the teachers in their operationalization of the content construct (Fulcher, 2012). In addition, it seems that more rater training would be beneficial, as this is reported to have positive effects on reliability (Taylor & Galaczi, 2011).

An avenue for further research is the question of the interface between language and content. As current research is particularly concerned with how the language and content constructs interrelate (Snow & Katz, 2014), it would be relevant to explore further how the teachers understand this interrelation in EFL/ESL teaching.

References

[1] Anderson, L. W., & Kratwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. New York: Longman.
[2] Bachman, L. F., & Palmer, A. (2010). Language Assessment in Practice. Oxford: Oxford University Press.
[3] Bachman, L. F., & Palmer, A. S. (1996). Language testing in practice. Oxford: Oxford University Press.
[4] Bejar, I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, 31(3), 2-9. doi:10.1111/j.1745-3992.2012.00238.x.
[5] Brinkmann, S., & Kvale, S. (2015). InterViews: Learning the craft of qualitative research interviewing. Thousand Oaks: Sage.
[6] Brown, A. (2000). An Investigation of the Rating Process in the IELTS Oral Interview. Retrieved from http://www.ielts.org/pdf/Vol3Report3.pdf (accessed 27 April 2016).
[7] 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.
[8] Byrnes, H. (2008). Assessing content and language. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education* (Vol. 7, pp. 37-52). New York: Springer Science+Business.
[9] Bøhn, H. (2015). Assessing spoken EFL without a common rating scale: Norwegian EFL teachers' conceptions of constructs. Sage Open. October-December 2015. Retrieved from http://sgo.sagepub.com/content/spsgo/5/4/2158244015621956.full.pdf.
[10] Chamot, A. U. (2009). The CALLA handbook: Implementing the Cognitive Academic Language Learning Approach (2nd ed.). White Plains, NY: Pearson Education.
[11] Douglas, D. (2000). Assessing languages for specific purposes. Cambridge: Cambridge University Press.
[12] Douglas, D. (2010). Understanding Language Testing. Oxon: Hodder Education.
[13] Echevarría, J., Vogt, M. E., & Short, D. J. (2008). Making content comprehensible for English learners: The SIOP model (3rd ed.). Boston: Allyn & Bacon.

[14] 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). Frankfurt: Peter Lang.
[15] Frost, K., Elder, C., & Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing*, 29(3), 345-369. doi:10.1177/0265532211424479.
[16] Fulcher, G. (2012). Scoring performance tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing* (pp. 378-392). Oxford: Routledge.
[17] Fulcher, G., & Davidson, F. (2007). Language Testing and Assessment. Oxford: Routledge.
[18] Green, A. (1998). Verbal protocol analysis in language testing research: A handbook (Vol. 5). Cambridge: Cambridge University Press.
[19] Green, A. (2014). Exploring language assessment and testing: Language in action. Oxon: Routledge.
[20] Hasselgård, H., Lysvåg, P., & Johansson, S. (2012). English grammar: Theory and use (2nd ed.). Oslo: Universitetsforlaget.
[21] Kratwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, 41(4), 212-218. doi:10.1207/s15430421tip4104_2.
[22] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174. doi:10.2307/2529310.
[23] Luoma, S. (2004). Assessing Speaking. Cambridge: Cambridge University Press.
[24] McNamara, T. (1996). Measuring Second Language Performance. Harlow: Longman.
[25] Met, M. (1998). Curriculum decision-making in content-based language teaching. In J. Cenoz & F.Genesee (Eds.), *Beyond bilingualism: Multilingualism and multilinugal education* (pp. 35-63). Philidelphia, PA Multilinugal Matters.
[26] Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). Qualitative data analysis: A methods sourcebook. Los Angeles: Sage.
[27] Norwegian Directorate for Education and Training [UDIR]. (2014). Rundskriv Udir-02-2014 - Lokalt gitt muntlig eksamen [Circular Udir-02-2014 - Locally administered oral exams]. Oslo: Author. Retrieved from http://www.udir.no/Regelverk/Finn-regelverk-for-opplaring/Finn-regelverk-etter-tema/eksamen/Udir-2-2014-Lokalt-gitt-muntlig-eksamen/ (accessed 11 April 2016).
[28] 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. Cambridge: Cambridge University Press.
[29] Saldaña, J. (2013). The Coding Manual for Qualitative Researchers (2nd ed.). London: Sage.
[30] 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.
[31] 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: Cambridge University Press.
[32] Yildiz, L. M. (2011). English VG1 level oral examinations: how are they designed, conducted and assessed? Oslo: L.M. Yildiz.
[33] Yin, R. K. (2016). Qualitative research from start to finish (2nd ed.). New York: The Guilford Press.

Henrik Bøhn received a Master's degree in English (1997) and a PhD in English education (2016) from the University of Oslo, Norway. He has worked as an English teacher in upper secondary school in Norway for three years and as a lecturer and researcher at the tertiary level in Norway for 17 years. His research interests include language assessment, language acquisition and intercultural communication.