DOI: 10.18261/978-82-15-03074-6-2019-19

18

PhD revisited: What is to be assessed?

Teachers' understanding of constructs in an oral English examination in Norway[1]

HENRIK BØHN
University of South-Eastern Norway

ABSTRACT In this chapter, the research design of and main findings from Bøhn's (2016) doctoral study are presented. The study used educational and psychological measurement theory and a mainly qualitative methodological approach to investigate teachers' understanding of what should be assessed in an oral exam in English in Norway. The findings indicated that the teachers generally agreed on the main aspects of student performance to be assessed, but disagree more on the more narrowly defined aspects. On the basis of the results the chapter discusses implications for oral assessment in English language teaching and possible avenues for further research.

KEYWORDS language assessment | spoken English | English language teaching | validity

1. The chapter presents the overall results of a doctoral study (Bøhn, 2016), focussing on summative oral assessment practices in Norway. This is an article-based thesis, with three published articles (Bøhn, 2015; 2017; 2018), and the thesis in its entirety – with theoretical, methodological and empirical details – can be found here: https://www.duo.uio.no/handle/10852/53229.

INTRODUCTION

Assessment[2] holds a prominent place in education. It can be defined as "the planned and systematic process of gathering and interpreting evidence about learning in order to make a judgement about that learning" (Isaacs, Zara, Herbert, Coombs & Smith, 2013, p. 1). In the last 20 years or so, assessment has received increasing attention from educational authorities, researchers, teaching practitioners and the general public in many countries. This includes English language teaching (ELT) in Norway, where there has been focus on assessment criteria, teachers' scoring consistency, and the relationship between the subject curriculum and final assessment, among other things (Directorate for Education and Training, 2010a; Yildiz, 2011).

Assessment is a complex matter, however, and the practice is wrought with a number of philosophical, political, social, ethical and technical issues. Very generally, these issues revolve around broad questions such as *why* a particular assessment is used, *what* is being assessed, and *how* the assessment is designed and used. In the doctoral dissertation reported in this chapter, which investigated the assessment of student performance in an oral English exam, the emphasis was mainly on "what". This is an important question in the sense that good quality assessment depends on a clear understanding of which aspects of a performance or a learning process that should be evaluated. Research has shown, however, that raters often find it difficult to agree on exactly what should be in focus (Eckes, 2009).

In international language testing, this problem has long since been recognized, and a lot of resources have thus been spent on developing high-quality language assessment (e.g. the Cambridge Assessment English tests or the TOEFL test). In such assessment contexts, raters are always provided with fairly detailed *rating scales*, or "scoring rubrics" (Ginther, 2013). Such scales specify both which performance aspects (i.e. criteria) should be attended to, and descriptions of how these aspects should be rated at each level of performance. Furthermore, in such testing situations raters are *trained* in order to help them know what to look for and how to score performance.

In the Norwegian educational context, however, the situation has traditionally been somewhat different in terms of English oral exams. In these exams there are no national rating scales or rater training. Rather, the authorities have left it to the local level (i.e. the county governors) to administer oral exams. This has led to dif-

2. The terms "assessment", "testing", and "evaluation" are frequently used with different meanings. In this chapter, however, I follow Bachman and Palmer (2010), who use them more or less interchangeably.

ferent practices across counties. In some counties teachers have had access to a common rating scale, and have had some rater training. In other counties there have been no common scales and no training. Such differences are potentially problematic for dependable and valid assessment results (see e.g. Nusche, Earl, Maxwell, & Shewbridge, 2012).

Against this background, the doctoral study reported in this chapter explored the following broad research question: *How do teachers in Norway understand the "what" to be tested in an oral English exam at the upper secondary level?* The focus was on the oral English exam taken by students in their first year of the general studies programme (GSP), and in the second year at the vocational studies programmes (VSP).

THEORY

In test theory, the "what" to be evaluated in language assessment is commonly referred to as "attributes", "constructs" or "traits". According to Weir (2005, p. 1), constructs can be defined as the "underlying […] abilities we wish to measure in students'. This means, for example, that when a student gives a performance in an oral exam, it is not the actual words being spoken that we are primarily interested in. Rather, we are interested in the student's underlying competences or skills, such as vocabulary, fluency, accuracy or pronunciation. These, then, are unobservable entities that cannot be assessed directly. In order to be assessed, they need to be operationalized, i.e. made concrete, before they can be evaluated. Fluency, for instance, can be operationalized by the observable features "pauses", "fillers", "false starts" etc. (Brown et al., 2005, p. 23).

According to Bachman & Palmer (2010), constructs are identified on the basis of a "frame of reference", such as a theory of language, a needs analysis or a syllabus (pp. 212–213). In English language teaching in Norway, it is first of all the latter category which defines the constructs, as teaching and assessment are supposed to be grounded in the English subject curriculum.

As for the operationalization of the constructs to be assessed, this is commonly done with the help of rating scales, as mentioned above. The use of rating scales in language testing is generally believed to enhance the validity of the assessment (Fulcher, 2012). There are different definitions of "validity", but today most language assessment specialists agree that it concerns the quality of the inferences that can be made from the assessment results (see e.g. Newton & Shaw, 2014). In this sense, validity has to do with *score meaning*. For example, if a student in Norway is awarded a "4" on the oral exam, one may ask what this mark means.

According to the Regulations to the Education Act, a "4" means that the student has "good competence in the English subject" (Norwegian Ministry of Education and Research, 2009, § 3–4, my translation). But then one could go on to ask: "What kind of competence has been assessed?" Here, the Regulations of the Education Act specify that it is the competence aims of the subject curriculum that decide which competence(s) that are to be focused on. This, however, leads us back to the question of constructs and how these constructs have been operationalized.

Good validity, then, requires that raters only attend to those aspects that are meant to be assessed. Whenever raters fail to take into account performance features that should be tested, this will be a "threat" to the validity of the scores. In such cases the results will be affected by *construct underrepresentation.* Conversely, if raters start attending to performance features that should not be tested, this will create *construct-irrelevant variance* in the assessment results (Messick, 1989). Construct underrepresentation and construct-irrelevant variance are therefore something that should be avoided.

A related question is the issue of reliability, or dependability. Simply put, reliability can be said to indicate the extent to which the same raters would award the same score, or mark, to the same performance. For instance, if one rater gives a performance a "4", and another gives the same performance a "2", this would be an example of poor reliability. Differently put, such assessment discrepancy means that the mark "4" does not mean the same thing. In this respect, reliability affects validity since it impinges on the quality of the inferences that can be made from the marks, or assessment scores.

REVIEW

As the main focus of the present investigation is on rater cognition, or raters' *orientations* in foreign or second language (L2) speaking assessment, it is relevant to consider studies that have looked into this phenomenon. In the assessment research literature there is ample evidence that scoring outcomes (i.e. marks) are often affected by raters' subjective understanding of how performance is to be judged (Bejar, 2012). This phenomenon is commonly referred to as *rater variability* (McNamara, 1996).

Rater orientation research in L2 language assessment has shown that raters pay attention to both construct-relevant and construct-irrelevant features when judging performance (Hsieh, 2011; Orr, 2002; Pollitt & Murray, 1996). For example, raters have been shown to heed construct-irrelevant performance aspects such as

age and gender (Orr, 2002), effort (Brown, 1995), interest and personality (Ang-Aw & Goh, 2011), physical attractiveness (Pollitt & Murray, 1996) and voice quality (Hsieh, 2011). There is also evidence of construct underrepresentation in a number of tests, as examiners fail to pay attention to criteria that should be considered, such as content-related performance aspects (Cai, 2015).

Although there is a rich body of research on language assessment generally, very few studies have examined English speaking tests in contexts where rating scales have not been provided. Only two international studies (Brown, Iwashita & McNamara, 2005; Pollitt & Murray, 1996) and one Norwegian study (Yildiz, 2011) have been identified. Brown et al. (2005) and Pollitt & Murray (1996) found that raters were attending to performance features such as linguistic resources (vocabulary, grammar, phonology), fluency, and content, whereas Yildiz (2011) discovered that the raters were focusing on "Language competence", "Communicative competence", "Subject competence", "Ability to reflect and discuss independently" and "Ability to speak freely and independent of manuscript". Yildiz's study is particularly interesting in this discussion since she investigated an assessment context which is almost identical to the present one, i.e. an oral English exam at the upper secondary level. However, the study was quite small, being an MA study including only 16 teacher informants.

Beyond these studies, there is research indicating that raters may have more or less common perceptions of how performance should be assessed. Some studies, for example, have demonstrated relatively good correspondence between raters' orientations (Brown et al., 2005), whereas others have recorded substantial rater variability (Orr, 2002). An important question in this respect is how such differences can be explained. Some studies have suggested that the differences may be a matter of rater background characteristics, such as rating experience or first language background (Kim, 2009). Others have indicated that there has not been sufficient rater training (Brown, 2012). Relatedly, there may be problems with how the rating scales are to be interpreted (Eckes, 2009).

METHODOLOGY

In order to investigate the teachers' understanding of constructs, this study used a predominantly qualitative, exploratory research design with an *inductive theoretical drive* (Morse & Niehaus, 2009). This means that the overall direction of the investigation was guided by the inductive analysis of data, which were collected through the use of qualitative methods in the first stage of the project. Two of these constructs, pronunciation and content, were analysed is some more detail, and a

quantitative instrument, i.e. a questionnaire, was used as a part of the study to investigate specific questions related to the pronunciation construct.

DATA COLLECTION

The investigation was carried out in three phases, each representing a separate study. In Phase 1, a student in the Health and Social Care vocational study programme was filmed as she was taking her oral exam. The video-clip was then distributed to a group of teachers who were asked to score the performance and justify their decisions. Semi-structured interviews were used to elicit the teachers' understanding of constructs when rating oral performance. On the basis of the results from the analysis in Phase 1, it was decided to look further into the pronunciation[3] and content constructs in Phase 2 and Phase 3 as there turned out to be noticeable rater variability regarding these two constructs. In Phase 2 another group of teachers were therefore recruited to watch the same video-clip and to answer a questionnaire regarding the assessment of pronunciation and intonation. In addition, semi-structured interviews were used to complement the data collection. Finally, in Phase 3 verbal protocol analysis (VPA) was used to examine a third group of teachers' understanding of the content construct. VPA is a method in which participants are asked to verbalize their thoughts, or "speak their mind", when carrying out a task (Green, 1998). Again, the video-sequence recorded in Phase 1 was used. In addition to the VPA, semi-structured interviews were used to collect data about the teachers' understanding of content. Table 18.1 gives an overview of the studies in each of the three phases.

3. Here the term *pronunciation* covers both segmental (i.e. the pronunciation of individual sounds) and suprasegmental features (such as intonation, stress and rhythm).

TABLE 18.1. An overview of the three phases of the investigation.

	Phase 1	**Phase 2**	**Phase 3**
Research focus	Teachers' understanding of *constructs generally*	Teachers' orientations towards aspects of the *pronunciation* and *intonation* constructs	Teachers' understanding of the *content* construct
Data collection	Semi-structured interviews	◗ Semi-structured interviews ◗ Questionnaire	◗ Verbal protocols ◗ Semi-structured interviews
Number of participants	24 interviewees (also interviewed in Phase 2)	◗ 24 interviewees (also interviewed in Phase 1) ◗ 46 questionnaire respondents	10 verbal protocol and interview informants
Data analysis	Qualitative and quantitative content analysis (Galaczi, 2014)	Qualitative, using magnitude and provisional coding (Miles et al., 2014), and quantitative, calculating descriptive statistics	Qualitative, using provisional coding (Miles et al., 2014)

PARTICIPANTS

The teacher participants in all the three studies (n=80) were fully qualified upper secondary school teachers of English. They were recruited by means of purposeful sampling (Creswell, 2013), in order to obtain variation in the samples with regard to age, gender, first language, teaching experience, county and study programme affiliation. There were 25 male and 55 female teachers. Some were involved in the vocational studies programmes only (n=19), some were involved in the general studies programme only (n=14), and the majority were involved in both programmes. Most participants had Norwegian as their L1 (n=64). The rest were L1 speakers of English (n=6), Swedish (n=4), Danish (n=1), Finnish (n=1), Mandarin (n=1), Romanian (n=1) and Russian (n=1). Their rater experience in the oral English exam ranged from none to more than six exams.

DATA ANALYSES

The interviews and the verbal protocols were investigated using the computer programme QSR NVivo10. Both interview and verbal protocol data were transcribed,

checked and returned to the participants for respondent validation (Bryman, 2012). The questionnaire data were analysed using IBM SPSS Statistics.

The data gathered in Phase 1 were analysed using qualitative and quantitative content analysis (Krippendorf, 2013). Analytical categories were developed from the teachers' statements, by means of coding, without any explicit theoretical framework to support the analysis. In this sense, the analysis was predominantly inductive. For example, a statement like "And if they can't [find the word], they should try to circumvent it, rather than switching into Norwegian" was coded as "Compensatory strategies". The qualitative analysis involved an in-depth scrutiny of teacher statements, and a comparison between different categories and statements, in order to capture the participants' understanding of the constructs and their interrelationships. The quantitative analysis meant that the different categories were subsequently counted in order to give an indication of how prominent the different constructs were.

In Phase 2 both the questionnaire and the interview data were analysed deductively in the light of the theoretical framework centred on the concepts of *nativeness* and *intelligibility*. The starting point for the analysis was the question of whether students need to have a (near-) native accent in order to obtain a top score. In the research literature the idea of learners trying to approximate native speakers is sometimes referred to as the "nativeness principle" (Levis, 2005). Historically, this principle has had a strong position in English language teaching and assessment. However, a competing principle, the "intelligibility principle", has gained ground in later years. This principle holds that the aim of pronunciation teaching and assessment should not be for learners to speak like natives, but rather to make themselves understood, i.e. to be intelligible.

In order to explore the teachers' orientations towards nativeness and intelligibility, these concepts were operationalized in different ways in the interviews and in the questionnaire. In the interviews the following question was asked: "Some teachers say that a near-native speaker accent is important in order to get a top score. What is your comment on that?" In the questionnaire, nativeness was operationalized in three different items and intelligibility was measured with the use of two items. An example of the former was: "Good 'native speaker' accent is important in order to achieve a top score", and an example of the latter was: "If it is difficult to understand what the student says, I will automatically mark him/her down from a 6 [i.e. the top score]". In addition, a set of items relating to four specific pronunciation features which have been found to relate to intelligibility were also included in the questionnaire. These features were *segmentals*, i.e. the pronunciation of individual sounds, *word stress*, *sentence stress*, and *intonation*. An

example of an item measuring segmentals was: "The correct pronunciation of individual sounds is important in order to achieve a top score (for example /hedeɪk/ and not /hedeɪtʃ/ for 'headache'; /wɒtʃɪŋ/ and not /wɒʃɪŋ/ for 'watching')". The responses to all items were given on a five-point Likert scale going from "Completely disagree" to "Completely agree".

The questionnaire responses were investigated by using descriptive statistics such as means and standard deviations in order to find out how the teachers rated the importance the different pronunciation and intonation constructs. The interview data regarding nativeness were explored through the use of *magnitude coding* (Miles, Huberman & Saldãna, 2014). It involved the assignment of coded teacher statements along a four-point scale going from "not at all" in agreement with the nativeness principle to "to a large extent" in agreement. The issue of intelligibility in the interview responses were analysed by using *provisional coding* (Miles, Huberman & Saldãna, 2014). This entailed the initial establishment of categories denoting intelligibility, such as "comprehensible speech" and "understanding", followed by a search for teacher statements that matched these categories.

In Phase 3, conceptualizations from Bloom's revised taxonomy of educational objectives (Anderson et al, 2001) were mainly used to analyse the data. In this taxonomy content is understood as a two-dimensional construct. It involves a cognitive process dimension and a knowledge dimension. The former refer to students' cognitive *skills or abilities* and include the following cognitive aspects: "remember", "understand", "apply", "analyse", "evaluate", and "create". The latter concerns *subject matter* and involves "factual", "conceptual", "procedural", and "metacognitive" knowledge (Kratwohl, 2002). The taxonomy was also found suitable as the competence aims of the English subject curriculum in Norway to a large extent reflect this type of thinking. The competence aims typically consist of verbs denoting the cognitive process dimension and noun phrases representing the knowledge dimension. The following competence aim at the upper secondary level (year 1 or 2) illustrates this: "[The student should be able to] discuss and elaborate on the growth of English as a universal language". In this competence aim the cognitive process dimension is realized by the verb phrase "discuss", and the knowledge dimension is denoted by the noun phrase "the growth of English as a universal language".

Bloom's revised taxonomy was also found relevant in the sense that the different aspects of the cognitive process dimension to some extent reflect a scale of increasing complexity. In this view, "understand" is a more complex cognitive process than "remember", "apply" is more complex than "understand", and so on.

Thus seen, the cognitive process dimension can represent an assessment scale going from lower to higher levels.

The data obtained from the VPAs and the interviews in Phase 3 were then analysed in two cycles with the use of provisional coding (Miles et al., 2014). In the first cycle, the statements were coded on the basis of the construct categories developed in Phase 1, such as "Fluency", "Grammar", and "Vocabulary", as well as the analytical framework largely built on Bloom's revised taxonomy. This framework was used for classifying statements relating to content, where verb phrases represented the cognitive process dimension and noun phrases denoted knowledge, or subject matter. An example is the following statement: *She didn't get the chance to, sort of, talk about the English language as a world language and an international language*. Here the verb phrase "talk about" was classified as a cognitive process category (cognitive ability), and "English as a world language and an international language" was coded as a subject matter aspect. In the second cycle, all the phrases relating to subject matter were sifted out in order to examine what kind of knowledge the teachers were concerned with, as this was a major focus of the third study.

RESULTS

The analyses of interview transcripts in Phase 1 concerning constructs generally showed that the informants paid attention to a large number of different aspects when assessing student performance. 38 categories were identified in the analyses presented in the doctoral thesis, including both main constructs, sub-constructs and sub-subconstructs. Overall, the results showed that the teachers focused on two main constructs, namely *Communication* and *Content*. *Linguistic competence* (belonging to Communication) and *Application, analysis, reflection* (belonging to Content) turned out to be the two most significant sub-constructs.

Linguistic competence involved aspects such as grammar, vocabulary and pronunciation, whereas Application, analysis, reflection referred to the ability to apply knowledge, as well as to be able to analyse and reflect on various issues. Table 18.2 shows the results from the quantitative content analysis, where the counts indicate the number of times each category was mentioned. (It is worth mentioning that many of the sub-constructs listed in the table, such as Linguistic competence, comprise different sub-subcategories, like for example Vocabulary, Grammar and Phonology, which are not included in this table.)

TABLE 18.2. Number of reference counts for the different statements pertaining to constructs and sub-constructs.

Constructs	Criteria	Reference counts
Communication	(General reference to communication)	28
	Linguistic competence	240
	Compensatory strategies	24
	Listening comprehension	21
	Take initiative	15
	Communicative structure	2
	Adapt communication to situation and audience	6
	Cohesion	2
	Ability to repair	2
	Social competence	2
	Sum Communication	**342**
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 can be seen in Table 18.2, the teachers mentioned categories related to Communication nearly twice as often as they did categories related to Content (342 as against 189). Within these two constructs there were 240 teacher statements relating to the sub-construct Linguistic competence, and 44 statements relating to the category Application, analysis, reflection. It should be pointed out, however, that the number of counts does not directly express the strength of correlation between statements and the significance of a category. Still, it gives an indication of how important the teachers found the different constructs to be.

Both the quantitative and qualitative analyses indicated that there was good correspondence between the individual teachers' understanding of the main constructs, but that there was some more discrepancy regarding the sub-subconstructs, particularly pronunciation. In addition, there were indications that the teachers weighted the content construct differently. The teachers mainly involved in the general studies programme tended to put more emphasis on content, whereas the teachers mainly working at the vocational studies programmes were less concerned with this construct. Beyond this, there was evidence that some teachers heeded construct-irrelevant performance features. For example, there were teachers who listed effort and level of preparedness as relevant assessment criteria. Additionally, it was found that some teachers agreed on which criteria were to be applied, but disagreed on how performance was to be assessed regarding these criteria. For example, two teachers agreed that fluency was a relevant criterion. However, one teacher thought that the performance of the student in the video-clip was "fluent", whereas another found it to be "fairly choppy".

The results from the analyses in Phase 2, which looked more closely at the teachers' assessment of pronunciation and intonation, indicated that there was strong agreement on the question of intelligibility, i.e. the students' ability to make themselves understood. For instance, 37 of the 46 questionnaire respondents strongly or completely agreed that students should be graded down from a top score if it was difficult to understand what they were saying. The analysis of the interview transcripts supported this finding, as 11 out of 24 informants stressed the importance of "clear pronunciation" and "comprehensible speech" in their discussion of general criteria.

As for nativeness, or the importance of speaking with a near-native accent, there was much more variation among the teacher participants. Of the 46 questionnaire respondents six strongly disagreed that it did matter, seven strongly agreed that it did, eight moderately disagreed, and 11 moderately agreed. The largest group of respondents, 13 teachers, neither agreed nor disagreed. Figure 18.1 visualizes the responses to this item.

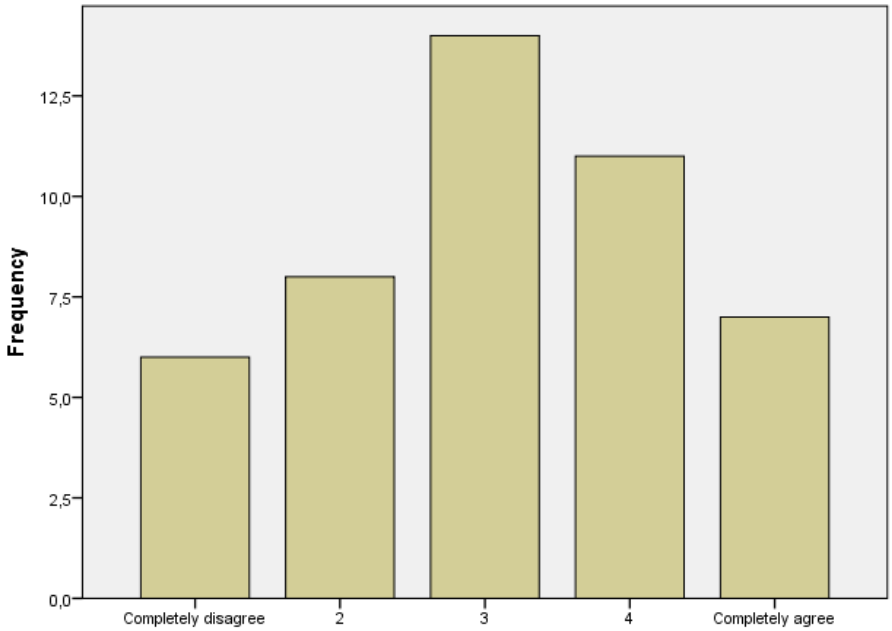

FIGURE 18.1. The distribution of responses to the item: "A strong Norwegian accent will mark the student down from a top score" (n = 46).

The interview analysis corroborated the findings from the questionnaires. Five respondents did not at all see nativeness as important, six found it to be of considerable importance, five thought it was of little importance, and five believed it was of some importance. A statement from a teacher who strongly opposed to using nativeness as a criterion said:

> I don't kind of expect Norwegian students to be native speakers. I don't think it is ethical for the teachers to give a lower grade just because the student hasn't got a proper British accent, or an American accent. It is my personal opinion.

However, there were also statements from some of the teachers that displayed an ambivalent attitude towards nativeness. The following exchange between the researcher and an informants serves as an example:

> Informant: Anyway, I think it is quite o.k. that [the students] don't speak perfect British English or American English, since English has become sort of a global language. This means that we must accept that pronunciation has been localized in different parts of the world.

> Researcher: But [...] would you say that it would be important to have a native speaker accent in order to obtain a 6?
>
> Informant: If you do, that's the best thing, but if you don't ... I don't think it's a must.

Here the informant starts by downplaying the relevance of the nativeness principle. However, when asked specifically about this issue, she admits that native speaker approximation is preferable, although it is not a "must".

Finally, in terms of the four phonology features that were included in the questionnaire, the results revealed that the respondents moderately to strongly agreed that segmentals, word stress and sentence stress were important. As regards intonation, however, the teachers either found this performance aspect less important, or they were uncertain about whether to assess it. On the five-point Likert scale measuring responses to this construct in the questionnaire, the results yielded the average score of M = 3.07.

In Phase 3 the VPA and interview analyses supported the findings from Phase 1 that teachers largely view content as a matter of responding well to the task questions, as well as to analyse and reflect on subject matter. Thus, they also confirmed the assumption that Bloom's revised taxonomy (Anderson et. al., 2001) is a relevant way of understanding content when assessing performance. Regarding what type of subject matter they viewed as important, the analyses showed that the teachers had a very general understanding of this issue. This may reflect the fact that the English subject curriculum does not specify a lot of factual knowledge to be learnt. Rather, it points to a number of wide-ranging subject matter aspects, such as "discuss and elaborate on culture and social conditions in several English-speaking countries". As one informant put it:

> 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.

The analysis furthermore indicated that a consequence of this type of thinking is that the cognitive process elements, such as to apply, analyse, and evaluate, become more important than specific subject matter knowledge, since it is unrealistic to expect students to remember details from all kinds of potential topics. A quote from another informant supports this claim:

> 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.

Another interesting point concerning this quote is the formulation "at a certain level… and not just surface level knowledge". This statement supports the hypothesis that the teachers tend to think in terms of a taxonomy, since the goal is for students to reach the higher levels in the taxonomy. In other words, a high score in the exam requires higher-order thinking skills, not just "surface level knowledge".

Two final points are worth making concerning Phase 3. First of all, the study to some extent supported the finding from Phase 1 that teachers at the general studies programmes place more emphasis on content than do teachers at the vocational studies programmes. Secondly, the teachers' understanding of content was fairly consistent with the content constructs identified in the subject curriculum. However, one instance of construct-underrepresentation was found. The informants were hesitant about the assessment of learning strategies, which is clearly defined as a competence aim in the curriculum. As one informant put it: "No, that is not to be tested… It is a meta-science".

DISCUSSION: CONTRIBUTION TO THE DIDACTICS FIELD

Most importantly, the findings in this doctoral dissertation should be discussed against the backdrop of the study, i.e. the Norwegian assessment context, where no common rating scales or common rater training existed at the time of the data collection. In the language testing literature the lack of such assessment aids is generally thought to be problematic for the validity and reliability of the scores (Brown, 2012; Fulcher, 2012). However, as I will return to below, there may also be arguments against the use of too "standardized" assessment procedures, especially in educational settings where assessment is closely linked to learning.

EMPIRICAL CONTRIBUTION: ASPECTS OF SCORE VALIDITY

Overall, the investigation found that the teachers generally understood the main constructs in the same way. However, there were also examples of construct underrepresentation and construct-irrelevant variance, which threatened the validity of the scores. Four areas turned out to be particularly problematic. Firstly, there was evidence that the teachers in the vocational studies programmes downplayed the content construct, and in doing so, underrepresented it. This may be due to the fact that students in these programmes are, on average, at a lower proficiency level than students in general studies programmes. Therefore, many teachers seem to prioritize language over content aspects, both when teaching and assessing performance, as one teacher in the interview made clear. This suggests that teachers in the vocational programmes see the English subject as more "language driven", whereas teachers in the general studies programmes see it as somewhat more "content driven" (Met, 1998). Secondly, there were indications that some teachers paid attention to construct-irrelevant features, such as effort and level of preparedness. According to a government circular, these aspects are not to be considered in final assessment, such as in oral exams (Norwegian Directorate for Education and Training, 2010b). Thirdly, there was variability regarding the assessment of pronunciation, especially related to the question of whether students need to approximate native speakers in order to achieve a top score. This finding supports Levis' (2005) claim that teachers either tend to focus on nativeness or intelligibility in pronunciation pedagogy (cf. above). Fourthly, there was the problem of teachers agreeing on the constructs to be assessed, but disagreeing on how performance should be scored with regard to these constructs. Whenever teachers say that, for example, fluency is important, but then fail to agree on what kind of performance is indicative of good fluency, this threatens the validity of the scores. Finally, the fact that teachers in the general studies programme put more emphasis on content than do the teachers in the vocational studies programmes is problematic for validity.

THEORETICAL CONTRIBUTION

The main theoretical contribution of this doctoral dissertation relates to its conceptualization of the content construct in language assessment. Largely based on Bloom's revised taxonomy of educational objectives (Anderson et al., 2001), this conceptualization first of all sees content as a construct comprising a subject matter (or *what*) dimension and a skills (or *how*) dimension. Due to the nature of the assessment context, where the subject curriculum for the most part describes sub-

ject matter in very general terms, it was found that teachers end up emphasizing the cognitive skills dimension since it seems to matter less to the teachers which topic the student has knowledge of, as long as he or she is able to reflect on that knowledge. This relates to the theoretical notion of *higher-order thinking skills*, which becomes important for students who aim for the highest marks in the exam.

IMPLICATIONS FOR TEACHING AND ASSESSING ENGLISH

The findings from this doctoral study are relevant in the discussion of both how to teach and how to assess student performance in ELT. Three aspects are particularly salient and were brought to the fore by the author at the time of the publication of the thesis. Firstly, it was recommended that national rating scale guidelines and better, more systematic rater training are introduced. As mentioned above, there are indications that this will allow for improved validity and reliability in assessment generally. More specifically, it was suggested that rating scale *guidelines*, rather than fixed rating scales, are provided. The reason for this is that fixed scales may make the teachers focus too strongly on the criteria for assessment rather than on the competence aims of the subject curriculum when teaching English (see e.g. Throndsen et al., 2009). As for national rater training, the Directorate for Education and Training could, for example, provide benchmark samples in the form of videotaped student performances in mock exam situations. Such samples could then help teacher raters get a common perception of what characterizes performance at the different grade levels.

Secondly, it was recommended that the pronunciation construct needed to be better defined, both for teaching and assessment purposes. The current subject curriculum is not clear on this score, referring for example to "patterns of pronunciation" (Norwegian Ministry of Education and Research, 2006/2013). Hence, it needs to be decided whether or not the nativeness principle should have relevance in assessment. Research suggests that nativeness is of less importance for communication (Derwing & Munro, 2009), but more research on which features are important for intelligibility is needed before clear recommendations in this area can be given.

Thirdly, since many raters in the study were concerned with the assessment of higher-order thinking skills, such as analysis and reflection, teachers were advised to allow for the practice of such skills in the classroom. For example, it would seem efficacious to provide students with tasks that allow for scrutiny, synthesis, contemplation and evaluation, and working on understanding concepts and the relation between them (see e.g. Chamot, 2009).

SUGGESTIONS FOR FURTHER RESEARCH

The starting point for this investigation was the assumption that the lack of rater training and rating scales in the Norwegian educational system can pose validity problems in oral English exams. However, validity and the process of *validation*, i.e. efforts to make assessment processes and outcomes more valid, continue to be challenging in all assessment contexts, since there are so many different aspects that impact the scoring outcomes (O'Sullivan, 2014).

One such challenge concerns the assessment of pronunciation and how, for example, one can understand the relationship between nativeness and intelligibility, and related concepts such as "correctness" and "error". As has been pointed out in this chapter, teachers struggle to operationalize the pronunciation construct, which is quite understandable given that the curriculum is so vague on this score. One obvious topic for further research would therefore be the assessment of intonation. Another could be the attitudes towards nativeness and intelligibility at the lower secondary school level.

Another underresearched area is the relationship between language and content, and how raters distinguish between these two constructs (cf. Snow & Katz, 2014). In addition, in the Norwegian context, it would be relevant to look further into how Norwegian teachers assess English when giving overall achievement marks. These are also issues related to the validity of assessment outcomes and the quality of assessment practices in schools.

REFERENCES

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.

Ang-Aw, H. T., & Goh, C. C. M. (2011). Understanding discrepancies in rater judgement on national-level oral examination tasks. *RELC Journal, 42*(1), 31–51. doi: 10.1177/0033688210390226

Bachman, L. F., & Palmer, A. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Baird, J.-A., Hopfenbeck, T. N., Newton, P., Stobart, G., & Steen-Utheim, A. T. (2014). State of the field review: Assessment and learning. (Case number 13/4697). Oslo: Knowledge Centre for Education.

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

Brown, A. (2012). Interlocutor and rater training. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing*. Oxford: Routledge.

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

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.

Bryman, A. (2012). *Social research methods* (4th ed.). Oxford: Oxford University Press.

Bøhn, H. (2016). What is to be assessed? Teachers' understanding of constructs in an oral English examination in Norway (Unpublished doctoral thesis). University of Oslo, Oslo, Norway. Retrieved from https://www.duo.uio.no/handle/10852/53229

Cai, H. (2015). Weight-based classification of raters and rater cognition in an EFL speaking test. *Language Assessment Quarterly, 12*(3), 262–282. doi: 10.1080/15434303.2015.1053134

Chamot, A. U. (2009). *The CALLA handbook: Implementing the Cognitive Academic Language Learning Approach* (2nd ed.). White Plains, NY: Pearson Education.

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

Derwing, T., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language teaching, 42*(4), 476–490.

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.

Fulcher, G. (2012). Scoring performance tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing* (pp. 378–392). Oxford: Routledge.

Ginther, A. (2013). Assessment of speaking. In C. E. Chapelle (Ed.), *The Encyclopedia of applied linguistics* (pp. 234–240): Wiley Blackwell.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge: Cambridge University Press.

Harlen, W. (2012). On the relationship between assessment for formative and summative purposes. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 87–101). London: Sage.

Hsieh, C.-N. (2011). Rater effects in ITA testing: ESL teachers' vs American undergraduates' judgements of accentedness, comprehensibility, and oral proficiency. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 9*, 47–74. Retrieved from Spaan Fellowship website: http://www.cambridgemichigan.org/sites/default/files/resources/SpaanPapers/Spaan_V9_Hsieh.pdf

Isaacs, T., Zara, C., Herbert, G., Coombs, S. J., & Smith, C. (2013). *Key concepts in educational assessment*. London: SAGE Publications Ltd.

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187–217. doi: 10.1177/0265532208101010

Kratwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212–218. doi: 10.1207/s15430421tip4104_2

Krippendorf, K. (2013). *Content Analysis* (3rd ed.). Thousand Oaks: Sage.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*(3), 369–377.

McNamara, T. (1996). *Measuring Second Language Performance*. Harlow: Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York: Macmillan.

Met, M. (1998). Curriculum decision-making in content-based language teaching. In J. Cenoz & F. Genesee (Eds.), Beyond bilingualism: Multilingualism and multilingual education (pp. 35–63). Philadelphia, PA Multilingual Matters.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook*. Los Angeles: Sage.

Morse, J. M., & Niehaus, L. (2009). *Mixed method design: Principles and procedures*. Walnut Creek, CA: Left Coast Press.

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational and psychological assessment*. London: Sage.

Norwegian Directorate for Education and Training (2010a). *Erfaringer og vurdering av eksamen 2010 og 2011*. Retrieved from https://www.udir.no/Upload/Brev/Eksamen/5/Erfaringer_og_vurdering_av_eksamen_2010_og_2011.pdf

Norwegian Directorate for Education and Training (2010b). *Rundskriv Udir-1-2010: Individuell vurdering i grunnskolen og videregående opplæring etter forskrift til opplærngsloven kapittel 3*. Oslo: Directorate for Education and Research.

Norwegian Directorate for Education and Training (2017). Eksamensveiledning – om vurdering av eksamensbesvarelser. Retrieved from http://www.lokenasen.gs.ah.no/images/nyheter/2016-2017/Eksamen/ENG0012_Engelsk_Eksamensveiledning.pdf

Norwegian Ministry of Education and Research (2009). Forskrift til opplæringslova [Regulations to the Education Act]. Retrieved from https://lovdata.no/dokument/SF/forskrift/2006-06-23-724/KAPITTEL_4#KAPITTEL_4

Norwegian Ministry of Education and Research. (2006/2013). Læreplan i engelsk [English subject curriculum]. Oslo: Author. Retrieved from http://data.udir.no/kl06/ENG1-03.pdf?lang=eng

Nusche, D., Earl, L., Maxwell, W., & Shewbridge, C. (2012). *OECDs gjennomgang av evaluering og vurdering innen utdanning: Norway. [OECD review of evaluation and assessment of education: Norway]*. Retrieved from http://www.oecd.org/edu/school/Evaluation-and-Assessment_Norwegian-version.pdf

O'Sullivan, B. (2014). Assessing Speaking. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (Vol. 1, pp. 156–171). Chichester, UK: Wiley-Blackwell.

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.* Cambridge: Cambridge University Press.

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.

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 Oslo: http://www.udir.no/Upload/Forskning/5/Bedre_vurderingspraksis_ILS_rapport.pdf?epslanguage=no

Weir, C. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave MacMillan.

Yildiz, L. M. (2011). *English VG1 level oral examinations: How are they designed, conducted and assessed?*, (Unpublished MA thesis). University of Oslo, Oslo, Norway. Retrieved from https://www.duo.uio.no/bitstream/handle/10852/32421/YildizMaster.pdf?sequence=2&isAllowed=y