USABILITY OF THE TESTING CRITERIA FOR ENGLISH PROFICIENCY IN COMPETENCE TESTS: Assessor perceptions

Master's Thesis
Seija Purhonen

University of Jyväskylä
Department of Languages
English
August 2010

JYVÄSKYLÄN YLIOPISTO

Tiedekunta – Faculty Humanistinen tiedekunta	Laitos – Department Kielten laitos

Tekijä – Author
Seija Purhonen

Työn nimi – Title
USABILITY OF THE TESTING CRITERIA FOR ENGLISH PROFICIENCY IN COMPETENCE TESTS: Assessor perceptions

Oppiaine – Subject englanti	Työn laji – Level Pro gradu -tutkielma
Aika – Month and year elokuu 2010	Sivumäärä – Number of pages 101 sivua + 2 liitettä (7 sivua)

Tiivistelmä – Abstract

Tutkimuksen tarkoituksena on selvittää arvioijien näkemyksiä englannin kielen taidon arvioimiseen laadittujen näyttökriteerien käytettävyydestä hotellivirkailijan ja matkailun ohjelmapalvelujen ammattitutkinnoissa. Ilman ammatillista pätevyyttä pitkään työssä olleet aikuiset voivat hankkia ammattitutkinnon Suomessa osoittamalla ammattitaitonsa näyttötutkintojen avulla. Näyttöä vastaanottava oppilaitos laatii näyttökriteerit tutkinnon perusteissa määriteltyjen arviointikriteerien pohjalta.

Tarkastelun kohteina tässä tutkimuksessa eivät ole vain arvioijien yleiset näkemykset näyttökriteerien käytöstä pitkällä aikavälillä, vaan myös heidän mielipiteensä hyödyllisistä ja hyödyttömistä kriteereistä sekä kriteereihin toivotuista lisäyksistä ja poistoista. Lisäksi selvitetään, onko kriteerien käytössä havaittu ongelmia ja tulevaisuuden kehittämistoiveita. Tutkimusaineistona on viisi puolistrukturoitua teemahaastattelua. Haastateltavat edustavat yhtä oppilaitosta, minkä vuoksi tutkimus toteutetaan tapaustutkimuksena. Aineiston analyysimenetelmänä käytetään sisällönanalyysiä, koska tutkimuksen lähtökohtana on haastattelujen sisältö.

Tutkimuksen tulokset osoittavat, että englannin kielen näyttökriteereitä hotelli- ja matkailualan ammattitutkinnoissa pidetään käytettävyydeltään melko hyvinä, jos niihin tehdään tiettyjä muutoksia. Kielelliset, ammattispesifit ja aktiivista kulttuurista ymmärtämystä esiintuovat ominaisuudet tekevät joistakin kriteereistä hyödyllisiä, kun taas epämääräisyys saa aikaan hyödyttömyyttä. Epämääräisiksi tai väljiksi havaittuihin kriteereihin tarvittaisiin selkeyttä, samoin kuin hyväksytyn ja hylätyn suorituksen välisen rajamaaston kuvaamiseen. Näyttökriteerien tulkinnan ja työelämässä vaadittavan kielitaidon arvioinnin ongelmia voitaisiin vähentää arvioimalla mallisuorituksia arvioijakoulutuksen aikana. Kielenopettajilla pitäisi aina olla mahdollisuus osallistua ammattitutkinnossa edellytettävän kielitaidon arviointiin.

Asiasanat – Keywords Language testing, assessor perceptions, Competence-based Qualifications, interview, content analysis

Säilytyspaikka – Depository Kielten laitos

Muita tietoja – Additional information

TABLE OF CONTENTS

1 INTRODUCTION

Many adults face the problem of either not having participated in or completed some form of professional training even though they may have a long working history in their specific field of work. In order to provide adults with an opportunity to demonstrate their professional skills and gain a qualification irrespective of how the skills have been acquired, different ways to complete vocational qualifications have been introduced in several countries. In Finland, a system called Competence-based Qualifications (referred to hereafter as CBQs or the CBQ system) was launched in 1994. In many vocational sectors, such as hospitality, the use of a foreign language is quite readily counted as part of a worker's professional skills, which is why varying requirements of foreign language proficiency are included in many qualification requirements. In CBQs, professional skills are shown and assessed in competence tests, which are assessment situations that usually take place in the workplace of a candidate taking the test.

Unfortunately, the assessment criteria that provide the guidelines for assessing foreign language proficiency in CBQs are often rather imprecise. The following example from the Further Qualification for Hotel Receptionists (Hotellivirkailijan ammattitutkinto 2001) demonstrates this quite well. In each module of the qualification for receptionists, customer service is the target of assessment that includes foreign language proficiency. The assessment criteria for English are listed under customer service. When loosely translated into Finnish, the criteria state that besides serving customers in Finnish or Swedish, the successful candidate should provide service to customers using fluent English on the intermediate level (Hotellivirkailijan ammattitutkinto 2001:12-13). It is further specified that the intermediate level refers to the Finnish National Foreign Language Certificate.

These kinds of criteria, however, raise many questions as to what someone's fluent use of English on the intermediate level in customer service actually means. Therefore, more practical instruments, i.e. sets of pass-fail descriptions, referred to as *testing criteria* in this study, are designed by training providers responsible for competence test arrangements. The testing criteria then serve as the key tool for assessors when determining whether candidates pass or fail their tests.

The present study focuses on the use of the English testing criteria in two qualifications on the second level of CBQs, i.e. the further vocational qualifications. More specifically, it is the assessors' own perceptions of the usability of the criteria that are under investigation here. As a research topic, it provides me with an opportunity to explore a practical issue that has been referred to as a target of development among colleagues. My interest in examining the assessors' point of view originates in my own work experience on CBQs at a training institute for adults. Having worked as an assessor of English proficiency in competence tests, I have been involved in frequent discussions with my colleagues about applying the testing criteria designed for assessing English proficiency in the hospitality sector. Questions have been raised as to whether this type of criteria function the way they should from the assessors' point of view. In addition, the pass-fail descriptions have been the object of general criticism.

Therefore, the starting point for this study is to find out how assessors as everyday implementers of this particular assessment tool describe its usability. The research questions do not only focus on finding out about the general perceptions provided by the assessors interviewed, but issues such as useful and unhelpful criteria are also examined. Further targets of investigation are desired additions to or removals from the criteria, perceived problems, and expectations for future development regarding the use of the criteria. The data for this study are collected by conducting semi-structured interviews with five assessors of English proficiency. Content analysis is then used to analyse the data so that the communicative connections of the assessor perceptions (Holsti 1969, Titscher 2001) can be understood.

Assessment of foreign language proficiency in CBQ competence tests has mostly been examined in studies for vocational teacher education but not in other studies at universities. In addition, assessors' viewpoints on the testing criteria for English in CBQs have only been observed in a recent study by Härmälä (2008) but have not been the main object of investigation in any research yet. Härmälä's research indicated that there is a growing need to pay more attention to how foreign language proficiency requirements and assessment criteria are determined and expressed in the qualification requirements of CBQs.

Need for development regarding the assessment scales and criteria of foreign language proficiency has also been suggested by rater perceptions elsewhere. Raters have been reported to be very critical towards the scales used (Lumley 2002, Tarnanen 2002) and to be willing to participate in determining what kind of changes would be desirable. However, it must be noted that that raters often perceive the importance of single rating criteria in different ways (Eckes 2008). This is a problem that has been widely discussed in research related to communicative language testing (see e.g. Huhta 1993, Weir 1993). From the angle of the reliability and validity of assessment, it has also been argued (Orr 2002: 153) that paying more attention to rater training is vital.

The current study falls within the context of testing Languages for Specific Purposes, i.e. LSP testing, which has been much debated in recent language testing literature. One of the disputed issues has been how the concept of specific purpose language testing should be defined (see e.g. Douglas 2000 and 2001, Davies 2001). Some researchers, such as Douglas (2000: 2), emphasise that LSP testing means communicative testing with the test candidate’s perspective taken into account. Drawing on the diverse history of LSP testing (Davies (2001: 133-137), however, even the theoretical justification for LSP testing has been disputed.

A more detailed review of the theoretical issues related to language testing and assessment, as well as previous studies on assessor perceptions, follows in chapter 2 of the current study. Chapter 3 examines the working principles of the Finnish competence-based qualification system and, in particular, foreign language proficiency testing in two qualifications. The topics of chapter 4 are the research questions and the methods of data collection and analysis. Chapter 5 reports the findings. A summary of the findings with discussion is offered in chapter 6, which also includes an evaluation of the present study and suggestions for further research.

2 BACKGROUND TO THE STUDY

The work of assessors in the competence tests of competence-based qualifications can be characterised as passing or failing candidates according to their performance with the help of relevant testing criteria. This characterisation contains three elements which are central to the topic of the present study. The first one is the actual competence test, including the question of what is being tested and how. The second element involves the choice of the assessment procedure along with relevant testing criteria. The final element is the assessor, who makes the observations and decisions of passing or failing. To connect these elements through a theoretical framework, the following sections will correspondingly concentrate on language testing, foreign language proficiency assessment and related examples of previous studies on assessor perceptions. In addition, to provide further information on how these issues relate to the practices of gaining vocational qualifications in Finland, the working principles of the Finnish competence-based qualification system will be introduced in chapter 3.

2.1 Language testing

Focusing on competence-based testing of English in CBQs places the present study within the field of applied linguistics. More specifically, it is placed within the context of testing and assessing English for Specific Purposes. Competence-based language testing in CBQs comprises, as mentioned in chapter 1, assessment situations that usually take place in the workplace of a candidate and do not require candidates to participate in teaching prior to testing. The assessment process makes use of real-life testing and communicativeness. It is therefore relevant to examine communicative language testing more closely in the following section. Another aspect of language testing that is directly applicable to the current study is English for Specific Purposes testing, i.e. ESP testing, which will be discussed in section 2.1.2.

2.1.1 Communicative language testing

In this section, the research of Bachman (1990), Bachman and Palmer (1996), and Weir (1988 and 1993) will be examined to illustrate what communicative language testing actually means. All three researchers have been involved in developing influential models for understanding and applying communicative measures in language testing. In addition, the core of communicativeness in contemporary language testing will be reviewed (Fulcher 2000).

However, before examining the principles of communicative language testing and the impact of its expansion, let us briefly focus on its past as action against some features of language testing traditions. Fulcher (2000: 484-487) offers a description of the principles of language testing in the early 20th century and argues that changes in language testing were already encouraged by language test developers well before the bloom of communicativeness in the late 1970s and early 1980s. He explains that the origins of communicative language testing can be traced to the growing desire to remove the domination of multiple choice tests and to re-examine the concepts of validity and reliability as they had been perceived in the 1960s. It is pointed out that performance, with assessment procedures that were qualitative and subjective, became to be considered a key part of communicative testing. These early developments were a reaction against assessments that appeared to show little appreciation of test candidates as individuals. The most important concepts of the initial stages in communicative language testing were “1. real life tasks; 2. face validity; 3. authenticity; and 4. performance.” (Fulcher (2000: 484). Furthermore, Fulcher (2000: 490) remarks that the development of communicative language testing was useful for the testing of Languages for Specific Purposes (LSP) in that authenticity and context were given more emphasis in test development and research. The relationship of authenticity and real-life testing will be discussed in more detail in section 2.2.3 of this study.

Also dealing with the development of language testing during the 1980s and 1990s, Bachman (2000) describes how the introduction of the communicative approach and other advances in language testing served to broaden the view of language proficiency. He points out (Bachman 2000: 3-4) that this meant major changes in the research of language testing, with a gradual shift from merely measuring the four skills of speaking, listening, reading and writing into accepting a view of language use that includes the role of discourse and sociolinguistic features. A similar remark is made by Fulcher (2000: 489), who argues that making inferences from performance tests changed from simply forming predictions of how test takers would perform in real life into understanding the abstract nature of performances. McNamara (1996: 35) points out that it has been fairly common for communicative language tests to be based on performance.

As has been implied above, the impact of communicative language testing on the research and development of language tests in the late 20th century was significant. As Weir (1988: 9) already put it more than twenty years ago, “the emphasis is no longer on linguistic accuracy, but on the ability to function effectively through language in particular contexts of situation”. In the early 1990s, Weir (1993) formulated the essence of communicative language testing as follows:

> This communicative/real-life approach in testing might be said to be characterised by the following features (in no particular order): focus on meaning, contextualisation, activity has an acceptable purpose (reasonable suspension of disbelief), realistic discourse processing, use of genuine stimulus material, authentic operations on texts, unpredictable outcomes, interaction based, performance under real psychological conditions, e.g. time pressure and in assessment of performance on a task, judgements made on achievement of communicative purposes. (Weir 1993: 167)

Weir argues (1988: 11-12, 28) that apart from an emphasis on the differences in content or rating, a communicative language test does not necessarily differ very much from other tests. Regarding the test content he says that even though not unproblematic, the authenticity of tasks is considered very important. As regards the criteria, concentration on the whole and not on separate parts is recognised as an element of the directness of test criteria, as well as of content. Including a maximum number of real-life language characteristics in the criteria is presented as essential for communicative language tests. Furthermore, sociolinguistic aspects must not go unrecognised. Therefore, Weir urges test developers to make sure that the context in a test situation is a reflection of realistic, real-life language use. He underlines the question of “what and how to sample” (Weir 1993: 29) as something that has to be paid careful attention to in language testing. The challenges of test validation are recognised and it is suggested (Weir 1993: 28) that a better degree of validity is achievable, if three issues are paid due attention to in preparation for testing. Firstly, it should be properly defined what it is that a student actually has to do in the intended language use situation. Secondly, the performance conditions should be established explicitly. Finally, the required performance level should be clearly indicated.

In addition, Weir (1988: 8) briefly describes the main theoretical arguments by Hymes and by Morrow on communicative competence, explaining that their work in shaping communicative language testing has been central. Furthermore, Weir also acknowledges the work of Canale and Swain, whose definition of communicative competence has been modified by Bachman (1990) and by Bachman and Palmer (1996). Communicative competence according to Canale and Swain (Canale 1983, Canale and Swain 1980) was made up of four parts, which were named grammatical competence, discourse competence, strategic competence, and sociolinguistic competence. Building on the principles of Canale and Swain, Bachman (1990: 84) proposes a model of communicative language ability, i.e. CLA, which is portrayed as a framework with three elements: language competence, strategic competence, and psychophysiological mechanisms. Bachman argues that this model answers the need to bring an individual’s ability to use his or her competence in actual communication into language testing. Moreover, he explains that CLA contains elements familiar from earlier research on communicative competence. However, Bachman arranges those elements differently and in more depth than has been done in earlier research. He also gives reasons for his decisions, such as previous inconclusive empirical research (Bachman 1990: 86) or the importance of putting certain components together (Bachman 1990: 89).

A further development of Bachman’s model was made by Bachman and Palmer, whose principles have provided a theoretical starting point for many researchers and have been applied in developing English language tests, such as the IELTS (Clapham 2000: 149). Bachman and Palmer (1996: 61) argue that language ability is the feature that language testing is primarily concerned with. They add that the concept of *language ability*, i.e. “language knowledge and strategic competence, or metacognitive strategies” (Bachman and Palmer 1996: 62), should be perceived as one of the interactional parts of language use. Furthermore, in order to make conclusions about someone's language ability, clarifying the relationship between test performance and language use beyond the test is necessary. Language use (Bachman and Palmer 1996: 61) is described "as the creation or interpretation of intended meanings in discourse by an individual, or as the dynamic and interactive negotiation of intended meanings between two or more individuals in a particular situation". On the implementation of language tests, Bachman and Palmer (1996: 17) emphasise that "the most important quality of a test is its usefulness". They go on to present their view on the concept of test usefulness as a fusion of the following qualities: reliability, construct validity, authenticity, interactiveness, impact, and practicality. It can be concluded that of those qualities practicality also seems to be the main motivating force in Bachman and Palmer's approach to language testing.

Moving onto the 21st century, a view on what constitutes the core of communicativeness in language testing is provided by Fulcher, who describes its key elements as follows (Fulcher 2000: 489-493): authenticity, performance and real-life scoring. The first element, i.e. authenticity, is described as entailing three features. Firstly, the purpose of an authentic test has to be distinguished as communicative by test candidates. Secondly, the tests should not include deliberately simplified instructions. Furthermore, the test candidates should be tested with regard to the linguistic context as well as to their ability to deal with particular features of real situations, such as the formality of a situation or being able to communicate attitudes. It is pointed out, however, that tests are instruments designed on the basis of theories, and that they are not imitations of real life. The second element, i.e. performance in the tests, is characterised as interaction that requires the test participants to express themselves through language use. The produced language is seen as unpredictable by nature since it is created instantaneously. Finally, the third element determining communicative language tests, i.e. real-life scoring, is described as scoring of performances with regard to whether the desired communicative effect has been reached by the test takers through appropriate communicative behaviour in the test. In addition, setting up rating scales that affirm what is measured is considered critical.

There are, however, various problems which have been connected with communicative language testing (Huhta 1993: 84). Firstly, a problem that occurs before testing is knowing whether the various versions of a test differ in difficulty. Secondly, the social situation with its features affecting oral testing can be problematic. Thirdly, subjectivity of assessment is also suggested as one of the problematic issues. Moreover, ensuring test validity and reliability of communicative language tests has also been an area of concern for many researchers (see e.g. Weir 1988, Messick 1994 and 1996, Bachman and Palmer 1996, Douglas 2001). Furthermore, on problems concerning rating scale design Fulcher (2000: 493) claims that until the late 1990s much of it was done "purely on arm-chair notions of language development and structure". He sees the increasing empirical research to validate rating scales as a positive trend for the future of communicative language testing. In their discussion on scoring methods, Bachman and Palmer (1996: 219) also mention the use of rating scales as often being problematic, since scales are not always seen as efficient or very reliable. They suggest, however, that such problems can largely be overcome with proper planning.

2.1.2 Languages for specific purposes testing

As the present study involves the use and testing of English in two further qualifications within the field of hospitality, testing languages for specific purposes, i.e. LSP testing, will be examined in this section. Determining the notion of LSP and finding justification for LSP testing have been much debated in recent language testing literature (see e.g. Cumming 2001, Douglas 2000 and 2001, Davies 2001, Elder 2001, Swales 2000). For the purposes of this study, the work of Davies and Douglas will be described as examples of research with somewhat differing viewpoints on the theoretical basis of LSP testing. Later in this section, two examples of LSP focusing on English for specific purposes will be introduced.

As implied above, Davies (2001) and Douglas (2000, 2001) look at LSP testing from somewhat different points of view. Already at the beginning of his book Douglas (2000: 2-3) argues for the existing necessity of LSP testing. He claims that there is sufficient theoretical ground to justify the use of LSP tests, which fulfil the requirements of test reliability and validity. Maintaining that early forms of LSP testing were already seen in the first half of the 20th century, he adds that every LSP test does not necessarily include all the essential features of specific purpose testing. An LSP test is defined in the following way:

> A specific purpose language test is one in which test content and methods are derived from an analysis of a specific purpose target language use situation, so that test tasks and content are authentically representative of tasks in the target situation, allowing for an interaction between the test taker's language ability and specific purpose content knowledge, on the one hand, and the test tasks on the other. Such a test allows us to make inferences about a test taker's capacity to use language in the specific purpose domain. (Douglas 2000: 19.)

But how does one tell the difference between a specific purpose language test and a general language test? Douglas suggests the following way of differentiation. The traditional general purpose test tasks (Douglas 2001: 172) can be identified as being compiled using a theory-based course programme or a language theory as the foundation. As for LSP tests (Douglas 2000: 2), they have two distinguishing features. One feature is the authenticity of test tasks, meaning that the tasks used in tests should in essence be similar to those in real situations relating to the specific purpose in question. The interactive relationship of language knowledge and background knowledge is, however, mentioned as the most important feature. This relationship is seen as a fundamental element of what Douglas (2000: 33-36) calls *specific purpose language ability*, which consists of language knowledge, strategic competence and background knowledge.

Building his framework for LSP testing on Weir, Widdowson, Bachman and Palmer, Douglas (2000: 9-14, 28) emphasises that LSP testing means communicative testing with the test taker's perspective taken into account. The problem of how far to generalise inferences from LSP tests is acknowledged, with the implication that no language test is either purely general or specific, as the extent of specificity varies in all tests. Therefore, language testers should distinguish performance from ability and thus come closer to involving test takers in tasks that are typical of the intended language use situation. The key point in LSP testing (Douglas 2000: 27) is to understand the aim of concentrating on a test taker's underlying language ability in a testing situation instead of the performance as an achievement. The emphasis is on communicativeness, in other words, specific purpose language ability. Since precision, such as technical jargon, is typical of specific purpose language use according to Douglas (2000: 19, 88), justification for LSP testing is found in this precise nature of specific purpose language and in the variation of performances produced in different contexts. For testing language use in specific contexts it is essential (Douglas 2000: 282) that not only the language knowledge of the test candidates is involved but also the background knowledge that they have, as well as their strategic competence.

With a different starting point to specific purposes testing, Davies remarks (2001: 133-137, 142) that the history of LSP involves such diverse entities as phrase books for tourists and German in the field of chemistry. The growth of the LSP testing movement in the 1970s is criticised for putting too much weight on direct testing and for paying insufficient attention to the implementation of tests. Davies questions the theoretical justification of traditional LSP testing, which is based on specific language varieties such as medical, legal or business English, because varieties are seen as indistinct and inconsistent. Furthermore, they overlap with each other and contain sub-categories within themselves, such as paediatric or surgical English within medical English. Examining the connection of language varieties to registers, Davies claims that the factor separating the varieties from each other is not the language as such but the content. The conclusion made is that LSP is not connected with register. Regarding the essence of LSP testing, it is argued that it "cannot be about testing for subject specific knowledge. It must be about testing for the ability/abilities to manipulate language functions appropriately in a wide variety of ways" (Davies (2001: 143). It is further implied that a general language test could thus be no different from a specific purpose test.

Concerning the practical problems in LSP testing, Davies (2001: 138, 143-144) refers to the difficulty of defining the content for test tasks, as well as to the challenges of administration and operationalisation in large scale tests, such as the IELTS. An additional challenge is found in the fixed nature of certain specific purpose languages, such as in air traffic control. For the persons involved, it is considered important that the specific language in such cases is based on a more general proficiency, so that they can take care of unexpected situations. Nevertheless, it is concluded that LSP tests can be considered feasible because pragmatic justification for using them exists and their predictive value is not weaker than that of general purpose tests.

Even though their starting points are different, as was seen above, both Davies (2001) and Douglas (2000) conclude that LSP testing does have a future if the debated issues are taken under careful inspection in test development. They both explain that developing an LSP test normally begins by making an appropriate needs analysis, and they acknowledge the communicative nature of LSP tests. However, Davies' conclusions are based on disputing the theoretical foundation of LSP tests and on recognising the practical need for LSP testing, whereas Douglas builds his case on a functioning theoretical framework for LSP testing.

For illustrations on the practices of LSP testing, let us now briefly turn to the research of McNamara (1996) and a study by Blue and Harun (2003). With a special focus on the Australian Occupational English Test, i.e. OET, McNamara (1996) connects LSP testing with second language performance testing. The OET is described as a government second language test for professionals in the field of health services. In his discussion on second language testing methods, McNamara (1996: 13-15) examines the research of Slater (1980) and Jones (1985), explaining that they identify three different kinds of tests in second language performance testing. The first one is direct assessment, which involves observing candidates in their place of work. As an example of this, McNamara mentions new teachers, who are not native English speakers, being evaluated on the basis of on-the-job performance before final employment. The second type, which is called work sample methods, also includes assessment at work. Aiming at standardised assessment, teachers' language proficiency is assessed as part of their teaching skills. The third test form, which is called simulation techniques, is not as concretely tied to the workplace as the first two test types. The particular features of this test type are that the test tasks are more abstract or imaginary and that predictions are made on how the test candidates would do in similar real situations. McNamara then combines the first two test types to be called *work sample tests*. Thus, the OET (McNamara 1996: 87, 93) is currently a work sample test of occupational English aimed at various groups of health professionals migrating to Australia, but in its earlier forms it used to be a general English proficiency test. The OET is the first of three stages in the registration procedure foreign health professionals have to go through to be able to practise their profession in Australia. McNamara (1996: 93) points out that a critical phase in the work sample test development is deciding on the test content.

For a portrayal of the content of a language for specific purposes, an investigation into hospitality English by Blue and Harun (2003) deserves to be introduced in this context. Examining the use of English as hospitality language in hotels in Britain, they define hospitality language as "*all* linguistic expressions which relate to and represent hospitality concerns" (Blue and Harun 2003: 74; emphasis original) and maintain that it had not yet been researched in depth. Furthermore, they describe it as evolving "from a combination of procedural, behavioural and linguistic acts, verbal and non-verbal, direct and non-direct" (Blue and Harun 2003: 89). The use of hospitality English is seen as ranging from hotels and travels agents to information desks and tourist sights.

Blue and Harun (2003: 77-78, 89-90) argue that due to globalised travel, English is often used and even required in an increasing number of hotel and travel service situations. The following language skills are listed as necessary for serving hotel guests, applicable to English as well as to other languages: addressing people, requesting and giving information, reacting to requests, dealing with prompts, using body language, taking care of difficult customers, and finding satisfying solutions to complaints. However, it is pointed out that since many customer service situations can be considered at least partly culture-bound, good service means being locally and cross-culturally well oriented. What is also emphasised is that hospitality English involves a large amount of variation, and cannot be regarded as something permanent and unchanging by nature despite its somewhat standardised form due to worldwide use.

On teaching the essentials of hospitality language, Blue and Harun (2003: 89) criticise existing textbooks for hotel staff for not being authentic enough and for containing language that is too simple. They call for more exposure to real-life communication. Although their focus is not on the testing of hospitality language, the authors do conclude (Blue and Harun 2003: 86, 87) that an appropriate proficiency of hospitality language should be required of staff in their native language as well as in a foreign language.

The practical approach in Blue and Harun's study (2003) on hospitality English appears to demonstrate the point made by Davies (2001) above that LSP testing could be justified specifically for its practical qualities and outcomes. In addition, Blue and Harun's (2003) argumentation promoting culturally aware customer service seems to agree with Douglas' (2000) emphasis on the role of background knowledge.

2.2 Assessment of foreign language proficiency

Language proficiency has been defined in many different ways, which is why it has not been easy for researchers to reach an understanding on how to measure it in general (Canale 1983: 334, Chalhoub-Deville 1997: 4). Let us therefore first consider the diversity of the concept of language proficiency before moving on to three topics relevant to the assessment of foreign language proficiency in the Finnish CBQs, i.e. assessment procedure (see 2.2.1), formulation of assessment criteria (see 2.2.2), and the relationship between authenticity and real-life assessment (see 2.2.3).

Some ways of understanding language proficiency, such as Bachman's (1990) model of communicative language ability, as well as Bachman and Palmer's (1996) revision of it, were already briefly introduced in connection with communicative language testing in section 2.1.1. According to Chalhoub-Deville (1997: 4), these models are examples of examining proficiency through its components. Similarly, the model proposed by Cummins (1979; 1983 as quoted by Chalhoub-Deville 1997: 5) is based on components of proficiency, the first of which is called *cognitive/academic language proficiency*, i.e. CALP, and the second is called *basic interpersonal and communicative skills*, i.e. BICS. It is further explained that BICS refers to informal contexts, whereas CALP concerns academic contexts. Chalhoub-Deville adds (1997: 4, 10, 11) that another way to define language proficiency is to focus on its levels, through which students' development can be portrayed. Furthermore, she points out that many models of second language proficiency have been questioned for their combination of components, insufficient empirical evidence, improper use of statistics, and for not being useful enough in practice due to their abstract or complex nature. Therefore, Chalhoub-Deville calls for a division to be made between operational models relating to contexts and theoretical models offering more general descriptions of language proficiency.

Yet another way to express the essence of language proficiency is formulated by Canale (1983: 334, 338-340), who argues that it should be determined in general terms rather than consisting of numerous separate components. Therefore, he proposes a combination of three features of language proficiency. The first feature, i.e. basic language proficiency, involves the general principles of all language use, called *biological universals*. It is pointed out that basic language proficiency does not only include universals concerned with grammar, but also those concerned with discourse, sociolinguistics and processing, as well as strategic universals. The second feature, i.e. communicative language proficiency, has an interpersonal and sociolinguistic emphasis, concentrating on how language is used by people in oral and written communication. The third feature, i.e. autonomous language proficiency, is concerned with grammatical rules and forms, concrete meanings, and individual language uses. Using creative language, dealing with problems or counting the change in a shop are examples of these.

Regarding the topic of the current section, not only is the concept of language proficiency understood in many ways, but also the meaning of the expressions *assessor* and *assessment* may differ depending on the user. On the one hand (Clapham 2000: 150), *assessor* and *assessment* may imply lesser concern for the formality, standardisation, reliability or validity of tests than the terms *tester* and *testing*, while on the other hand, they may be used in a general sense to include different methods and types of testing and assessment. Furthermore, *assessor* and *assessment* may also be used as the synonyms of *tester* and *testing*. In addition to the above-mentioned terms, there is yet another pair of terms used in many studies to express the same relationship: *rater* and *rating*. In the present study, no difference is implied between any terms of rating, testing and assessment, as they are used in a synonymous and general sense, respecting the original source.

2.2.1. Assessment procedure

When performing their task raters, testers or assessors of foreign language proficiency encounter the fact that test designers have already made decisions about the assessment procedure to be used in a particular assessment situation. In this section, three procedural issues also relevant in CBQ competence testing are examined: testing in and out of classrooms, deciding whether to carry out a holistic or an analytic assessment, and choosing between norm-referenced and criterion-referenced assessment. For an extensive general review of assessment procedures for evaluating foreign language proficiency, see for example the *Common European Framework of Reference for Languages* (2010).

In a study on rater behaviour, Tarnanen (2002: 48-51) comments on the first issue, i.e. testing in and out of classrooms. She suggests reviewing assessment from the point of view of either teaching or testing. In relation to teaching, she speaks of classroom assessment, and regarding testing she refers to assessment of skill levels. The latter is concerned with particular language use situations and the abilities and skills of individuals in those situations, whereas classroom assessment is more learning centred and also involves other features such as motivation, attitudes and social issues. It is pointed out, however, that these two do sometimes overlap. The clearest difference between them is identified in the methods of assessment. In skill levels assessment, the methods include test tasks, interviews and surveys. This type of assessment is explained to have a predictive character, even though it is seen as assessment of a particular situation. The methods used in classroom assessment include not only tests and interviews but also self-assessment, use of portfolios, observation, project work and different exercises. Regarding the present study, the overlap of the above-mentioned methods is common practice in the competence tests of the Finnish competence-based qualifications system.

The second procedural concern relevant in the present study is whether the assessment should be completed by using a holistic or an analytic rating scale. Holistic rating scales require the assessors to give one rating of a performance as a whole (McNamara 2000: 43). In connection with assessing writing, Cooper (1977 as quoted by Fulcher 2003: 89) characterises holistic assessment as “any procedure which stops short of enumerating linguistic, rhetorical, or informational features of a piece of writing.” He adds that special aspects of the writing being assessed may be paid attention to or rated but never by counting their occurrence. Contrary to the holistic assessment procedure, analytic assessment involves separate assessment of all the features of a particular performance (McNamara 2000: 43-44), and several rating scales are needed in order to include each feature. Thus, when assessing a speaking test, for example, this could mean that the assessors are required to rate such features as pronunciation, vocabulary, use of grammar, correctness and fluency separately. As to the usefulness and reliability of these two methods, Weigle (2007: 203) comments that based on research it is the analytic assessment which seems to be the more reliable one. She adds that getting feedback through the different scores will also help students learn about their strengths and weaknesses. She points out, however, that assessment can be carried out more quickly and efficiently by using holistic scales. Nevertheless, common practice in the Finnish CBQs tends to be that these two scales are often used together in various ways to reach the final assessment. As McNamara mentions (2000: 44), an assessment report might show a single score but actually consist of the scores of separately assessed features. This description corresponds with the assessment procedure of the second level of CBQs, i.e. further vocational qualifications, which is also the level focused on in this study. Assessment in the competence tests of further vocational qualifications is completed on several separate features but finally reported in the form of a single score, either a pass or a fail (see sections 3.2.2 and 3.2.3).

The third CBQ-related issue examined in this section is choosing between the procedures that Bachman (1990: 72-73) calls *norm-referenced* and *criterion-referenced assessment*. In norm-referenced assessment, a candidate’s performance is compared with the test results of other candidates. The reference points forming the norms in a test are drawn from the test results of either a special norm group or the actual group of candidates taking the test in question. The latter approach is illustrated by a case where the candidates falling within the top ten percent are given the highest score, while the bottom ten percent do not pass at all. Carrying out an assessment according to specific criteria based on agreed subject content or on set skill levels (Bachman 1990: 74, 211) means that it is criterion-referenced. While norm-referenced assessment is usually used to compare candidates, for example, in a selection procedure, criterion-referenced assessment is used to discover whether an individual possesses a particular proficiency. This is the case in assessing competences, for instance. It is thus possible that all criterion-referenced test candidates are above the set level or that they master the required content. Because of their different uses (Bachman 1990: 75), norm-referenced and criterion-referenced tests also differ in design as well as scale construction and interpretation.

Consequently, it is the criterion-referenced interpretation of test results that is used in the assessment procedure of the Finnish CBQs (Haltia and Hämäläinen 1999: 59), since the point of testing foreign language proficiency in competence tests is not to rank the candidates as such but to determine each candidate’s language proficiency with the help of preset level requirements.

2.2.2. Formulating the assessment criteria

Apart from making decisions about assessment procedures, test designers need to be able to formulate relevant assessment criteria for assessing second or foreign language proficiency. Weir (1993: 40) comments on the importance of criteria selection by saying that “the relationship between a task and the criteria that can be applied to its product, [sic] is an essential factor in taking decisions on what to include in a test of spoken or written production”. McNamara (1996: 18) points out that deciding how to evaluate performances is a vital step in designing tests and that the importance of comprehending the difference between a test and a criterion (McNamara 2000: 8) should be emphasised as well.

How then should a criterion be characterised when assessing language proficiency in a performance test, for example, in the competence tests of CBQs? Referring to performance assessment as a method requiring “actual performances of relevant tasks” (McNamara 1996: 6) instead of involving a demonstration of certain knowledge, two practical definitions for a criterion are suggested by McNamara. He defines a criterion (McNamara 2000: 8, 132) as “relevant communicative behaviour in the target situation” and as “an aspect of performance which is evaluated in test scoring, e.g. fluency, accuracy etc.”. The following figure, which is an adaptation of the original by McNamara (2000: 8), illustrates the kind of influence the criterion has on the test.

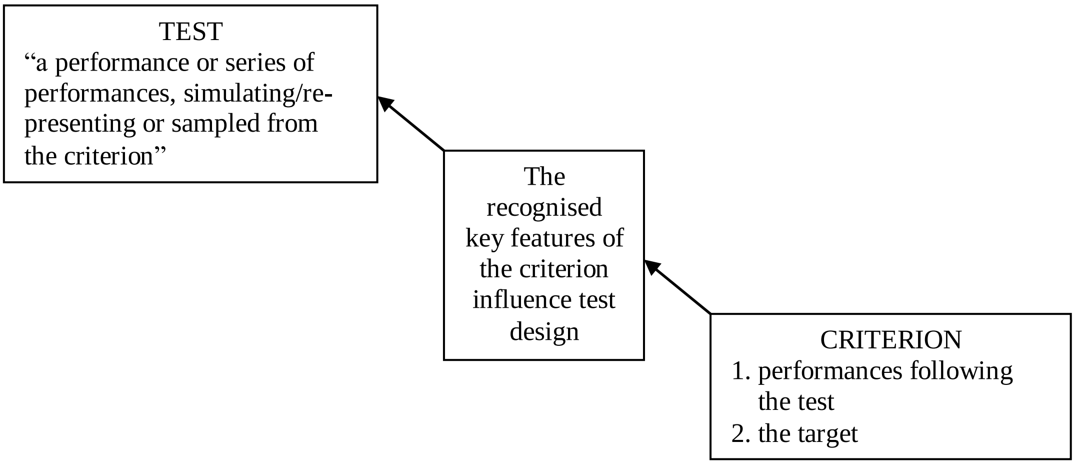

TEST
“a performance or series of performances, simulating/re-presenting or sampled from the criterion”
The recognised key features of the criterion influence test design
CRITERION
1. performances following the test
2. the target

Figure 1. Influence of the criterion on test design (adapted from McNamara 2000: 8)

What the figure implies is that performances in test situations can be used to make inferences about a candidate’s language use in a similar situation (McNamara 2000: 8-9), while bearing in mind the difference between the criterion and the test, as well as the fact that tests represent simulations of real-life situations. It is further pointed out that criterion behaviour is not observable as such, since the test tasks derived from the criterion cannot be real even though they may be fairly realistic (see also McNamara 1996: 10-12).

However, devising relevant criteria is not a simple operation. Concerning the formulation of assessment criteria for second language proficiency assessment in LSP contexts, Jacoby and McNamara (1999: 214) as well as Douglas (2001: 174) argue that many of the real-life criteria used in such assessments are not based on proper analysis of the spoken interaction among respective professionals. Instead, they point out, the criteria of LSP tests are often based on theories of language ability in the same way as in more traditional language tests. In their study, Jacoby and McNamara collected examples of *indigenous assessment criteria* (Jacoby and McNamara 1999: 214) from physicists in the United States and compared their results with the criteria in use in the Australian Occupational English Test. Indigenous assessment criteria are defined by Jacoby (1998, as quoted by Douglas 2000: 68) as criteria that the specialists of a particular profession apply to assessing a trainee’s communicative performance. One of the findings by Jacoby and McNamara (1999: 235-236) was that linguistic performance could not be separated from professional performance in indigenous assessments. In their view, this could mean that there is a clear difference between the criteria formulated by linguists and those formulated by subject professionals. It is pointed out that using criteria specific to a particular profession not only leads to the problem of generalisability but also raises the question of whether or not generalising in professional performance assessments is desirable. Through this and other similar issues raised the authors wish to encourage more research and cooperation regarding LSP criteria.

The views of Jacoby and McNamara are echoed by Douglas (2001: 173-174, 179-180), who maintains that drawing LSP assessment criteria from actual language use situations can offer test designers a better understanding of the range of criteria to use. He continues by suggesting that "an analysis of the indigenous assessment criteria in the specific purpose domain in which we are attempting to develop a language test could serve as a framework for the development of assessment criteria in the test domain". He points out that this would be a way of enhancing the good practices of language testing with the relevant features of real-life assessment. Douglas (2001: 181-183) also suggests several methods for analysing relevant target situations and defining the criteria. A recommendation offered to LSP test designers is to use indigenous criteria for two purposes: to serve as aids in making inferences from LSP performances and to enhance the linguistically oriented criteria.

Regardless of the way they have been gathered and defined, the assessment criteria of foreign language proficiency in performance tests are connected with tasks from the real world along with the question of authenticity, which will be examined more closely in the next section.

2.2.3. Authenticity and real-life tasks

The emphasis in the assessment procedure of the Finnish CBQs, including language proficiency assessment, is on assessing a candidate's skills primarily in actual working situations, in a working environment which should be real or as real as possible (Näyttötutkinto-opas 2003). This requirement necessitates an introduction to the debate about authenticity and real-life tasks in the field of second language testing.

Bachman (1990: 300-303, 308-310, 330) claims that defining authenticity is “one of the most difficult problems for language testing”. He continues by observing that for a long time there has been a genuine interest in language testing research to grasp the real meaning of authenticity by attempting to fit the core parts of language use into language tests. Two classifications for authenticity are offered. Firstly, authenticity can be described through what is called the *real-life approach*, which involves the aim of developing tests with as much imitation of the reality of non-test language performance as possible. The test output is then understood to be an inference of the test candidate’s future performance in an actual real-life situation. Secondly, authenticity is defined through the *interactional/ability approach*. In this approach, the importance of language ability and communicative use of language are emphasised. Instead of mirroring real-life performances with a predictive character, test performances are used to make inferences of the test candidates’ communicative language abilities. In several contexts, the real-life approach is criticised, for example, for its deficiencies in validity and for failing to recognise the difference between behaviour and language ability. However, Bachman (1990: 356) remarks that practical tests intended for homogenous groups can benefit from the principles of the real-life approach.

Together with Palmer, Bachman has further developed the definition of authenticity in language tests as part of the concept of *test usefulness* (Bachman and Palmer 1996: 18). The authors describe authenticity (Bachman and Palmer 1996: 23, 25-26) as “the degree of correspondence of the characteristics of a given language test task to the features of a TLU task”, separating it from interactiveness, which is defined as “the extent and type of involvement of the test taker’s individual characteristics in accomplishing a test task”. Interactiveness is presented as a feature of all test tasks (Bachman and Palmer 1996: 28-29) while authenticity always remains relative, which is why it is considered important to find a balance between the degree of authenticity and the other test features when designing language tests.

Bachman and Palmer’s description of the dual character of authenticity has been adopted into LSP testing by Douglas. The key element of LSP tests (Douglas 2000: 12-13, 20, 39, 57-58) is the authentic involvement of test candidates in all test tasks, which are connected with real situations of the target language. However, Douglas acknowledges the problem of using one test performance as the basis for making predictions, as well as the problem of dealing with background knowledge. He also calls attention to the different characters of the input data and responses in and out of tests, and maintains that the authenticity of the input and responses may be lost when their framework of occurrence changes from the original situation and interaction into that of language tests. Thus, their authenticity can only be maintained by examining the elements they have in common with the features of the target situation.

Along with Bachman and Palmer, other researchers, for example, Weir (1993), agree that no language test can be fully authentic. Some researchers have also presented critical views of how authenticity has been perceived in language testing contexts (see for example Spolsky 1985, Messick 1994, Lewkowicz 1997). Arguing that authenticity is perhaps not as important to other participants in the testing process as it is to assessment theorists, Lewkowicz (2000: 44, 50-52, 60) advocates systematic research into the qualities of authenticity. Furthermore, she raises a number of questions as to the role of authenticity regarding the characteristics of test tasks and their relationship with the target language use situations, while other issues raised include the way authenticity is perceived by different parties in the testing process. In this study on student perceptions, no basis was found for the assumption that all students would agree on the authenticity of a specific language test.

All the issues dealt with in sections 2.2.1 - 2.2.3 are present in the assessment procedures put into practice by assessors in the competence tests of CBQs. Since the present study is focused on assessor perceptions concerning the testing criteria in those competence tests, the next sections will concentrate on related studies carried out on the perceptions of assessors.

2.3 Previous studies on assessor perceptions

As described in the previous three sections, the work of raters or assessors in language proficiency testing involves a number of fairly complex issues concerning assessment procedures, criteria and tasks. Regarding studies involving raters and assessors, it appears that the research carried out in the past two decades has covered a wide variety of topics. Studies have been carried out for example, on performance assessment and rater behaviour by McNamara (1996), on possible differences between occupational expert raters' and language expert raters' perceptions of the proficiency of test candidates by Lumley (1998), and on changes in rater behaviour by Elder et al. (2007). However, the way assessors perceive the usability of foreign language proficiency testing criteria in a vocational setting has so far remained an unexplored issue in previous research.

In the present study, it is the work and perceptions of assessors that are investigated from the point of view of criteria usability. More specifically, this study examines assessor perceptions concerning the usability of a particular type of criteria, i.e. the foreign language proficiency testing criteria in Finnish competence-based qualifications. Since the testing criteria of Finnish CBQs are drawn from the more general assessment criteria, the next three sections will offer a review of previous studies on how assessors of language proficiency perceive the assessment criteria they use. Firstly, a view into the scope of problems perceived by assessors in applying given scales and criteria will be introduced in section 2.3.1, followed by a look at how assessors perceive the importance of single criteria in section 2.3.2. Finally, section 2.3.3 will concentrate on perceptions concerning the rating scales in Finnish CBQs and degrees of criterion importance.

2.3.1 Problems with given scales and criteria

With the aim to find out whether raters agree on a given rating scale and criteria, and on how they apply them, Lumley (2002: 257-263) conducted a study on the assessment process of the Special Test of English Proficiency, a language proficiency test used for immigration purposes by the Australian government. Concentrating on the writing section of the test, Lumley discovered that the rating process itself could be described as systematic but not without problems regarding the scale and criteria. Firstly, the scale wording was found to be problematic by the raters. Problems with the wording led one rater to value one criterion, i.e. clarity of meaning, over another, i.e. relevance of response. Another problem that came up was general dissatisfaction with the scale, causing raters to make decisions based more on feeling than on the given criteria. One rater, for example, conveyed that she was not happy with the rating scale, and compared the text she was going through with another text, trying to remember why she had rated the other one as a two on a scale of 0 to 5.

Furthermore, Lumley (2002: 263-268) discovered that there was a problem relating to a conflict between the actual features of a text and the level descriptors. Regarding the category of task fulfilment and appropriacy, one of the raters found it problematic to rate a text that she considered appropriate but too short. Lumley defined this problem as a missing construct, calling it the quantity of ideas in a text. Another missing criteria element defined was the lack of conjunctions and other linking relationships that build cohesion, which was used as a kind of specific self-made criterion by at least two raters. Summarising his findings, Lumley remarks that rating scales are not to be understood as authoritative interpretations of language proficiency. Furthermore, he promotes further research into the issues of validity, central role of raters in the rating process, as well as rater training.

Within the context of testing Finnish as a second language, Tarnanen (2002: 34, 43) carried out a study on raters and rating behaviour, including rater perceptions on the rating process and scales. The raters completed three rounds of rating adult students' written performances. For the first round, the rating scale of 1-6 was defined by the raters themselves, but during the other rounds a given scale was used. Following each round the raters were interviewed.

Tarnanen (2002: 195-227) reported that the given scale and criteria applied in her study were treated with a multitude of opinions and feelings by the raters. There were numerous perceptions which were expressed by all or most of the raters. For example, the scale contents received criticism from all the raters: the contents were seen as too vague and sketchy. Many changes to the scale were suggested by the raters, such as the addition of more concrete criteria and more detailed definition of performance level descriptions.

When faced with problematic borderline cases (Tarnanen 2002: 212, 226-227, 270-271), the criteria element that was most unanimously called attention to by the raters was the structural command of language. As to the types of structures in those cases, the raters often referred to a particular feature, such as correct verb forms. In general, the raters were found to value structural accuracy as perhaps the most important criterion of good writing. However, acknowledging the importance of rater training whenever raters make interpretations of rating scales, it was pointed out that due to lack of resources, the assessors in this particular study had no previous training in the use of the given scale.

2.3.2 Criterion focus

In a study on how raters perceive test-takers' performances, Orr (2002) made similar observations as Lumley (2002) and Tarnanen (2002) on raters not paying attention to all the rating criteria in the same way. Concentrating on the speaking section of the First Certificate in English, the raters' verbal reports of their rating process were used by Orr (2002: 144, 149-151) to understand the way raters make their decisions. The raters were not only found to focus on different criterion relevant elements in the test-takers' performances but also to include features that were not part of the criteria, such as body language and comparing one test-taker to another. Therefore, Orr emphasises the need for careful examination of language test characteristics and test results, pointing out that correct criteria application is of vital importance in assessment. Orr (2002: 153) argues that rater training should instruct raters to focus on relevant criteria and on the level of sufficient performance. Furthermore, he maintains that raters should also learn what elements the assessment criteria do not include.

Using a questionnaire, Eckes (2008: 163) explored how raters with assessment experience interpret the importance of rating criteria of a writing test. Raters of the Test of German, which is a test used in higher education in Germany to assess the language proficiency of foreign applicants, were asked to give their views on the importance of the assessment criteria they used. With no reference to any particular performance, a total of 64 raters gave their perceptions of nine criteria, stating whether a criterion was "less important, important, very important or extremely important". The criteria included the following nine features: degree of fluency, text structure, argumentation, connecting ideas, syntax, variation and precision of vocabulary, completion of task description, degree of accuracy, and summarising given information.

Based on how the raters valued the different criteria, Eckes (2008: 163, 167, 171-173, 177) asserted his hypothesis that distinct rater types can be found in at least large-scale tests. Based on the findings of the study, the raters were grouped into six clusters, each of which represented a certain focus on specific criteria. Thus, for example, the raters of cluster A put weight on syntax and vocabulary, whereas the raters of cluster C perceived the criteria related to text structure as most important. Furthermore, there were clusters where the raters perceived certain criteria as less important than others, such as the criterion of fluency for the raters of cluster E.

In addition, Eckes (2008) argues for the importance of rater training, in line with the studies of Lumley (2002), Orr (2002) and Tarnanen (2002) reviewed above. Eckes (2008: 173, 179) suggests that rater training could be the tool to help balance raters' assessment performance and to promote test validity. There were indications in the study as to the failure of the training, based on the differences in the way the raters perceived the importance of single criteria. Moreover, it is proposed that raters' perceptions of rating criteria should be studied in more detail, using personalised methods such as interviews.

2.3.3 Rating scales and importance of criteria in competence-based qualifications

The assessment of foreign language proficiency was examined by Härmälä (2008: 18, 108, 111-112, 117) within the context of the Finnish competence-based qualifications. The qualification in question was the Qualification of Business and Administration, which represents the basic level of the three CBQ levels (see section 3.1.1). The key issues under examination were the competence test tasks and their implementation, as well as the assessment criteria. More specifically, the focus was on evaluating what a sufficient performance in a competence test of foreign language proficiency involves. Furthermore, investigating the uniformity and types of criteria that different training institutes apply was another objective. When using the expressions 'assessment criteria of performances', *suoritusten arviointikriteerit*, and ‘institute-specific criteria’, *oppilaitoskohtaiset kriteerit,* Härmälä refers to the same type of criteria that are referred to as ‘testing criteria’ in the present study (see chapter 1).

Before conducting interviews in her study, Härmälä (2008: 124-126, 219) first surveyed the views of fourteen teachers and assessors of English and Swedish by using a questionnaire, which contained issues such as importance of particular criteria, and arrangements of preparatory training and competence tests. The teachers were asked to place the given criteria in an order of preference according to importance, and to add a criterion if they so wished. The criteria consisted of seven elements picked from well-known rating scales: fluency, idiomaticity, pronunciation, accuracy, size of vocabulary, written or spoken content, and delivery of message. Some additions were also suggested by the teachers, such as courage to use a foreign language. The results were grouped in two ways (see Table 1).

Table no 1. Importance and general ranking of criteria (Härmälä 2008: 126)

Importance	General ranking
1. delivery of message	1. delivery of message
2. written or spoken content	2. written or spoken content
3. fluency	3. accuracy, pronunciation
4. pronunciation	4. fluency
5. size of vocabulary	5. size of vocabulary
6. accuracy	6. pronunciation, idiomaticity
7. idiomaticity	7. idiomaticity

Firstly, according to the perceived order of importance, the most important criterion was delivery of message, followed by written or spoken content, fluency, pronunciation, size of vocabulary, and accuracy. On this list, idiomaticity came last. Secondly, the criteria were grouped according to the most general rank given to each criterion by the teachers. With this method, no significant changes were detected, except that pronunciation became classified either as slightly more important or considerably less important. Yet when focusing on the elements of excellent language proficiency in particular, the assessors perceived accuracy as the most important criterion.

Härmälä (2008: 154, 214-217, 235) found that the institute-specific criteria devised by the training providers differed not only in content but also in level requirements. What the assessor interviews revealed was that some elements of the institute-specific scale contents came from the teachers, and some were based, for example, on parts of the Common European Framework of Reference for Languages or the Finnish National Foreign Language Certificate. In addition, there were inconsistencies regarding the scale structure, so that scales of 1-5 and pass-fail, for example, were used for the same qualification. According to the CBQ assessment procedures approved by the Finnish National Board of Education (see section 3.1.1), the standard procedure on the basic level, i.e. in vocational qualifications, is to use a five-point assessment scale. What was found in the assessor interviews was that the five-point scale was not perceived as easy to use by the teachers and assessors of English and Swedish. One reason for this was the complexity of determining the correct level of a candidate’s proficiency. Therefore, some teachers suggested moving to a pass-fail scale, which would, according to their perceptions, enable a more straightforward assessment by fulfilment of the given criteria.

As to other difficulties perceived by the teachers and assessors in the study by Härmälä (2008: 154-155, 245), one had to do with the criteria descriptions. For example, one assessor wanted to know exact expressions or precise grammar points that should be required at a certain level. Furthermore, many assessors reported that they wanted proper training on how to apply the criteria descriptions. They also called for opportunities to discuss the use of the criteria with colleagues. Härmälä concluded that uncertainty concerning the actual required level of language proficiency was a key factor in making the criteria of the training providers so different.

With the studies reviewed above, it is implied that the way the assessors of language proficiency perceive the use of rating scales and assessment is not uncomplicated. Since specific interest in the present study is on the assessment of language proficiency in the Finnish CBQs, some of the most important features of the CBQ system will be examined in chapter 3, with a focus on assessment in two competence-based qualifications.

3 FINNISH COMPETENCE-BASED QUALIFICATIONS

As described in chapter 1, a new training system offering vocational qualifications for adults was launched in Finland in 1994. There were several reasons (Yrjölä et al. 2001: 34-36) why 1994 was the year when the renewed Act on Vocational Adult Education, which contained the launch of the new Competence-based Qualifications system (referred to in this study as CBQs or the CBQ system), came into force. Firstly, since there was an apparent need for skilled workers, concerns had been voiced in the business world about the professional skills of people who had undergone traditional vocational training. The training system was regarded as inflexible, and there was thus a need for closer cooperation between training institutes and enterprises. Secondly, the education officials responsible for developing adult education had a desire to keep the vocational training for adults separate from the training aimed for young people. Thirdly, the experiences of other countries, such as England, Scotland and Germany, and experiments made in Finland offered encouraging examples of options for development. Fourthly, since adult participation in vocational training was growing, it was necessary to change the old training system, which was considered to be structurally heavy and expensive.

During an overall evaluation of how the Finnish CBQ system works, Yrjölä et al. (2001: 35) interviewed persons who had influenced the structure of the new vocational training system for adults, and the interviewees expressed views similar to the ones discussed above on the founding of the new system: the need for renewal was based on deficient vocational training of adult workers, on employers' criticism of vocational training, and on a fear of labour shortage.

The framework of the new system (Yrjölä et al. 2001: 37) was developed by the Finnish National Board of Education, and when the Act on Vocational Adult Education came into force on the first of May 1994, there were 143 qualifications included. Nowadays all fields of vocational training have their own vocational qualifications, approximately 360 CBQ titles (Finnish National Board of Education 2010). The number of candidates in pursuit of the three levels of CBQs has grown from approximately 2000 in 1995 (Vihervaara 2001: 33) to an annual total of 62 000 in 2007 (Competence-based qualifications 5/2008). The three levels of CBQs will be introduced later in section 3.1.1 of this study.

Compared with other European vocational training frameworks, the Finnish CBQ system bears closest resemblance to the British National Vocational Qualifications, i.e. NVQs (Työss' on sun mittas' – ammatillisia tutkintoja koskevista käsitteistä 1994: 26).

3.1 Cooperation network of authorities, enterprises and training providers

As a functional key element, the CBQ system features close cooperation of authorities, representatives of enterprises and training institutes working together in the development and implementation of the qualifications. In the following, these main contributors to the CBQ system (see Figure 2) and their duties will be introduced. Starting from the national level and moving on to the implementation, the first to be examined are the national authorities.

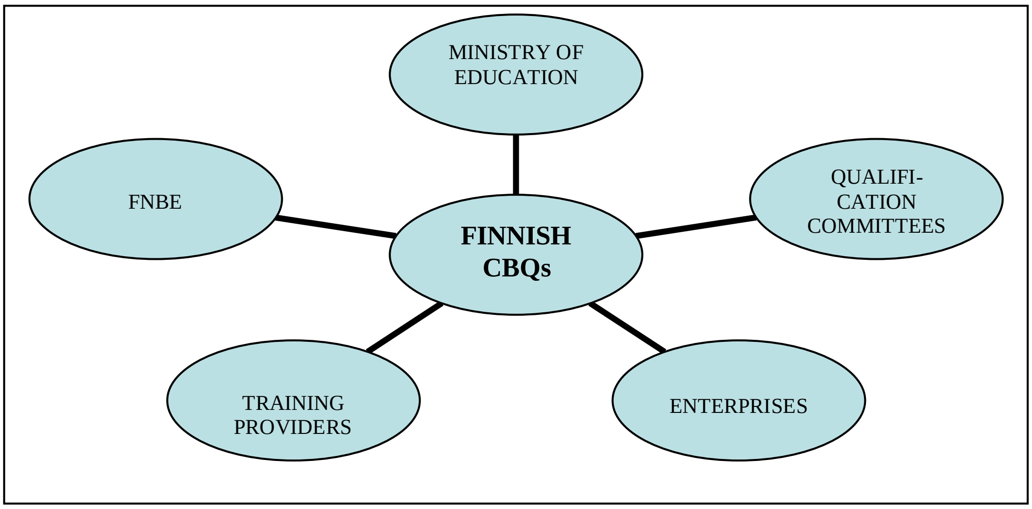


Figure 2. The CBQ system and its main contributors

Regarding the framework of the CBQ system (Competence-based qualifications 5/2008: 3-4), the Ministry of Education has the final decision making power concerning the choice of qualifications to be included in the qualifications framework, and the ministry is also responsible for any changes made to the framework in cooperation with representatives of Finnish enterprises.

Prior to final decision making (Näyttötutkinto-opas 2003: 16), training committees appointed by the Ministry of Education prepare statements for the ministry on qualifications and requirements. Details of qualification modules and competence requirements for each CBQ are established in the qualification requirements (Competence-based qualifications 5/2008: 3-4). The Ministry of Education is responsible for the approval of the qualification requirements, which also include assessment criteria for competence tests.

The second educational authority involved in developing CBQs, the Finnish National Board of Education, has several responsibilities (Näyttötutkinto-opas 2003: 16, 19). One of them is collecting proposals concerning qualifications to be included in the qualifications framework. Another responsibility is to draft qualification requirements together with representatives of enterprises and training institutes. This includes making decisions on the contents of the qualifications and sending the final proposal to the Ministry of Education. Yet another responsibility involves qualification committees: deciding which vocational branches have representation in the committees and nominating the committee members. A further area of responsibility of the Finnish National Board of Education is guiding and training members of qualification committees and training providers.

On the implementation level, enterprises and training institutes have their own forms of collaboration. Illustrated by the term tripartite cooperation, (Näyttötutkinto-opas 2003: 19), contributors representing employers, employees and training institutes are committed to developing the CBQs by working together as official experts in qualification committees, and as trainers and assessors. A key element in the tripartite cooperation is appreciation of the practical work done in workplaces. The Act on Vocational Education (Näyttötutkinto-opas 2003: 57-63) instructs training providers to pay particular attention to the needs of enterprises in planning and implementing CBQs. The tripartite cooperation described above does not, however, include any representation by national educational authorities (Yrjölä et al. 2001: 36). An explanation for this may be found in the shortcomings of the previous training system discussed earlier in section 3: there was an apparent effort to reduce the amount of administration in the new system and to bring enterprises right into the centre of it. In addition to participating in the implementation, all three parties are needed as advisers to the national authorities in issues concerning the development of the qualifications framework.

Representatives of enterprises participate in CBQ-related work in a number of ways (Näyttötutkinto-opas 2003: 16). Firstly, they are involved in drafting qualification requirements together with the National Board of Education. Secondly, together with training providers, they are responsible for the planning and implementation of CBQs, as well as for assessing candidate performance.

It is the responsibility of training providers, such as vocational adult education centres, to be in charge of the practical arrangements for competence tests (Näyttötutkinto-opas 2003: 16) , as well as to determine the contents of preparatory training (Finnish National Board of Education 2010). The main duties involved in arranging competence tests (Näyttötutkinto-opas 2003: 16) are appointing assessors and getting them acquainted with the assessment procedure, planning and implementing tests together with representatives of enterprises, and participating in candidate performance evaluation.

Qualification committees (Näyttötutkinto-opas 2003: 16) are bodies which monitor competence tests and their execution, give out qualification certificates and actively participate in developing the CBQ system and qualification requirements. There are typically three sectors of professionals represented in the committees: employers, employees and trainers. The Act on Vocational Education (Näyttötutkinto-opas 2003: 58) requires representation of self-employed professionals as well if self-employment is a typical feature of the profession in question.

The original purpose of having qualification committees bear the administrative duty in the implementation of the qualifications (Yrjölä et al. 2001: 36) was to ensure that test performances be evaluated in an unbiased way. Furthermore, this was a way to boost the employers’ and employees’ degree of commitment to the implementation of the tests.

3.1.1 Types of qualifications

A brief review of the qualification types is in place before examining the notion of competence and assessment practices of CBQs in the upcoming sections. Figure 3 illustrates the position of CBQs in the Finnish educational system.

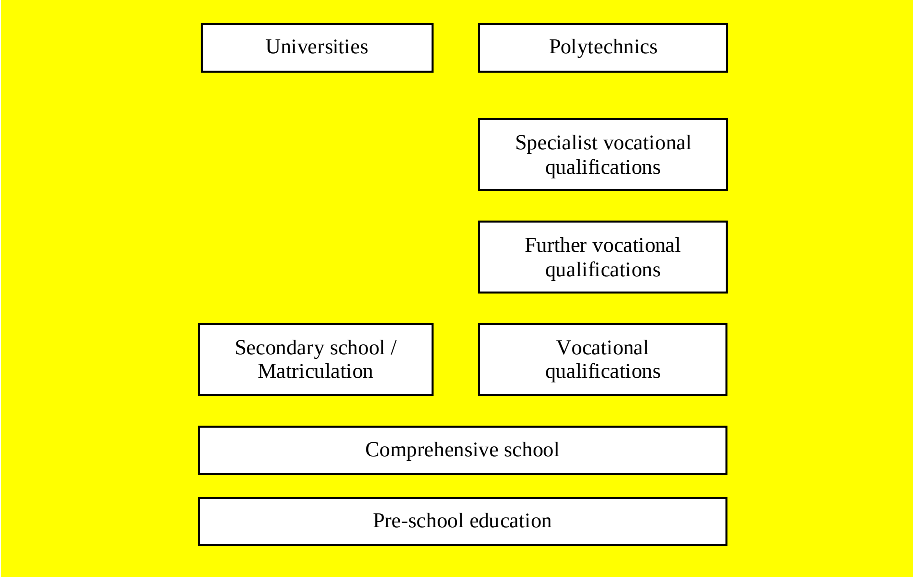

Universities
Polytechnics
Specialist vocational qualifications
Further vocational qualifications
Secondary school / Matriculation
Vocational qualifications
Comprehensive school
Pre-school education

Figure 3. CBQs and the Finnish educational system (Näyttötutkinto-opas 2003: 13, Vihervaara 2001: 30)

Within the Finnish vocational training system, the CBQs are arranged into three levels (Competence-based qualifications 5/2008: 4) depending on what kind of competence is expected of a candidate. As was pointed out above in section 3.1, the qualification requirements are separately determined for each CBQ by the Finnish National Board of Education. The level grouping meets the requirements set in the Act on Vocational Education (Yrjölä et al. 2001: 49). What follows next is a more detailed account of these levels according to the CBQ handbook published by the Finnish National Board of Education (Näyttötutkinto-opas 2003: 14).

Firstly, in qualifications of the basic level, vocational qualifications, a candidate should demonstrate mastery of such basic vocational competence which does not yet qualify him or her as a professional worker but which is required if the candidate is to pursue a further vocational qualification. There are no particular requirements concerning previous work experience at this point. The candidate’s performance is rated on a scale of 1 (satisfactory) to 5 (excellent). Completion of a vocational qualification entitles the candidate to apply to study at a polytechnic or a university.

Secondly, on the next level, i.e. in further vocational qualifications, a candidate is expected to demonstrate the vocational competence required of his or her particular field of work. This level of competence refers to the skills and knowledge a candidate has acquired either by working or studying on the basic level, with the addition of having completed further studies, and having work experience of about three years. On this level, as well as on the following one, the candidate's performance is rated as either a pass or a fail.

Thirdly, specialist vocational qualifications represent the most challenging level of CBQs, requiring the mastery of the most demanding tasks of a profession. The competence requirements on the specialist level refer to the skills that a candidate should possess after having first completed the basic level studies or after having otherwise acquired matching knowledge and, in addition, the candidate should have completed further studies and have five years of work experience.

It is further noted that despite the recommendations included in the level descriptions, the only decisive factor in gaining a CBQ certificate is a candidate's ability to provide evidence in competence tests that his or her competence meets the requirements of the qualification.

3.1.2 Notion of competence in CBQs

The concept of competence can be examined in various contexts. In section 2.1.1, competence was viewed in connection with communicative language testing. In order to gain an insight into the qualification requirements of CBQs, let us now look at some ways of how competence has been defined by parties involved in the development, implementation and research of CBQs.

When the CBQ system had just been launched in 1994, the Finnish National Board of Education, which had been the main developer of the CBQ framework, published a booklet wishing to clarify some key terms and expressions used in vocational qualifications. In this booklet (Työss' on sun mittas' - ammatillisia tutkintoja koskevista käsitteistä 1994), there are also comparisons to similar expressions used in some other European countries. Competence in a vocational context is described as professional skills acquired either through training or through work, and then demonstrated in a specific field and at a specific level. An individual's mastery of a specific profession or an area of expertise is seen as the basis for defining his or her competence.

An updated definition of competence by the Finnish National Board of Education can be found in the CBQ handbook (Näyttötutkinto-opas 2003). It says that the competence required in each profession is described by the labour market as part of the professional skills listed in the qualification requirements of each CBQ. The qualification requirements include detailed descriptions of the following elements (Näyttötutkinto-opas 2003: 54): qualification modules, required professional skills, targets and criteria of assessment, and ways of demonstrating competence. Anyone interested in checking the qualification requirements of a particular CBQ can do so on the internet homepage of the Finnish National Board of Education (see Opetushallitus – näyttötutkinnot 2010).

Regarding the notion of competence, three related studies are chosen to clarify the way competence is presented by CBQ researchers. The first two deal with the issue of competence by attempting to determine what it means, whereas the third one reviews various viewpoints and then reflects on them.

In a study on the progress of some CBQ pilot groups prior to the actual launch of the new training framework, Taalas (1995: 16), discusses the concept of competence as a general term as well as a vocational term. He argues that the Finnish concept of competence entails highlighting performance, competitiveness and comparison of individuals. Taalas maintains that general definitions of competence are made on the basis of duties or types of tasks, and through certain performance requirements, as in for example linguistic or vocational competence. He continues to claim that among the essential features of competence is individual achievement, which does not develop by itself or at random but as a result of learning, and should be related to a specific area of expertise.

Taalas (1995: 17, 22) then moves on to vocational competence. He argues that vocational competence is a combination of skills and knowledge that is inseparable from the learning process. The learning process is valuable, because it includes self-development, which is an important motivating factor for an individual improving his or her competence.

Furthermore, based on some model CBQ frameworks designed by the pilot groups during the study, Taalas (1995: 58) concludes that the parties involved in the implementation of CBQs consider it necessary to establish performance standards when defining competence. A candidate's competence is then demonstrated by a performance surpassing those standards. Degrees of expertise are considered equivalent to degrees of competence.

The relationship of competence and professional skills is examined by Turpeinen (1998) in a study on the assessment of CBQs. On the topic of competence, Turpeinen (1998: 12, 28) opens up the question whether professional skills always contain a type of general competence out of context or whether competence through its contents is always related to certain workplace requirements. She maintains that professional skills are not a steady level of competence, and continues to argue that competence essentially means a worker's ability to make use of and develop his or her skills and knowledge in various contexts. Regarding CBQs, Turpeinen (1998: 10) claims that the professional skills required in vocational qualifications and further vocational qualifications are defined as competence in such a way that they are valuable in themselves at workplaces. In addition, she also notes that tacit knowledge should be regarded as a part of competence and cannot be ignored in competence tests.

The third study was carried out by Haltia and Hämäläinen (1999), who interviewed assessors, trainers, CBQ candidates and members of qualification committees. They reached the conclusion that these groups had somewhat similar views and expectations on many issues regarding CBQs. They discovered, however, that the interviewees' views on the required level of competence varied somewhat in different CBQs. Identical professional practices were not easily found, since there was a great deal of variation both regionally and between workplaces. Defining the qualification requirements of a profession almost always entailed identifying and developing the vocational competence involved.

As Haltia and Hämäläinen remark (1999), the competence required in a CBQ is determined in its qualification requirements. Nevertheless, people involved in CBQ implementation may have a different understanding of competence in practice. What all the parties seem to agree on, however, when referring to competence in a vocational context is that it represents an individual’s skills and knowledge in a specific professional area, being tightly connected with learning and self-development.

3.1.3 Competence tests

Within the CBQ system, an adult candidate can gain a vocational qualification by passing competence tests. As mentioned in chapter 1, the idea of competence tests is to provide adults with the chance to demonstrate their professional skills irrespective of how the skills have been acquired and to become professionals who possess an official qualification certificate. One whole CBQ consists of several competence tests. According to Vihervaara (2001: 29), a competence test is a combination of methods used to determine the professional skills of a candidate in pursuit of a qualification. This combination consists of three parts: orientation phase, demonstration of professional skills, and assessment of professional skills according to qualification requirements. An orientation phase prior to testing is considered necessary in order to provide candidates with information on qualification requirements, assessment criteria, and practical arrangements.

As to the practical arrangements of competence tests, it is common practice that the implementation varies according to the wishes and facilities of workplaces and training institutes. Assessment of competence may thus take place either completely or only partly at workplaces, where concerns like issues of confidentiality have to be taken into consideration. The tripartite cooperation of employers, employees and training institutes discussed in section 3.1 is essential for finding the best arrangements and securing the quality of performance assessment.

Another issue concerning competence tests is terminology. Since the aim of the CBQ system is to provide work-based qualifications for people who are already employed, consistency in the choice of CBQ-related terminology further clarifies the position of the qualifications for everyone involved. In Finnish, the terminology has become fairly consistent by now, but when discussing the Finnish CBQs in English there is an inconsistency which will be examined next.

The above-mentioned inconsistency lies in translation differences. On its web site, the Finnish National Board of Education maintains a glossary of educational terminology in Finnish and in English. In the glossary (Opetushallinnon sanasto 2009), the current Finnish term *näyttö* has completely replaced the older term *näyttökoe*, which used to be the common term about a demonstration of competence in a CBQ, and which has been applied by various researchers during the existence of the CBQ system (see Taalas 1995, Turpeinen 1998, Haltia and Hämäläinen 1999, Vihervaara 2001). With the word *koe* removed, the terminology has become yet another element in distinguishing the CBQs from the old vocational training and qualification system that was considered inflexible and too distanced from the actual work.

The current recommended English translation for the Finnish term *näyttö*, when used in vocational adult education and training, is *competence test* (Opetushallinnon sanasto 2009). Similarly, publications in English regarding the CBQ system also talk about competence tests (see Competence-based qualifications 5/2008). However, for CBQs in vocational upper secondary education and training, *näyttö* has been translated as *vocational skills demonstration* (Opetushallinnon sanasto 2009). As was pointed out above, consistent terminology is an important element in clarifying the role of CBQs and distinguishing them as work-related qualifications. It would therefore be necessary to establish a common English translation for *näyttö* to be used in CBQs at both adult and upper secondary levels. Since such a term is not yet available, *competence test* is used in this study.

3.2 Two qualifications from the Tourism, Catering and Home Economics Sector

Two of the qualifications in the Tourism, Catering and Home Economics Sector have been chosen for this study: the Further Qualification for Hotel Receptionists and the Further Qualification in Tourism Activities. There are several reasons for this choice. Firstly, the adult education centre that is the training provider in this case study arranges competence tests in both of the two qualifications. Secondly, all the interviewees in this study are employees of this particular adult education centre. Thirdly, the interviewees and the present writer have all worked as assessors of English proficiency in the two above-mentioned CBQs. The foreign language proficiency requirements of these qualifications will be examined in section 3.2.1, and the guidelines of assessment as well as assessment practices will be discussed in sections 3.2.2 and 3.2.3. Before that, however, let us take a brief look at the structure of both CBQs,

All CBQs consist of more than one module (see section 3.1.3), some of which are mandatory and others optional. The whole qualification is completed when a candidate has finished all the required modules. The Further Qualification for Hotel Receptionists (*Hotellivirkailijan ammattitutkinto* 2001: 6-7) consists of one mandatory module, which is reception services, and of two optional ones, which a candidate can pick from a choice of six modules: conference services, lobby bar work, sales services, entrepreneurship, or a module from another CBQ. Proficiency in English is or can be included in all other modules except entrepreneurship. The Further Qualification (*Matkailun ohjelmapalvelujen ammattitutkinto* 2001: 6-7) consists of two mandatory modules and of two optional ones, which a candidate can pick from a range of nine modules. One of the optional modules is the use of foreign languages in travel-related customer service.

In the Further Qualification for Hotel Receptionists, the required foreign language skill level is embedded in the other professional requirements of each module. An example of this is the module of reception services (*Hotellivirkailijan ammattitutkinto* 2001: 12), where the target of assessment is customer service.

3.2.1 Foreign language proficiency requirements

The necessary skills in foreign languages have been expressed in different ways in CBQ requirements, of which the Further Qualifications in Tourism Activities and for Hotel Receptionists are good examples. On the practical level, however, the means of implementing the skills assessment are very similar (see section 3.2.3).

It is characteristic of the CBQ system that the qualification committees of each qualification establish their own requirements, and that there are no model performances available. In the Further Qualification for Hotel Receptionists, the requirements for foreign language proficiency are determined in a rather general manner. In those modules, in which customer service is mentioned, it is stated that in addition to Finnish or Swedish, the successful candidate should be able to serve customers in fluent English, as well as show fair command of at least one other foreign language. The required level for English is the intermediate level of the Finnish National Certificate of Language Proficiency (*Hotellivirkailijan ammattitutkinto* 2001: 12-13). It is pointed out by the Finnish National Board of Education (Finnish National Board of Education 2008) that the levels of the certificate correspond to those of the Common European Framework of Reference for Languages.

Contrary to the example given above, in the Further Qualifications in Tourism Activities, there is a separate module for foreign languages in customer service (see 3.2), with specific professional language skills required (*Matkailun ohjelmapalvelujen ammattitutkinto* 2001: 18). Firstly, the candidates should be able to use a foreign language to produce all the necessary background materials. Secondly, the candidates should be able to offer customer service in at least one foreign language, and customers should be served with regard to their cultural backgrounds. Furthermore, the candidates should know how to describe Finland and Finnish culture to customers, and to characterise their own operational environment. These requirements do not include any specification of a skills level, but are broken down into targets of assessment and assessment criteria (see section 3.2.3).

3.2.2 Assessment procedure

The assessment procedure of CBQs is based on what is stipulated in the Vocational Qualifications Act (Näyttötutkinto-opas 2003), which sets the general guidelines for CBQ assessment. Further procedural specifications follow in the qualification requirements of each CBQ. Detailed advice for organisers of competence tests are also available in various publications by the Finnish National Board of Education on the internet (see Opetushallitus - näyttötutkinnot 2010). These include certificate formats, advice on what needs to be registered and how, as well as hints on how to ensure the quality of the assessment procedure.

It is pointed out in the Act on Vocational Education (Näyttötutkinto-opas 2003: 56-63) that the emphasis of assessment should be on work performance, and that a candidate's skills should primarily be assessed in actual working situations. This includes a working environment which should be real or as real as possible. Furthermore, assessment requires systematic action in the collection of material, decision making and documentation, all of which should be in line with the professional skills and criteria defined in the qualification requirements.

In the handbook on the implementation of CBQs (Näyttötutkinto-opas 2003: 29), the Finnish National Board of Education specifies that different, primarily qualitative methods of assessment should be applied to get reliable results when assessing competence test performance. The recommended methods include observation, interviews, questionnaires, self-assessment, previous documented performances, and group assessment. It is further specified that by thus collecting all the relevant material, representatives of enterprises and the responsible training institute together make the final assessment on each performance. Moreover, the candidate will also prepare his/her own self-assessment, which may offer the assessors some useful information as well.

The qualification requirements of the Further Qualification for Hotel Receptionists and the Further Qualification in Tourism Activities both contain most of the stipulations described above. However, practical requirements have been added, and here are two examples of them. The first one concerns hotel receptionists, specifying that the areas of special attention concerning targets of assessment (*Hotellivirkailijan ammattitutkinto* 2001: 6) are core skills as well as mastery of working methods, tools, materials, processes, and background knowledge. The second one is for the qualification in tourism, dealing with the role of assessors (*Matkailun ohjelmapalvelujen ammattitutkinto* 2001: 19): assessors can participate in the service situations of a competence test as customers or as silent observers. Their role is to observe, interview customers, and have a conversation with the candidate, thus collecting assessment material.

The assessment scales are not the same on all levels of CBQs (Koulutusnetti 2010). On the basic level, i.e., in vocational qualifications, the scale is from 1 (satisfactory) to 5 (excellent). On the other levels, i.e., in further and specialist vocational qualifications, the scale is pass - fail. The testing criteria used in assessment were briefly introduced in chapter 1 and will be examined in more detail in the next section.

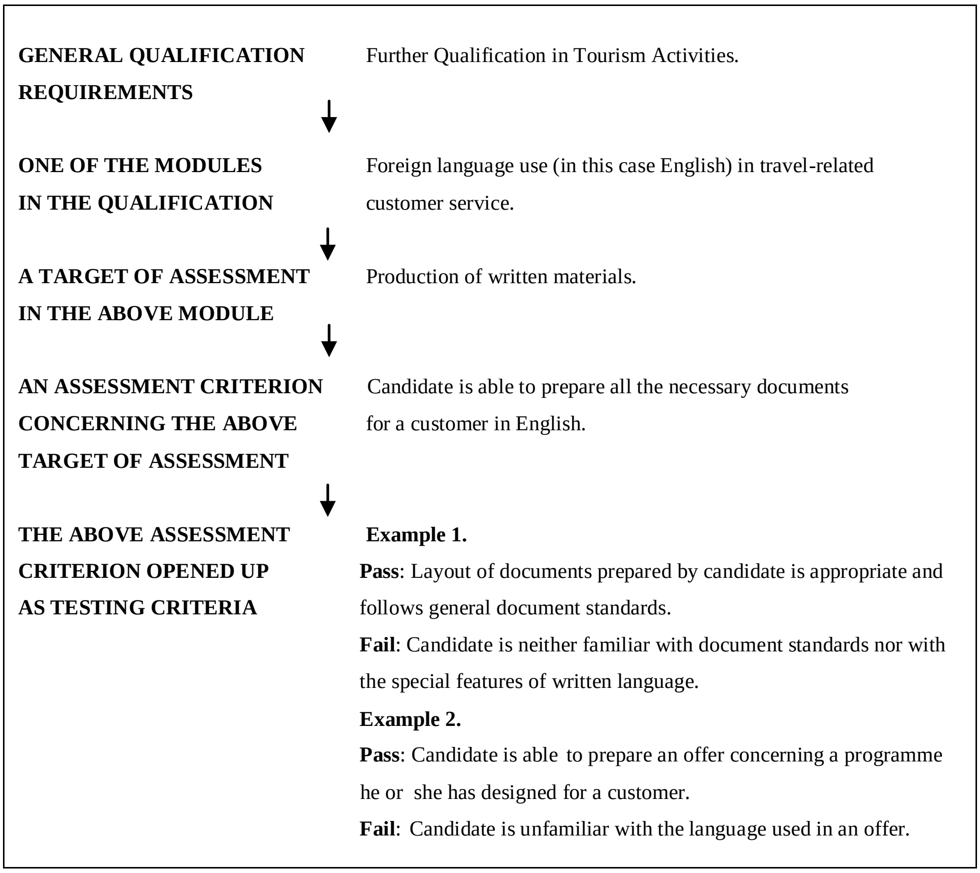
3.2.3 Practical application: the testing criteria

When awarding a CBQ candidate a qualification, assessors use the testing criteria as the key tool in a competence test. Before examining the function of the testing criteria, let us first determine their connection to some other elements that form the foundation of assessment in CBQs.

The foreign language proficiency required in a particular CBQ is described in the qualification requirements; more specifically, it is expressed as targets of assessment and assessment criteria (see *Hotellivirkailijan ammattitutkinto* 2001: 12-13 and *Matkailun ohjelmapalvelujen ammattitutkinto* 2001: 18). Targets of assessment refer to the core areas of competence in a qualification module (Näyttötutkintosanasto 2010: 2). Assessment criteria, based on targets of assessment, refer to definitions with which the levels of candidate performances can be assessed (*Hotellivirkailijan ammattitutkinto* 2001: 6). A common practice then is that each training provider organising a CBQ opens up the assessment criteria by transforming them into testing criteria for competence tests. When opened up, the assessment criteria become sets of sentences that describe acceptable and unacceptable performances, containing the practical essentials of the required competence. The testing criteria for a foreign language can either be a separate set or they may be embedded in the professional testing criteria designed by vocational trainers. As pointed out in chapter 1, assessment criteria provide the guidelines for assessment and are often rather imprecise. It is therefore the pass-fail descriptions, i.e., testing criteria, which are used by assessors when determining whether a candidate's performance is a pass or a fail. As there are many pass-fail descriptions in one competence test, getting one fail does not usually cause the candidate to fail the whole competence test. Regarding the actual tasks used in competence tests, their design is always the responsibility of each training provider.

The function of the testing criteria is best illustrated with some examples from the Further Qualification in Tourism Activities. In this particular qualification, there is a whole module on the use of foreign languages in travel-related customer service. The module contains three targets of assessment with six assessment criteria (*Matkailun ohjelmapalvelujen ammattitutkinto* 2001: 18-19). The first target of assessment is production of written materials related to tourism activities, and its assessment criteria require the candidate to be able to prepare documents such as offer, order confirmation, questionnaires concerning food or allergies, and customer feedback forms. The second target is the use of a foreign language in customer service, and its criteria require the candidate to be able to speak the foreign language fluently, to use specialised vocabulary if necessary, and not to be afraid of taking the initiative in conversations with customers. The third target is introducing Finland in customer service situations. Its criteria require the candidate to be able to introduce Finland and Finnish culture in a foreign language, to answer common questions presented by tourists, to take the customers' cultural background into consideration, to act in an unprejudiced way, and to give a positive impression of Finland and the services available.

Figure 4 clarifies the structural relationships between the elements that form the guidelines of assessment in CBQs. To pin it down to one of the qualifications in question, there are concrete examples concerning the first target of assessment from the Further Qualification in Tourism Activities, including its assessment criteria and testing criteria (*Matkailun ohjelmapalvelujen ammattitutkinto* 2001).



GENERAL QUALIFICATION REQUIREMENTS — Further Qualification in Tourism Activities.

↓

ONE OF THE MODULES IN THE QUALIFICATION — Foreign language use (in this case English) in travel-related customer service.

↓

A TARGET OF ASSESSMENT IN THE ABOVE MODULE — Production of written materials.

↓

AN ASSESSMENT CRITERION CONCERNING THE ABOVE TARGET OF ASSESSMENT — Candidate is able to prepare all the necessary documents for a customer in English.

↓

THE ABOVE ASSESSMENT CRITERION OPENED UP AS TESTING CRITERIA

Example 1.

Pass: Layout of documents prepared by candidate is appropriate and follows general document standards.

Fail: Candidate is neither familiar with document standards nor with the special features of written language.

Example 2.

Pass: Candidate is able to prepare an offer concerning a programme he or she has designed for a customer.

Fail: Candidate is unfamiliar with the language used in an offer.

Figure 4. Relationships between qualification requirements, targets of assessment, assessment criteria and testing criteria

As discussed earlier, one assessment criterion of the qualification in question requires a candidate to be able to prepare various documents for a customer (*Matkailun ohjelmapalvelujen ammattitutkinto* 2001: 18-19). This particular criterion has been opened up as practical pass-fail descriptions, i.e. testing criteria in examples one and two. The first pass-fail description specifies that when the layout of the documents prepared by the candidate is appropriate and follows general document standards, it is a pass. Failing means that the candidate is neither familiar with document standards nor with the special features of written language. The second pass-fail description specifies that a pass requires the candidate to be able to prepare an offer concerning the programme he or she has designed for the customer. A fail means that the candidate is unfamiliar with the language used in an offer. The assessor perceptions of using the testing criteria for the two qualifications in question will be analysed in chapter 5.

4 THE PRESENT STUDY

As described in chapter 3, the emphasis in candidate assessment in the Finnish CBQ framework is on real-life testing conditions and authenticity. The CBQ assessment model is designed to be enhanced by each training provider's own methods of implementing the assessment criteria. According to the CBQ assessment practices illustrated in sections 3.2.2 and 3.2.3, in the vocational qualifications of tourism and hospitality it means transforming the assessment criteria into practical descriptions, which serve as the testing criteria. In the following sections, the connection between the research questions and the use of the testing criteria will be indicated, and the research design will be discussed.

4.1 Research questions

The varying practices of using testing criteria in CBQs are a challenge to the trainers who work as assessors. As mentioned above, it should be of central importance that real-life testing conditions and authenticity be ensured in LSP testing (see Douglas 2001), such as the competence tests. In the present study, authenticity and real-life testing were examined in more detail in section 2.2.3. In my work as a trainer and qualified CBQ assessor, I have been involved in many informal discussions concerning criteria use, especially the criteria for English proficiency in the fields of tourism and hospitality. Colleagues not only from the institute where I work have raised questions about applying testing criteria in competence tests. It has been debated whether this type of criteria function as an assessment tool the way they should, whether some of the criteria descriptions should be reformulated, and whether there are any problems in the use of the criteria. The ultimate concern behind the questions has been how to use the criteria in varying testing conditions so that fair assessment of the candidates is ensured. Fairness of assessment has been recognised as an important issue in language testing by many researchers (see Messick 1994, Bachman and Palmer 1996, Davies 1997, Kunnan 1997, Elder et al. 2007).

There are no previous studies with a focus on how assessors of English proficiency perceive the usability of the testing criteria in competence tests. Various other assessor perceptions have, however, been the object of investigation, as can be seen in the studies reviewed in section 2.3. Therefore, the starting point for this study is to find out how the everyday implementers of this particular assessment tool describe its usability. As a topic, it provides an opportunity to explore a practical issue that assessors consider a target of development in their work. The research questions are as follows:

1. Which testing criteria do the assessors perceive as helpful when making their decisions and which not? Why?
2. What are the assessors' general perceptions on the long-term use of the testing criteria?
3. Is there something missing from the testing criteria or do they contain something unnecessary?
4. Are there any problems in the use of the criteria?

4.2 Research design

The behaviour, views and perceptions of professional educators have been investigated with both quantitative and qualitative methods. For example, in a study by Chambers and Richards (1993), the views of French teachers on pupil performances and marking criteria were identified through semi-structured interviews. In addition, the speaking tests which were the source of those views were marked on quantitative scales. Orr (2002), whose research was reviewed in section 2.3.2 of this study, used both quantitative and qualitative methods when examining raters' decision-making process. The scores from speaking tests were analysed quantitatively, but verbal reports from raters describing their thoughts during the rating process were dealt with qualitatively.

Similarly, quantitative as well as qualitative methods were applied in the following two studies. Cumming et al. (2004) investigated how university ESL instructors viewed some new TOEFL task types. The students' test scores were presented quantitatively. A questionnaire was used to collect the instructors' views on whether the student performances had been effective or not. Finally, in the interviews, the instructors were asked to give their views on how well the students' performances during the study reflected their typical classroom performances. In a study on the international school teachers’ perceptions of international-mindedness, Duckworth et al. (2005) remarked that using both quantitative and qualitative methods enabled the researchers to obtain a sufficient amount of macro- and micro-level information. They prepared a quantitative survey on personal information and professional backgrounds, whereas a qualitative survey was used for collecting beliefs concerning, for example, what an international teacher should be like.

Since the sample in the present study is very small, and the aim is to focus on the information content that teachers in the role of assessors can provide, a qualitative approach was considered most suitable for the current purposes. According to Hirsjärvi et al. (1997: 165), one of the features of qualitative research is that the data are typically gathered in natural situations. Another feature is that the researcher uses his or her own observations in order to gather data that will open up new approaches to the target of study. Moreover, discussions with carefully chosen participants are more valuable in measurement than other means. The whole research is seen as flexible, with the human being as its flexible instrument. In addition, the information gathered in those discussions serves to illustrate individual views of the phenomenon in question (Tuomi and Sarajärvi 2003).

Because every training provider is responsible for preparing their own sets of testing criteria for different competence tests, only assessors working for the same institute use the same criteria. Therefore, it was appropriate to do a case study. The format of a descriptive case study enables the gathering of detailed information of a small group of cases connected with each other (Hirsjärvi et al. 1997: 130). This case study connects assessor perceptions on the testing criteria in two further vocational qualifications. In addition to what was disclosed about the choice of these two qualifications for the present study in section 3.2, a few further remarks are in place as to why it was relevant to explore perceptions on both of the testing criteria. Firstly, the content of the criteria in the two CBQs is very similar, even though the pass-fail descriptions are expressed in different ways. Secondly, as all the assessors had been using both criteria, it was interesting to let them examine the two and see whether they would pick up similar points from them or compare them in some way. Moreover, the preceding informal conversations suggested that the information given by the assessors in this study might turn out to be useful to other assessors of English proficiency in developing testing criteria for competence tests.

As mentioned above, some assessors had already been involved in informal conversations concerning the English testing criteria, which is why it was feasible to offer them relatively free hands to continue expressing themselves. This could be done by interviewing. The reasons for this choice are discussed in more detail in section 4.2.2. The motivation for analysing the data using content analysis will be discussed in section 4.2.3.

4.2.1 Participants

As interviewees in this study there were five assessors of English who all work with adult students. The first reason for choosing these five particular persons was that they were all trainers in the same adult education institute as the present writer. Secondly, they all had more than five years of work experience as teachers of English in adult education and more than three years as assessors in competence tests, so they were very familiar with the Finnish CBQ system. Furthermore, they had all been working as teachers and assessors in the hotel and tourism sectors of CBQs, and had expressed their concerns about the development of the CBQ system and the testing criteria in several informal meetings. Some time before the interviews, the present writer had been working with two of the interviewees assessing students of the Further Qualification in Tourism Activities, and had cooperated with another interviewee in general CBQ development work. They were thus the first to be contacted. As the first interviewee was approached, she also named two other trainers as potential interviewees. The first contact with the interviewees was made either by e-mail or in personal communication at the adult education institute, and they all agreed to participate in the study.

As to collecting background information on the interviewees, using a questionnaire was first considered, as has been done by some researchers, for example, Tarnanen (2002) and Härmälä (2008). However, no written questionnaire was used this time for two reasons. Firstly, the present study involved only a small group of participants. In addition, the interviewees and the present writer were all colleagues, who were already partly familiar with each other’s backgrounds. The missing information was therefore gathered in person right after each interview, the details already known earlier were confirmed, and everything was recorded in a text file on the present writer’s computer. Since the interviewees were second language proficiency assessors like the ones in Tarnanen’s study in the example above, the personal data recorded involved the same topics but were modified to suit the needs of this study. The topics used by Tarnanen (2002) were: name, sex, age, mother tongue, education, experience with Finnish as a second or foreign language, and experience of assessing writing. The background data gathered for the present study were as follows: name, sex, age, mother tongue, education, experience as trainer, and experience as assessor of English proficiency in CBQs.

For the interviews, the participants were numbered as assessors 1-5. Three of the assessors were women and two were men. At the time of the interviews, their ages ranged from thirty-eight to fifty-two. There were two interviewees whose mother tongue was not Finnish: one was a native English-speaker and the other came from a country where English is an official language. Their interviews were held in English, while the others were held in Finnish. Through their work, all interviewees were familiar with both of the testing criteria in their original Finnish form, so that commenting on the criteria presented no problem. Three assessors had a Bachelor’s degree, while two had a Master’s degree. All other interviewees except one also held the national qualification for assessors of Finnish CBQs. In addition, two of the assessors were qualified interviewers for the Finnish National Foreign Language Certificate, and one was a trained interviewer and assessor for Cambridge Business English Certificate.

Before their participation in this study, the interviewees had been working for the same adult education institute as the present writer for varying periods, but in all cases for more than five years. Two assessors had also worked as language teachers in other countries than Finland. At the institute, all had been teaching English to students of various occupations, as well as to personnel of numerous enterprises. Their work experience with CBQ assessment varied from three to seven years. Furthermore, one of the assessors had been involved in defining the testing criteria for some CBQs, and two assessors had participated in other development work of CBQ English testing material.

4.2.2 Data collection: semi-structured interviews

To record assessor perceptions for the present study, personal semi-structured interviews were used. In what follows, other methods of recording perceptions and the motivation for choosing the interview will be discussed.

Stimulated recall is introduced by Gass (2001) in connection with second language learning as a form of verbalisation done either immediately or some time after a language event. The nonrecent type of recall is described as useful in cases where a longer period of, for example, learning strategies is commented on. Gass explains that the comments are collected with the support of audio, video or written material produced earlier by participants. On the usefulness of the method she remarks that even though stimulated recall does provide an additional angle to research, the time lapsed between the recall and the commented event may seriously influence the validity of the data. Using stimulated recall to gather perceptions in the current study was not considered relevant, since there is no direct connection to particular events. Furthermore, the assessors do not comment on any particular cases of assessment, even though their perceptions are naturally related to their experiences.

Other examples of data collected concerning raters' views can be found in the studies of Lumley (2002) and Orr (2002). They used simultaneous think-aloud data and verbal protocols, respectively, to collect raters' thoughts during the rating process. As the perceptions in the present study do not focus on the rating or assessment process itself but on using an assessment tool, i.e. the testing criteria, neither of the two methods mentioned above was considered suitable. Furthermore, in this case study there is no direct connection to a simultaneous rating or assessment procedure.

When exploring tertiary level students' perceptions of ESL speaking task difficulty, Elder et al. (2002) gave the students a questionnaire to be completed after each speaking task. In addition to task difficulty, there were questions concerning attitudes and familiarity with that particular task type. The actual test ratings were analysed quantitatively using a computer programme. Considering that the focus of the current study is on the information content of assessor perceptions, using questionnaires as a method of data collection was not applicable. As explained in section 4.2.1, information from a questionnaire used by Tarnanen (2002) served only as a source of determining which type of background information would be relevant.

For a small case study such as the present one, the semi-structured interview was considered most appropriate for data collection. As a method, Hirsjärvi and Hurme (2001: 34-38) maintain that the interview is very flexible and allows for reasons and explanations of opinions or answers. When comparing the interview with other methods, they point out that it can be more motivating for participants than a questionnaire and that along with the intended themes it can provide additional information, such as of connections between things. Furthermore, the interview provides a researcher with an open window to assessors' experiences and perceptions (Tarnanen 2002: 46). Of the different interview types (Hirsjärvi and Hurme 2001: 43-48, Tuomi and Sarajärvi 2003: 76-79), the semi-structured interview appeared to suit the current purposes best, since it was assumed that the interviewees were willing to elaborate on the interview topics, and thus the order and wording of the questions could be varied.

Adhering to the subject matter of the research questions, the interviews were constructed on four main themes: the criteria of the two CBQs in general, the usefulness of individual criteria, desired changes to the criteria, and any problems encountered. They were cut down into six concrete questions, in two of which the question was further divided into more detailed parts a) and b). The correlation between the research questions and the interview questions will be examined next. The full interview schedule can be found in appendix 1.

Interview questions 1 and 2 were connected with the first research question. Since the aim was to find out about whether this type of criteria were usable in assessment work, it was important to know which criteria really help the assessors when making their decisions, and which criteria were not considered helpful at all. Interview question 3 was connected with the second research question. This was a modification of what Tarnanen (2002: 43-44) asked the assessors' in the semi-structured interviews of her study. Tarnanen enquired the assessors' general views about the assessments they had completed, and in the present study the question was on the assessors' general perceptions of the criteria usability. Interview question 4 was connected with the third research question. It was based on the previous informal discussions among assessors, suggesting that some changes to the criteria content were hoped for. Therefore, in questions 4a) and 4b), it was important to discover what exactly the interviewees' suggestions would be. Interview question 5 was also divided into two parts. Question 5b) was connected with the fourth research question. Again, this was a modification from Tarnanen (2002: 44). Her question was on whether there had been any problems in the assessment process and how the problems had been solved. For the current study, the assessors were asked to describe any problems and their solutions concerning the criteria use. In order to make describing more concrete and easier for the assessors, a preceding question, i.e. 5a), was added to help the assessors to recall recent, actual assessment situations that they could refer to. The need to look to the future in question 6 arose from the informal conversations mentioned earlier. The last interview question was connected with the third research question in that expressing their views on any needs for development the assessors could offer some concrete ideas for renewal. It was also connected with the fourth research question, as remedies to some problems could be suggested.

The interviews were conducted during the last quarter of 2004 in the premises of the adult education institute, either in the interviewee's office or in one of the meeting rooms. There were no interruptions. Each assessor was interviewed individually, and no group interviews were used, unlike in some other studies. For example, Leung and Teasdale (1997) applied semi-structured interviews with panels of participants to investigating how teachers differentiated between student performances. Another example is Grant (2000), who also used semi-structured group interviews, as well as post-interview evaluations, in his study. He investigated New York state elementary and secondary level teachers' perceptions of what the changes in the state testing system mean to them. Forming a panel or group for the present study was not applicable considering its size and character as a case study, in which each assessor's perceptions were to be examined. The interviews lasted 45-60 minutes, were tape-recorded and later transcribed. The present writer was the interviewer in each case. Such interviewee statements that contained information subject to client or company confidentiality were excluded from the transcripts. Those occurrences in the transcription texts were marked accordingly, using the transcription protocol. During the interviews, the testing criteria of both CBQs in question were at hand to be commented on. Before the interviews, as the interviewees confirmed their participation, they were informed that the focus would be on their perceptions of using the two criteria. They were also asked to prepare themselves by recalling their latest experiences of using the criteria.

In order to find out whether the interview protocol was functional, a pilot study was carried out with a trainer who was not working with the CBQs in question. It helped the interviewer in being prepared to give the interviewees room to elaborate on topics of their choice, thus obtaining as much valuable information during the interviews as possible. It was also helpful in determining how much time would be needed for the interviews. However, timing was not always easy, although it was presumed that the participants might wish to elaborate. Since all interviewees were sincerely eager to discuss their use of the testing criteria, two interviewees started talking about their views and experiences even before the tape-recorder had been switched on. This was nonetheless in line with the assessors' assumed need to exchange views with colleagues. The topics that were prematurely started were returned to during the taped interview, following the interview schedule.

4.2.3 Data coding and analysis using content analysis

To discover how the assessors of English proficiency perceive using the testing criteria, it was necessary to conduct the analysis in such a way that connections of meaning behind the assessor perceptions could be found and understood. Content analysis was chosen in order to achieve this aim. In his discussion on the research designs of content analysis, Holsti (1969: 24) states that it is always the message of a communication which is content analysed, even though the results can be used in different ways. He then continues by presenting six questions to be used in the analysis. The last of these questions, "why", he establishes himself, after first listing five other questions by citing Lasswell, Lerner and Pool (1952: 12): "who says what, to whom, how, and with what effect?". Content analysis is also recommended by

Titscher et al. (2001: 66), when the centre of attention is on the communicative content of the data, as is the case in the present study.

Tuomi and Sarajärvi (2003) describe the role of content analysis in qualitative research not only as a framework for analysing data but also as three ways of carrying out the analysis: inductive, deductive or theory-based analysis. The method of analysis in this study was inductive, since the objective was not to build a theory or follow a previously presented one.

On the topic of data analysis, Lumley (2002: 254) reported using a complex coding system for categorising the think-aloud data, with categories which were partly similar to the rating scale used. The data in the present study emerged from the interview transcripts through the coding and clustering of units of meaning (Tuomi and Sarajärvi 2003: 112-115). The data were then categorised and examined on the basis of the research questions (Titscher et al. 2001: 58-59, 66). This was done by first forming sub-categories by clustering the assessor comments that were found to be similar, after which the main categories with themes were formed by clustering the sub-categories (Tuomi and Sarajärvi 2003: 111). The connections between the themes were examined in order to find answers to the research questions.

To present the categorisation in a clear and easily readable form, the layout of the categorisation in the present study is an adaptation from Williams et al. (2001), who used content analysis in their study examining EFL teachers' and students' perceptions on why students fail or succeed in learning English. In the description of their study, the sections of text include the main data categories presented in tables, along with the frequency of occurrence. In a similar way, the main categories of the present study with their thematic characterisations are presented in a table at the beginning of each section. The frequency listing in this study is used not to demonstrate the quantity of references as such but to give visibility to the themes that the assessors paid attention to in their choice of criteria. In the sections of chapter 5, reference to individual criteria is indicated as the themes are opened up and analysed.

Since assessors' views on criteria usability typically remain unverbalised in testing situations, bringing out these perceptions will therefore make the testing of English proficiency in CBQ competence tests more transparent, following an approach established by Tuomi and Sarajärvi (2003: 106). They describe two ways in which the world is experienced in studies using content analysis. On the one hand, there are studies in which the human being's position to reality and the targets of investigation is that of an outsider. On the other hand, some studies include the human being as a part of the world presented in them. In the latter approach, the main goal is not considered to be the truth as such but rather the comprehension of the invisible elements of reality through human experiences. Therefore, the analysis used in the present study follows the principles of the second approach. As users of the tool they work with, the assessors as well as the present writer are insiders to the world they comment on.

Nevertheless, with the small number of subjects in this study, there are concerns as to the subjectivity of data interpretation and reliability of the results (see section 6.2). The chosen method is, however, useful in gaining insights into the assessors' work by exploring a practical issue that is also a part of the present writer's everyday work environment.

Since the present study concentrates on the content rather than the form of the data drawn from the interviews, it will not be relevant to apply transcription symbols of high precision, such as pauses with their length or the tone of voice. In this respect, the data transcription conventions resemble those used by Tarnanen (2002: 45) when transcribing assessor interviews. She mentioned that all kinds of expressions from the interviews, including laughter or interviewer feedback, were taken along into the transcripts. The present study will, however, include symbols for overlapping speech and emphasis in order to provide additional details of the speakers' elicitation process. The symbols used in this study are listed in Table 2.

Table 2. The set of symbols used in transcribing the data (Adapted from Tarnanen 2002:45).

*** unintelligible word/words

ma- unfinished word

bold original emphasis by speaker

bold with italics key point highlighted by the present writer

[--] information under client confidentiality

[ ] overlapping speech

[…] a shortened example

(laughter) laughter or other noises by the interviewee or interviewer

For the transcripts and for the analysis, the interviewed assessors were identified by numbers from 1 to 5. The present writer was coded as the Interviewer. All through chapter 5, the Further Qualification for Hotel Receptionists will be referred to as the hotel qualification for short, and the Further Qualification in Tourism Activities will be called the tourism qualification. The testing criteria of the hotel qualification were coded as H1 – H7, and the criteria of the tourism qualification as T1 – T13. In order to allow for the assessors own voices to be heard properly, some of the interview samples included in chapter 5 are rather long. Therefore, the original Finnish interview extracts and their translations into English can be found in appendix 2.

5 FINDINGS

As explained in section 3.2.2, the testing criteria are the key tool that assessors use in competence tests. This chapter will contain the findings of this study concerning assessor perceptions on the usability of those criteria. The four research questions presented in section 4.1 will be the basis of the analysis: 1) Which testing criteria do the assessors perceive as helpful when making their decisions and which not? Why?, 2) What are the assessors' general perceptions on the long-term use of the testing criteria?, 3) Is there something missing from the testing criteria or do they contain something unnecessary?, and 4) Are there any problems in the use of the criteria?

The findings will be introduced applying the arrangement used by Williams et al. (2001), as indicated in section 4.2.3. Following the interview schedule of the present study, the first to be examined will be the perceptions on the useful and the unhelpful testing criteria.

5.1 Useful criteria

The interviewees were first asked to determine which criteria were most useful for them when awarding candidates a pass or a fail for their performances. Furthermore, by gathering the assessors' reasons for choosing particular criteria, it became possible to examine the connections between the choices in more detail than by analysing the criteria in an order of preference[1].

The themes which emerged to characterise the useful criteria (see Table 3) were as follows: linguistic features, cultural sensitivity, and clarity and measurability of the criteria. Each of these themes will be analysed in its own section (see 5.1.1 – 5.1.3), with indication to the particular testing criteria the assessors referred to.

[1] The English translations of all Finnish samples can be found in Appendix 2.

Table 3. Themes characterising the most useful criteria

Themes	Frequency
Linguistic features	10
Cultural sensitivity	6
Clarity and measurability	4

5.1.1 Linguistic features

Linguistic features was the theme which characterised most of the criteria perceived as useful by the assessors. The criteria which were believed to represent linguistic entities were emphasised by some assessors, whereas the use of professionally and structurally correct language was mentioned by others. These two sub-categories will be examined in more detail in the sections below.

Forming a linguistic entity

Regarding customer service situations as linguistic entities and not just checking the command of separate tasks were the issues which most assessors brought up when asked about useful criteria. The practical skills of a person working in tourism services were regarded worth emphasising. This was why the candidates' fluency and command of linguistic entities were argued for by assessor 1 in example (1) with reference to criterion T2 of the tourism qualification, which requires a candidate to carry out a full programme for tourists in English:

(1)

Assessor 1: elikä siinä on nyt ide- ideana se että hän ***pystyy huolehtimaan koko siitä tilanteesta elikä se on itse asiassa se kattaa koko sen idean*** siinä mielessä että et hänen pitää pystyä huolehtimaan jos ei hän tiedä sanoja hänen pitää myöskin öö pystyä kiertään asiat

The usefulness of assessing linguistic entities as a part of the assessment work was also acknowledged by assessors 2, 3 and 5. However, their views were based on another criterion, T1, which requires a candidate to design a full programme for a customer in English. As illustrated by example (2), assessor 3 noted that the key issue in all assessment was to focus on linguistic entities:

(2)

> Assessor 3: no tää että pystyy suunnittelemaan ohjelmakokonaisuuden englannin kielellä ni sehän ***viittaa suoraan siihen kielitaitoon jota me me ollaa niinku totuttu muutenki arvioimaan***
> Interviewer: mmh
> Assessor 3: siis se että arvioidaan ***kielellistä kokonaisuutta***

The recognition of the significance of linguistic entities in assessment work seemed to be in line and to promote the holistic view of assessment prevalent in general CBQ assessment guidelines. This view is commonly defined in the first chapter of the qualification requirements (see *Hotellivirkailijan ammattitutkinto* 2001 and *Matkailun ohjelmapalvelujen ammattitutkinto* 2001), where the targets of assessment are outlined as being related to the command of professional entities.

Use of professionally and structurally correct language

Strong support was found for the bond of correct grammatical structures and field-specific English. Appropriate language use, however, was not only thought of as remembering the right words and structures in the right place, but also as avoiding too complicated expressions. The effects of using difficult terminology and incorrect structures were brought up by assessor 1. Her argument was that inappropriate field-specific language would spoil the communicativeness of a marketing leaflet designed by a candidate, and might even result in a customer's decision not to use the services of that particular tour organiser. Referring to criterion T8 of the tourism qualification, which requires the candidates to compose a marketing leaflet in English, assessor 1 pointed out that official documents were often problematic for candidates, as can be seen in example (3):

(3)

> Assessor 1: ja ***ei virhei- kielivirheitä*** ku se on esite siinä ei sen pit- kieli pitää olla ymmärrettävä ei siinä saa olla olla näitä näitä tota liian vaikeita sanoja ja ennen muuta termejä et ***tämmösissä virallisissa asiakirjoissahan*** on juuri se että että jos ei sulle ihan selkeetä oo se suomenkaan asiakirjatermistö ni ***sä sit haksahdat***

In addition, assessing the candidates' command of business letters was seen as challenging by assessors 2 and 3, who made the point that candidates did not always regard learning the standards as important, if they had ready-made forms at their disposal at work.

Furthermore, the differences between spoken and written language were the reason why assessing a candidate’s personal way of using written language in a leaflet was considered useful by assessor 5, also relating to criterion T8. Example (4) shows how she defined her point:

(4)

Assessor 5: ja näissä pystyy kattoo sitte että että tota myöskin ton kohteliaisuusfraseologian ja ja sen että tota öö moni asiakaspalvelutyössä oleva ihminen hallitsee niinkun sanotaan semmosen niinku puhekielen aika kivasti
Interviewer: mm
Assessor 5: mut sit se että miten miten tota ***miten tää näytönantaja osaa sit muokata sen kirjalliseen muotoon*** niin se on kans aika aika olennainen

It seemed easy for the assessors to know how the above-mentioned criteria helped them in their work, but at the same time clearer guidelines on the criteria use were hoped for. The impression of dissatisfaction was strengthened by the assessors’ reasoning for their choices. Even though the usefulness of specific purpose language use was agreed on, there was dissatisfaction with not knowing how to instruct the students and how to define the threshold of a pass and a fail. For demonstrating field-specific writing skills, criterion H1 of the hotel qualification was observed by assessors 2 and 5 as being useful but problematic. This criterion requires the candidate to be able to design and use the necessary documents in English. For assessor 2 it sometimes meant asking the candidates to rewrite their work if they had tried to use the ready-made forms that many enterprises have. The difficulties of real-life testing are reflected on in example (5):

(5)

Assessor 2: there’s ***no clear guideline on what exactly you should tell the students*** […] if I compare it with yki
Interviewer: mm
Assessor 2: the the it’s a test it’s a testing situation so no dictionaries no materials but here they can use any materials they can use all dictionaries ***because it’s a working environment*** and when we actually write letters ***we use several resources***

The gap between the pass and the fail in H1 was described as disproportionate by assessor 5, as portrayed in example (6):

> (6)
>
> Assessor 5: siinä sit sanotaan et ei hallitse kielen perusrakenteita ja ammattisanastoa ni tässä onkin tavallaan niinkun annettu jo se alataso että minkä alle ei voi mennä [...] jotenkin se että että ***tää on aika karkee tää näyttöjen välinen ero osaa ja ei osaa***

Other themes with issues which generated discontent will be identified in the sections dealing with unhelpful criteria, features to be excluded from the criteria, and perceived problems (see sections 5.2, 5.5 and 5.6).

5.1.2 Cultural sensitivity

Issues related to culture surfaced on many occasions during the interviews, but on the question of useful criteria they clearly formed a theme of their own: cultural sensitivity. It was viewed as a kind of cultural awareness, but also as a form of culturally sensitive descriptive ability. These sub-categories will be opened up in the next two sections.

Active cultural awareness

On the connection between culture and useful criteria, two ways of being culturally aware were referred to by the assessors. For the purposes of the present study, these two views have been defined as active and passive cultural awareness. It became clear through the assessors' comments that passive cultural awareness, such as dressing up correctly for different occasions, attentiveness to different greeting habits, or sticking to the politeness norms of one's home country would not be enough in competence test situations. Instead, active cultural awareness was promoted by the interviewees, meaning that the candidates should readily combine their knowledge of customs and cultural differences with the English they use at work. Example (7) offers a concrete illustration of such a combination:

(7)

Assessor 1: et osaako tää henkilö öö erottaa erilaiset kielen ilmaisut öö sillon kun ei oo suomen kieli kyseessä elikä jos suomessa sanotaan et mee tonne ***nin englantilainen ei koskaan sano mee tonne***

Furthermore, knowing what one had to find out about before going into a service situation was the ground on which culturally polite customer service was built according to assessors 1 and 3. Two criteria were connected in their comments, i.e. H3 from the hotel qualification and T5 from the tourism qualification. Both criteria require a candidate to use polite English, respecting the customers' varying cultural backgrounds. Several examples of culturally active service were mentioned by assessors 1 and 3, such as active use of polite phrases, choice of field specific vocabulary, using or not using titles and names, as well as polite ways of starting and ending a conversation. The importance of understanding politeness norms was emphasised by assessor 3, as demonstrated by example (8):

(8)

Assessor 3: et se tulee aika nopeesti oikeestaan siit viestimisestä läpi että millä tavalla on niinku sisäistäny ***sen kohteliaisuusnormin joka kuuluu siihen kulttuuriin*** ja joka tulee esiin suoraan sen kielen kautta

In addition, politeness was claimed by assessor 5 to be the key to good service for international customers in English. Referring to criterion H2 of the hotel qualification, having a candidate advise customers on a hotel's services was seen as a good way to test it.

The assessors' reasoning for their choices caused some overlap with another sub-category, i.e. *Forming a linguistic entity* (see section 5.1.1), in that politeness, for example, was regarded as a linguistic as well as a cultural factor.

Descriptive ability

Cultural sensitivity was also perceived as being a type of descriptive ability, indicating the candidates' ability to use their oral English skills in describing their own country and its cultural features. Referring to criterion H4 of the hotel qualification, the argument of assessor 2 suggested that hotel receptionists had to be culturally sensitive of their customers' backgrounds and needs when discussing topics related to Finland. A candidate's ability to have a conversation with a customer about Finland and Finnish culture was considered a good way to measure oral proficiency. Assessor 2 had, however, sometimes had slightly complicated experiences of such conversations, when the customer's interests had not met those of the candidate, as example (9) illustrates:

> (9)
>
> Assessor 2: the only problem is that err if the questions that we're asking are not really familiar to them
> Interviewer: alright
> Assessor 2: like *** one student when we talked about mannerheim and she doesn't really care about mannerheim
> Interviewer: [(laughs) yes]
> Assessor 2: [so she doesn't] want to to talk about mannerheim but it's part of the näyttö that they know something ***but can we fail her if her language is good but she's not interested in mannerheim***

Although similar experiences concerning Finnish culture were not commented on by other interviewees, their other comments on cultural sensitivity reflected their awareness of the possibility of cultural complications in competence test situations.

5.1.3 Clarity and measurability

Use of the testing criteria can be challenging in unexpected ways, as indicated by the observations of assessor 2 in sections 5.1.1 and 5.1.2. Similarly, finding certain elements, i.e. clarity and measurability, in the testing criteria was also considered challenging but necessary when making decisions about passing or failing. Compared with other criteria, the topic of criterion T4, i.e. answering questions about Finland, was perceived as clear and easy to use by assessors 2 and 4. It was conceived as a clear task for any candidate and specific enough as a topic. All through the interview, assessor 4 in particular was consistent in his argumentation for the clarity and measurability of good testing criteria. Example (10) is an illustration of his views on measurability:

(10)

> Assessor 4: it's easier to measure things which are sort of ***specific behaviours*** than it is to sort of measure something which is kind of err more more to do with tacit knowledge for example or sort of vaguer kind of abilities

Two other criteria which were considered useful by assessor 4 for the same reasons were criterion T3, i.e. giving advice to customers on safety matters, and H5, i.e. giving information about the local area and its services. Assessor 4 even suggested how to make better use of this type of criteria, as indicated in example (11):

(11)

> Assessor 4: maybe these could be lumped together in one sort of err ***category*** perhaps ***of being able to answer questions*****.**

Interestingly, both of the assessors who brought up the issue of criteria measurability, i.e. assessors 2 and 4, were qualified interviewers for YKI, the Finnish National Foreign Language Certificate. At some point during their interviews, they both compared the CBQ competence tests with the YKI, expressing their dislike for what they described as vague CBQ testing criteria.

5.2 Unhelpful criteria

Regarding the criteria perceived as unhelpful (see Table 4), one the themes, i.e. difficulties with verification, included criteria that were also described as helpful by some assessors. The other theme which came up was the irrelevance of certain criteria as tools of assessment.

Table 4. Themes characterising the unhelpful criteria

Themes	**Frequency**
Irrelevance	8
Difficulties with verification	4

5.2.1 Irrelevance

Irrelevance of a criterion as an assessment tool was the reason why all assessors claimed two criteria of the hotel qualification, i.e. H6 and H7, to be difficult and unhelpful. The relevance of these criteria, which require the candidates to follow the rules of their working community and to show appreciation of their own work, was questioned because they were not considered to measure foreign language proficiency. The same line of argumentation for removing H6 and H7 from the set can be found in section 5.4.2.

Two particular cases when assessment was considered difficult were brought up by assessor 1. Firstly, reliable evaluation of a candidate's behaviour in the working community could not always be done. Secondly, the candidate could end up unfairly failing the whole competence test. The comments by assessor 1 in example (12) illustrate the assessors' shared view of irrelevance as assessment tool as well as assessment difficulties:

(12)

Assessor 1: mutta ***ei nää musta kuulu kielen osaamisalueisiin***
Interviewer: mm-h joo
Assessor 1: et se on sikäli paha jos matkailukielen käyttö hotellin asiakaspalvelussa hä- öö hylätään ***jos kaks oo o on hylätty*** okei jos siitä hylätään esimerkiks että minkälainen hän on työyhteisön jäsenenä et siihen ei oo saatu oikein kunnollista arvioijaa taikka sitten miten hän arvostaa omaa työtänsä ***ni hä- häneltä on matkailukielen näyttö hylätty***

Furthermore, referring to criterion H6, measuring the required content during the competence test was perceived difficult by assessor 4, as demonstrated by example (13):

(13)
Assessor 4: it's something you only know ***after seeing somebody in action in a workplace over over a very long time***
Interviewer: uhuh
Assessor 4: err so how do you know if somebody is ready to help other people I mean can you spot the you know in a t- ***in an examination can you spot if somebody's not ready to help*** somebody else

The general impression drawn from the interviews in this study was that the assessors wanted to participate in defining what was part of language proficiency in the CBQs and, in particular, what was not. Thus, naming criteria H6 and H7 as unhelpful and later choosing to exclude them from the criteria (see section 5.5) matched this impression.

5.2.2 Difficulties with verification

Some of the skills required in the testing criteria were considered difficult to verify, because the topics were seen as vague and therefore unhelpful. It was not considered easy by assessor 4 to know whether the candidates could actually do what was required by criteria H3 and T5 unless they were caught not doing it. Both criteria require the candidate to respect the customers' varying cultural backgrounds. The difficulty experienced by assessor 4 is shown in example (14):

> (14)
>
> Assessor 4: like this kulttuurien eroja you can only really it's something which you can only really s- see when it's not there if you like you know if somebody's obviously sort of racist or whatever or err insensitive to cultural needs perhaps in some way then ***it's noticeable by its absence***

In addition, uncertainty about how the required skills should be demonstrated also made criterion H4, i.e. describing Finland and its cultural features, difficult to verify for assessor 4.

Thus, the very same criteria which were viewed as useful parts of cultural sensitivity by assessors 1, 2 and 3 in section 5.1.2 were considered unhelpful by assessor 4 due to difficulties in verification resulting from vagueness. This kind of variation in what each assessor actually pays attention to in the criteria highlights the importance of the so-called assessment discussion. The discussion is held at the end of each competence test, to provide the assessors room for sharing their views and arriving at a joint decision. Another comment on the connection between the vagueness of the criteria and verification difficulties was made by assessor 5, who wanted to see more accuracy in the skill descriptions of the qualification requirements, particularly in the travel qualification.

5.3 Perceptions of long-term use

Since all interviewees had been working for more than three years as assessors, it was interesting to examine their long-term perceptions of using the testing criteria. It turned out that the general usability of the testing criteria was regarded as relatively good, but not without amendments, as will be shown in the following sections. Comments were made on the usefulness of the testing criteria, as well as on the pass-fail descriptions of performances, and on the assessors' own interpretations of the criteria (see Table 5).

Table 5. Perceptions of general usability

Themes	**Frequency**
Usefulness for assessing speaking and writing	2
Pass-fail descriptions	2
Personal interpretations of criteria	2

5.3.1 Usefulness for assessing speaking and writing

The testing criteria were regarded as suitable for assessing speaking and writing skills by assessors 1 and 2. However, on speaking skills they both agreed that the actual skill level needed in many international hotels nowadays was often closer to advanced rather than intermediate, which is the required level in the hotel qualification requirements designed by the qualification committee (see section 3.2). On the quality of the testing criteria they used in their work, assessors 1, 2 and 3 agreed that it was clearly better than those of some other training providers, since language trainers had been involved in defining the criteria. It is, however, common in some training institutes that no language trainers are involved, if vocational trainers design the criteria for the whole CBQ. This was strongly criticised by assessor 1, as can be seen in example (15):

(15)

Assessor 1: että se aukikirjottaminen öö siinä pitäs **ehdottomasti** olla kielenopettaja joka kirjottaa niitä auki [….] koska ***jos* hyvin *on laadittu kriteerit*** ni sillon se ***arvioija ei tarvii olla niin kauheen kauheen hyvä öö näyttökokeen arvioija*** koska et sä et sä välttämättä poimi niitä näyttökokeen arvioijia kielen arvioijia joka paikasta ei niitä **ole**

The role of the language trainer in CBQs will come up in more depth in connection with interview question 6 regarding future developments in section 5.7.2.

5.3.2 Pass-fail descriptions

Using criteria that include pass-fail descriptions (see section 3.2.3) with statements describing what a successful candidate can do was found convenient by assessors 4 and 5. However, assessor 4 pointed out that something needed to be done to the huge gaps between passing and failing. Regarding the gaps, there was too much room for subjective assessment according to assessor 5. As example (16) illustrates, the wish for more precision was expressed almost as an introduction to the upcoming theme of defining acceptable and unacceptable performances (see 5.4.1). :

> (16)
>
> Assessor 5: se että tota öö että ne kriteerit on niinku ***avattu tommosiks hyväksytty hylätty öö pareiks*** esimerkiks ni se on mun mielestä ***erittäin hyvä*** mä oon havainnu sen toim- niinkun hyväksi toimintaperiaatteeks
> Interviewer: mm
> A5: ***mut se tarkkuus sitä kaipaan***

5.3.3 Personal interpretations of criteria

The assessors' right to make personal interpretations of the testing criteria was both supported and criticised. Assessor 3 made interpretations of the criteria because she did not consider the criteria very precise at all, and because she had long work experience as trainer and assessor, as explained in example (17):

> (17)
>
> Assessor 3: no mitenköhän mä selittäsin tän ku must tuntuu et mä luen nää kriteerit useimmiten siinä vaiheessa ku mä suunnittelen sen valmen- valmistavan koulutuksen mut sit ***ku mä oon tehny tätä niin pitkään*** ni esmes ku ***mä teen yrityksille niit kokeita*** ni ni mä oon niinku tavallaan tehny sit vielä sen ***oman kriteeristöni***
> Interviewer: mm
> Assessor 3: joka niinku pohjaa tietysti tähän ja pohjaa sit siihen kokemukseen mikä mul on ollu […] ***mä en ehkä*** henkilökohtasesti itte ***oo seurannu näitä kriteerejä ihan just tän mukaan*** mitä täs on

In contrast, having to make interpretations because the testing criteria were not objective enough was difficult for assessor 5. Her concern was that two assessors using the same criteria for the same performance should arrive at the same result.

5.4 Additions to the criteria

The aim of the fourth research question was to find out if the assessors thought that something should be added the testing criteria or if there was anything they wanted to remove as unnecessary. The suggested additions to the criteria will be examined first.

All assessors had something to add to the criteria. The most wanted additions focused on the question of defining the borderline between acceptable and unacceptable performances (see Table 6). In the other desired additions, field-specific English was again united with cultural skills, in a similar way as in section 5.1 on useful criteria.

Table 6. Additions to the criteria

Themes	**Frequency**
Defining acceptable and unacceptable performances	6
Field-specific English	4

5.4.1 Defining acceptable and unacceptable performances

The threshold between acceptable and unacceptable performances aroused wordy comments from the assessors. More precision into the performance descriptions was unanimously called for, since the current descriptions were not considered accurate enough to help reduce the degree of rater subjectivity in measuring language proficiency.

Making sure that the testing criteria could be used in a meaningful way was a concern shared by assessors 1, 3 and 4. Establishing genuinely precise criteria in order to achieve more objective assessment decisions was suggested by assessor 1, as indicated by example (18):

(18)

Assessor 1: ne pitää olla selkeet **kenelle tahansa** kielenopettajalle pitäis olla selkeet kenelle tahansa ee muun aineen kri- näyttökokeen vastaanottajalle ne pitäis olla ihan selkeet
Interviewer: joo
Assessor 1: että ei se ei se riitä että tulee toimeen se ei oo mikään kriteeri tulla toimeen eli kuka päättää kun sitten että et mitä on se toimeen tulon raja ja ja just nimenomaan se että se on kova paikka ***jos sä hylkäät millä perusteella sä perustelet että että ei tullu toimeen***

A related point promoting precision in order to close the gap between the descriptions of a pass and a fail was made by assessor 3 (example 19):

(19)

Assessor 3 : jos aattelee näitä ääripäitä nin ne ne vaatimukset on aika erilaiset sit että
Interviewer: mm mm
Assessor 3: ni sellasta täsmällisyyttä toivos näihin kriteereihin ettei kenenkään niinku opettajan tai niinku arvioijan tarkotan ***tarviis jäädä epätietoseks*** siitä että tota että mitä mikä niinku missä kulkee se raja että ***kuka ja millä perusteella on hyväksytty ja kuka ei***

In addition, the wish to diminish the effect that the assessors' work experience had on their assessment decisions was why precision was argued for by assessor 4. Example (20) is an illustration of his views:

(20)

Assessor 4: ideally for err any assessor ***you need to know exactly where the line falls*** you know where something is unacceptable and where it's acceptable
Interviewer: mm
Assessor 4: and ok if you have very sort of skilful err experienced people who have a lot tacit knowledge and they've been doing it for a long time and so on and they kind of maybe intuitively know what is acceptable and unacceptable and that's fine if your s- your system is running purely on people like that ***but inevitably you're gonna have people who don't have very much experience*** and who are new to the system and ***how do they know***

Furthermore, inclusion of concrete examples of acceptable and unacceptable performances in the criteria was suggested by assessor 5 to help the assessors in their work. Yet another solution, i.e. nationwide standardisation of the criteria for English proficiency, was suggested by assessor 2 in order to use them for all occupations.

5.4.2 Field-specific English

Having more criteria connected with field-specific English skills was hoped for by assessors 1, 3, and 4, and even putting them in a separate section was suggested. A section divided into speaking and writing skills was argued for by assessor 4, with the aim of obtaining clear measurement of specified skills. A slightly different proposition for a section connecting cultural knowledge to field-specific English was offered by assessor 1. Her point was that cultural knowledge was too often understood only as familiarity with the differences between cultures, such as Japanese and American culture, instead of including the cultural features belonging to a language, as demonstrated in example (21):

(21)

Assessor 1: esimerkki tästä vois olla se et mä olin mä olin tuolla [--] vastaanottamassa näyttöö joss oli kaks kaks tota kaks meijän harjottelijaa […] mä kuuntelin sen respan joka siellä oli siis ***ammatissa toimiva ihminen öö eikä mikään harjottelija*** eikä mikään just sinne tullu ja kyse oli siis [--] jota se on nyt aika arvostettu kuitenki ihan keskellä helsinkiä
Interviewer: mmh
Assessor 1: ja tota sinne tuli tähän öö puhelu asiakas sieltä jostain huoneesta siel oli jotain hässäkkää ja ja ku hän sitten vastas siihen nin hän sano jotain tähän tyyliin että ***i come wait there*** siis tyyliin [tähän näin]
Interviewer: [(nauraa)] aha joo joo
Assessor 1: ***tää on se*** mitä mitä mä niille ***mitä mä siellä haen*** ja jo usein niinku sanonkin sitte että mietipä uudestaan mitä varten mitä varten englantilainen ei sano tai **yleensä** mene suoraan asiaan vaan että […] could you wait ni joku tulee sinne pian

In addition, using English in collecting information from various sources, such as the internet, was considered worth adding to the testing criteria in some form by assessor 3. She explained that gathering information was difficult if a candidate did not have good field-specific language skills.

5.5. Exclusions from the criteria

Having to assess the candidates' commitment to their working communities was clearly the most unwanted theme in the testing criteria. The same two criteria, i.e. H6 and H7 of the hotel qualification, which were described as unhelpful in section 5.2.1 as irrelevant assessment tools, surfaced under this theme. In addition, preparing the necessary background materials for customers was not considered to be a consistent section of the tourism criteria, and therefore removing parts of it or redesigning it was perceived as necessary. Assessor 3 was the only one who did not wish to remove anything, while the others were very critical. The features to be removed (see Table 7) will be described in the following sections.

Table 7. Exclusions from the criteria

Themes	**Frequency**
Commitment to the working community	5
Preparing background materials for customers	1

5.5.1 Commitment to the working community

Following the rules of one's working community and being its co-operative member were not regarded as a relevant part of a foreign language competence test. Removing criteria H6 and H7, which were the ones used to assess those skills as well as the candidates' appreciation of their own work, was therefore suggested by assessors 1, 2 and 5. The two criteria were not thought to measure language skills, nor was it even regarded possible to assess someone's commitment to his or her working community when testing language proficiency. The view expressed by assessor 2 in example (22) was shared by the others:

(22)

> Assessor 2: it doesn't really tell us it it does tell us something about communicative skills but ***it's not the use of foreign language***

However, the content of the above-mentioned criteria as such was not questioned by the assessors, only their appropriate use in CBQ competence tests.

5.5.2 Preparing background materials for customers

Designing documents, such as a marketing leaflet or an offer, for clients in English might seem an odd topic for removal, since it already appeared in section 5.1.1 as one of the helpful criteria. Nevertheless, the criteria referring to preparing a marketing leaflet and an offer, i.e. T8 and T9 of the tourism qualification, were regarded worth removing by assessor 4 because they did not clearly measure any specified skill. He found the whole section on preparing background materials too vague and in need of redesigning. Example (23) is an illustration of his view on vagueness:

> (23)
>
> Assessor 4: and again you have the problem of […] ***what does it mean not to know err the the language in a err an offer*** for example

During his interview, it proved too difficult for assessor 4 to decide which other criteria to take out, as he would have wanted to change the whole section to measure either writing skills or spoken ability. Even though the other assessors' reasons for their exclusions were quite different, it can still be concluded that their general aim was to remove features that could not be measured as part of language skills.

5.6 Problems

The problems encountered in using the two sets of English testing criteria culminated in expressions of uncertainty. The uncertainties brought up by the assessors involved the problematic relationship between the qualification system and the testing criteria, as well as particular problems in criteria implementation (see Table 8).

Table 8. Problems

Themes	**Frequency**
Uncertainties concerning the qualification system	3
Uncertainties in applying the criteria	3

According to the planned interview schedule (see Appendix 1), the assessors were asked not only about problems but also about the last time preceding the interviews when they had assessed candidates for either the hotel or the travel qualification. This was done to help the assessors recall actual assessment situations and to describe the problems more concretely, as well as to give examples of possible solutions used. However, the assessors perceived it difficult to link each problem with a specific assessment situation.

As mentioned in section 4.2.2, the interviews were conducted during the last quarter of 2004. Assessors 1 and 2 recalled having assessed many candidates for the hotel qualification during the year 2004, but did not name any particular date or time period. Assessors 3 and 4 had been working with qualifications of different professions, which was why it turned out difficult for them to recall a date for the last time they had assessed candidates for the hotel or travel qualifications. Assessor 5 recalled having assessed candidates for the travel qualification in summer 2004. Nevertheless, the assessors were asked to identify any problems in using the criteria despite the lack of exact temporal reference points, which unfortunately gave the present study a more general perspective into the perceived problems than was intended. It is also possible that the lack of reference to concrete assessment situations had an effect on the assessors' choice of problematic issues, so that bringing out general problems was perhaps easier.

5.6.1 Uncertainties concerning the qualification system

Operating within the Finnish CBQ system (see section 3.1) was perceived as problematic by assessors 2, 3 and 4. All the problems concerning the system were also ultimately connected with the testing criteria. Therefore, a brief summary of the main CBQ assessment principles is in place before examining the perceived problems.

Working as an assessor in CBQs means that one has to be familiar with the basic principles of the CBQ system, qualification requirements, and use of each testing criteria set. In short, the CBQ system is constructed on the premise that training providers open up the assessment criteria of each qualification by transforming them into testing criteria for competence tests carried out in workplaces (see section 3.2.2), where the real-life assessment takes place. This means that the assessment criteria of each CBQ are turned into practical descriptions of acceptable and unacceptable performances. As explained in the above-mentioned sections, the assessment criteria of qualification requirements are the ground on which competence tests are built, but the actual assessment is done using the testing criteria.

As to the perceived problems, the English proficiency requirements designed by the travel and hotel qualification committees were considered uneven by assessors 2, 3 and 4. Assessor 3 argued that not all targets of assessment for English seemed to represent the same skill level, which had a harmful effect on the assessment process and made the testing criteria use more difficult. In her experience, one example of this was writing business letters, which in real life required a higher level of skills than sought for in the qualification requirements.

Further criticism on the CBQ system was given for not establishing standards for foreign language assessment, which led to unwanted variation in the testing criteria according to assessor 2. In example (24), his frustration of having to deal with the mix of qualification requirements, assessment criteria, and testing criteria is described:

(24)

Assessor 2: but another problem is that they have a system and they don't really have err ***what we need is criteria of what is fluent and what is not fluent*** they they just say that they are able to err function or serve customers in fluent english
Interviewer: uhuh
Assessor 2: err as long as the customer probably understands it that's fine but what is the criteria there's no criteria so you have to it's your own common sense probably or judgment ***your personal judgment about what is err the appropriate competence*** for these people […] so who **gives** these levels err ***is the secretaries' work less demanding in terms of use of foreign language than receptionists' or travel agents'***

Moreover, assessor 4 went as far in his criticism as to express his doubts about the whole Finnish CBQ system. His point being the need for better recognition of the workers' skills in workplaces, he reflected on the way the Finnish society functions as illustrated by example (25):

(25)

Assessor 4: and and what I was saying earlier about problems again then maybe you know perhaps what I feel is that perhaps the ***the whole sort of concept of these err examinations with their testing criteria is perhaps a bit unnecessary***
Interviewer: uhuh
Assessor 4: do you really do they really need to have these qualifications in the first place
Interviewer: uhuh
Assessor 4: I think err I mean ***at a deeper level*** I think err this it has something to do with the sort of finnish culture and finnish society and the sort of ***importance of qualifications***
Interviewer: uhuh
Assessor 4: and err sort of ***lack of recognition for tacit knowledge and experience*** […] those things often go so unrecognised in finland and there's this kind of concern that err

> every e- there should be a qualification for everything and ***if you don't have a qualification then you shouldn't be able to do it***

Apart from the very critical view of assessor 4 in example (25), the above-mentioned uncertainties appeared to be connected with the discomfort felt towards the lack of support from the system for the assessors' work. The desire for support was already noticeable in earlier expressions of discontent during the interviews, for example, in the comments of assessors 2 and 5 regarding imprecise instructions on how to conduct competence tests and on the threshold of a pass and a fail (see section 5.1.1), as well as in the comments of assessors 1, 3, and 4 on defining acceptable and unacceptable performances (see section 5.4.1).

5.6.2 Uncertainties in applying the criteria

Contrary to the problems concerning the CBQ system, the problems in criteria implementation were more easily connected with actual assessment situations. The problem of uncertainty in giving a candidate comprehensible reasons for the decision to fail him or her was brought up by assessor 1. Example (26) is a description of a situation when assessor 1 found it difficult to explain her decision clearly to a candidate who had tried hard but just did not have the required skill level:

> (26)
>
> Assessor 1: mutta sillon kun tulee näitä joista ***selvästi näkee et ei ei niinku sinne päinkään oo taitotaso*** nin sillai jää miettimään että ***mitä mä tästä nyt sitten pistän hylätyks***
> Interviewer: mm mm
> Assessor 1: et se on kun se ***kuitenki täytyy aina sille henkilölle perustella***
> Interviewer: nii
> Assessor 1: että esimerkiks sen takia että jos hän haluaa parantaa hän haluaa joka tapaukses tehdä sen uudestaan nin ***mitä hänen pitää harjotella***

Uncertainty was also found in interpreting the testing criteria in an assessment situation. Recalling an assessment discussion from the summer before the interview, assessor 5 mentioned that she could not feel sure that all the assessors had the same understanding of the criteria, as shown by example (27):

(27)

Assessor 5: kun sitä käytiin sitä arviointikeskustelua siinä siin oli ***kaks muuta arvioijaa*** ja minä nin tota me käytiin sitä keskustelua siinä niin öö sen niinkun huomas että me oltiin niinku aikasemmin nää kriteerit jo luettu ja omaksuttu tietysti mutta se että ei meillä ollu niinkun mun mielestä kun me käytiin läpi sitä sitä suoritusta sitä näyttöö niin ***en mä voinu olla varma että ymmärretäänks me nää kriteerit samalla tavalla***

Another kind of problem originated in the preparatory training and the reservations some students had about the correlation of their jobs to the testing criteria. Assessors 1, 2 and 3 had faced this problem in recent student feedback and had reacted strongly to it, as demonstrated by the reaction of assessor 3 in example (28):

(28)

Assessor 3: siinä näyttötilanteen aikana ei tuu useimmitenkään niinku näiden kriteerien kanssa ongelmia mutta mutta sit joskus siel valmistavas koulutukses tulee […] siel oli tullu tämmöstä palautetta että ei he et ***että en*** **minä** ***joudu omassa työssäni mitään tämmösiä asiakirjoja laatimaan*** nin tota mä oisin sanonu tohon et kuule ei tää oo mikään tää ei ole kurssi joka valmentaa sinua sinun omaan tämänhetkiseen työhösi
Interviewer: mm
Assessor 3: vaan ***tässä on nou- noudatettava näitä mitä on tutkinnon perusteissa***

According to the assessors, none of the problems concerning criteria application were solved during the actual assessment situations. However, different ways to deal with the uncertainties brought up as problems in sections 5.6.1 and 5.6.2 were offered by the assessors when they were asked about needs for development (see section 5.7).

5.7 Future developments

When asked about any need for development regarding the criteria, the themes raised by the assessors dealt with the future role of language trainers in the assessment process, the need for training and standards, and the wish to adjust testing criteria to work content (see Table 9). These themes will be examined more closely in the next sections.

Table 9. Future developments

Themes	**Frequency**
Need for training and standards	5
Role of language trainers in assessment	2
Adjustments to criteria according to work content	1

5.7.1 Need for training and standards

The biggest concern regarding the future was the current lack of training and standards. Even though the assessors were aware of the training providers' responsibility to personalise the testing, it was found that the official bodies within the Finnish CBQ system did not offer enough training to teachers and assessors involved in the assessment process. More training and standards were called for by assessors 2, 4 and 5 in order to make the assessments easier and more reliable. Example (29) conveys the criticism on the training given to CBQ assessors and a concrete wish to get training in the interpretation of the criteria:

> (29)
>
> Assessor 2: well I would like them to also give training to teachers like ***real training on how to evaluate the use of foreign language at work***
> Interviewer: uhuh
> Assessor 2: like they do in yki where they where they well they **give** us ***training for this näyttömestari but it's all about organisation about the structure but but it's not about how to really evaluate or how to interpret the criteria*** now they they tell you it's up to you to interpret the criteria *** there's no **standard** way of ***you can do your own form*** that you would as long as you you have to interpret the criteria so as long as the board approves but the board doesn't give any exact rule […]

Another concrete suggestion of how to improve assessments by jointly assessing performances was offered by assessor 4, as demonstrated by example (30):

> (30)
>
> Assessor 4: you really need to have sort of ***regular training for the teachers and assessors*** where they are ***for example looking at videos and assessing together*** the performance of err candidates
> Interviewer: uhuh
> Assessor 4: and ***deciding together*** you know that's acceptable that's unacceptable that's a grade three or that's grade four or five

Agreeing with her colleagues on the need for training, assessor 5 also wanted the training to include the possibility to evaluate a performance together with other assessors, so that they would all be able to share their understanding of the criteria.

The request to add more precision to the testing criteria (see section 5.4.1) was extended to a wish for wider standardisation in connection with the need for more training by assessors 2, 4, and 5. In a comment arguing for standardisation in example (31), assessor 2 referred to levels from the basic level of vocational qualifications (see 3.1.1).

(31)

Assessor 2: whoever is the organiser *** the board of education probably that they ***should set err standard criteria for language skills***
Interviewer: uhuh
Assessor 2: because they ***could be used for all different vocations and professions***
Interviewer: uhuh
Assessor 2: so ***what is a two what is a one what is three*** is it the same as yki because it doesn't say so and we as language teachers just probably think that it's the same

In agreement with his colleague, example (32) shows assessor 4 speaking for standardising skill levels:

(32)

Assessor 4: and those rates it´s very important that you have ***standardised assessment*** I think that's something which I think err is often lacking
Interviewer: uhuh
Assessor 4: and err to to get that sort of standard level of assessment where you know ***everybody understands what they mean when they give a pass or a fail*** […]and […] if there were ***in-between criteria*** presented somehow ***on a scale of err different behaviours*** what is the sort of what is a fail what is a borderline pass

Furthermore, assessor 5 strongly suggested that standardisation should include cooperation by the different parties involved in CBQ development, as shown in example (33):

(33)

Assessor 5: ja mä toivosin semmosta ***ruohonjuuritason yhteistyötä*** että ne ihmiset jotka jotka ni et olis niinku semmonen jonkinlaista kunnon yhteistyötä oltiin me sitten minkä tahansa tason toimijoita ni että ***enemmän sellasta avoimuutta tähän***

5.7.2 Role of language trainers in assessment

In the interviews, assessors 1 and 3 voiced their concern that language trainers might soon not be able to be involved in foreign language proficiency assessment in CBQs. It was known at the time of the interviews that the hotel qualification, for example, would not include independent foreign language competence tests in the future. In spite of this, the involvement of language trainers in the assessment process was firmly encouraged by assessors 1 and 3. Assessor 3 was very upset that assessments of foreign language proficiency had already been done without any involvement by language trainers.

Having expressed her uncertainty concerning future assessment conventions, assessor 1 suggested cooperation between vocational trainers and language trainers, starting from drafting the qualification requirements, as pointed out in example (34):

(34)

Assessor 1: ainakin sen voi öö sen öö kielitaito öö ***taitokuvauksen siihen tutkinnon perusteisiin niin tehdä kielen arvioijan kanssa***
Interviewer: joo
Assessor 1: että arvioijilla on helpompaa ja sitte se ***että molemmat ymmärtää sen asian samalla lailla***
Interviewer: joo joo
Assessor 1: et se on must semmonen suur kehittämisen paikka

In addition to being involved in drafting qualification requirements, language trainers should be engaged in opening up the testing criteria as well as in the actual vocational assessment according to assessors 1 and 3. In example (35), one way of doing it in practice is introduced:

(35)

Assessor 3: ja sit näis näytöis on tota tosiaan se et siinä arvioinnissa **pitäis** ***olla mukana jossain ominaisuudessa myös kielen opettaja***
Interviewer: aivan joo
Assessor 3: et vaikka se tapahtus sitte tällä taval että ni niitä ***nauhotetaan ja sit joku vaikka ni kuuntelee*** sitte [ja antaa oman]
Interviewer: [mm-h mm-h]
Assessor 3: lausuntonsa siitä

It was further suggested by assessor 1 that vocational and language trainers could hold discussions on what knowing how to assess a candidate’s English proficiency means in practice.

5.7.3 Adjustments to criteria according to work content

Wishes for development regarding the relationship of the testing criteria to the work content of the candidates of the hotel and travel qualifications were offered by assessor 1. She suggested that the testing criteria should be modifiable so that tasks that do not belong to the job description of hotel receptionists or travel agents could be excluded. Example (36) is an illustration of her point:

> (36)
>
> Assessor 1: hiukan tarkemmin kannattas ottaa ehkä huomioon mitä ne joutuu siinä tekemään että öö ***jos öö kertakaikkisesti ei koskaan joudu töissään mitään tuollasia kirjeitä kirjottelemaan*** nin miks ne on sitten sisällytetty tohon tohon näytön kriteereihin et ku siel on eri ihmiset jotka jotka sen hoitaa

However, between the present moment and the time the data for the present study were gathered, there have been changes in the assessment procedures of CBQs as well as in the requirements of the tourism qualification. Foreign language competence tests are no longer held as independent parts of vocational skills testing. Furthermore, the new qualification requirements for the tourism qualification, which came into force in 2006, only require candidates to make themselves understood in the languages they need, but there is no mention of foreign language proficiency levels at all. The requirements of the hotel qualification have not changed.

When the assessors where asked about any needs for development regarding the criteria, it was hoped that they might have some concrete ideas of what to develop, and how to remedy some problems. It turned out that they were able to formulate concrete proposals regarding the problems. Firstly, the uncertainty concerning testing criteria application (see section 5.6.1) could be diminished by establishing common testing criteria or some type of nationwide standards on, for example, what kind of language skills are considered fluent. Secondly, regarding the problem between the candidates’ work content and the testing criteria (see section 5.6.2), it was suggested that the testing criteria could be adjusted according to relevant job descriptions.

Furthermore, other concrete proposals were made on the theme of future needs. It was suggested (see section 5.7.1) that training be given on how to assess foreign language use at work and to how interpret the criteria, and that trainers could participate in defining standard criteria by watching videos of performances and assessing them together.

6 CONCLUSION

The aim of the present study was to discover how assessors of English proficiency perceive the usability of the testing criteria used in Finnish competence-based qualifications for adults. The motivation for this study originated in the informal discussions I had with some colleagues on our experiences of assessing English proficiency in competence tests. This study offered me an opportunity to explore and record assessors' perceptions of using the testing criteria, which are a key assessment tool in CBQs. By using semi-structured interviews, the assessors were first asked about the criteria they perceived as useful or as unhelpful, after which they were enquired about their general perceptions on the long-term use of the criteria. Furthermore, the assessors were invited to point out if there was something missing from the criteria, or if the criteria contained something unnecessary. In addition, the assessors were requested to reveal perceived problems in the use of the criteria. In connection with the desired additions, removals, and problems, the interviewees were also asked whether they could identify any issues to be developed in the future. The findings are summarised and discussed in section 6.1 in relation to previous studies. Finally, the implementation of the current study as well as the connections between this study and suggested further research will be the topics of section 6.2.

6.1 Summary of the findings and discussion

The overall perceptions on the long-term use of the testing criteria conveyed that the assessors regarded the criteria as relatively convenient for assessing speaking and writing. Most of the testing criteria perceived as useful by the assessors can be characterised as focusing on linguistic features. Strong support was found for the bond of correct grammatical structures and field-specific English. While a previous study by Tarnanen (2002) revealed that structural accuracy was valued by raters as perhaps the most important criterion of good writing, it was discovered in the present study that assessors perceived avoiding excessively complicated expressions as an equally important part of fluent field-specific written communication. In addition, most of the assessors viewed criteria that enabled the assessment of linguistic entities, such as designing or carrying out a full programme for tourists, as useful.

The other criteria perceived as useful were identified by their connection with cultural sensitivity. The assessors felt that passive cultural awareness needed to be separated from active cultural awareness. Therefore, respectively, the assessors distinguished mere attentiveness to different greeting habits and sticking to the politeness norms of one's home country from combining one's knowledge of customs and cultural differences with the English used at work. Thus, for example, active use of polite English phrases while respecting the customers' varying cultural backgrounds, as well as choosing to use or drop titles and names, were considered culturally active behaviour by the interviewees. This finding agrees with previous research on the use of hospitality English (Blue and Harun 2003: 88-89) referred to in section 2.1.2, which also acknowledges that good service means being locally and cross-culturally well oriented, since many customer service situations can be considered at least partly culture-bound. The type of active cultural sensitivity described by the assessors in the current study is particularly important because hospitality English is not something permanent and unchanging (Blue and Harun 2003: 77) due to the large amount of variation that it involves.

The criteria that were deemed as unhelpful were regarded so either on the grounds of being irrelevant as assessment tools or because of difficulties in verifying them. All assessors agreed that following the rules of one's working community and showing appreciation of one's own work were not suitable criteria for assessing foreign language proficiency in a competence test. Having to assess the candidates' commitment to their working communities was clearly the most unwanted topic in the testing criteria. Thus, the same two criteria that all assessors had already categorised as unhelpful were the first to be removed from the set. It was not even regarded possible to assess someone's commitment to his or her working community while testing language proficiency. As to verification difficulties, they were mostly based on the perceived vagueness of a criterion, such as describing Finland and its cultural features.

The issue of imprecision or vagueness continued to surface in other findings of the present study. In particular, this was the case with desired additions and removals of criteria, problems, and future developments. Regarding desired additions, more precision into the descriptions of acceptable and unacceptable performances was unanimously called for. Although the current form of the pass-fail descriptions was supported, the assessors did not consider the descriptions accurate enough to help reduce the degree of rater subjectivity. Vagueness was also seen as a reason for removing some criteria, such as the one where a pass was described as a candidate's ability to prepare an offer concerning a programme he or she has designed for a customer and a fail as a candidate's unfamiliarity with the language used in an offer.

The findings on vagueness correspond with those of Tarnanen (2002: 195-227), who reported that most raters viewed the content of the given scale as too vague and sketchy. In her study, the raters wanted to add more concrete criteria to the scale and to define the performance level descriptions in more detail. In the present study, also designing qualification criteria for English to be used for all occupations was suggested. It was regarded as a remedy to the problem of vagueness before the definition of testing criteria. A similar suggestion was also made in a previous study by Härmälä (2008: 245-246) on foreign language assessment in CBQs. She proposed that at least some of the qualifications could share consistent level descriptions because the language proficiency requirements were vague and the institute-specific criteria devised by training providers differed not only in content but also in level requirements.

As to the problems perceived by the assessors in using the testing criteria, the CBQ system was strongly criticised for the lack of adequate support for the assessors' work. A finding substantiating this claim was also reported in the study by Härmälä (2008: 249), who discovered an alarming lack of cooperation between the main organisers of CBQs, i.e. authorities, representatives of enterprises and training institutes. Another problem brought up by the assessors in the present study was that the standards for foreign language assessment in CBQs were not well-established. The English proficiency requirements designed by the travel and hotel qualification committees were considered uneven, so that not all targets of assessment for English seemed to represent the same skill level. According to Härmälä (2008: 247-248), the qualification requirements usually only describe the required level as sufficient, without an explanation of, for example, whose understanding of sufficiency is referred to. Yet another problem raised by the assessors of the current study was related to uncertainties in applying the testing criteria. Some assessors doubted whether all the assessors in an assessment situation had the same understanding of the criteria. The previous studies of Lumley (2002), Eckes (2008) and Härmälä (2008) confirm this doubt. Lumley (2002: 266-267) discovered that even though the raters seemed to have a fairly good common understanding of the scale contents, the way the raters emphasised and applied the scale descriptors sometimes differed. Eckes (2008: 171-172) noted that not all raters valued the criteria in the same way; for example, one cluster of raters perceived the criteria connected with syntax and vocabulary as most important. Härmälä (2008: 154) found that the raters experienced using written criteria descriptions as difficult because they had no opportunities to discuss the use of the criteria with colleagues.

In reference to research questions three and four (see 4.2.2), which were associated with criteria additions and removals as well as perceived problems, the assessors were also asked to comment on any future needs regarding the use of the testing criteria.

The findings imply that there is not enough proper training for assessors of language proficiency in CBQs. The current training was criticised for concentrating too much on the structure of the tests and on how to organise them. More training in interpreting the testing criteria and in evaluating the use of foreign language at work was called for in order to make the assessments easier and more reliable. An inclusion of joint assessment of sample performances was also suggested. These findings are supported by those of Härmälä (2008: 154) who established that since many raters of CBQs viewed deciding on the right skills level difficult, they wanted proper training on how to apply the criteria descriptions. However, contrary to the findings of this study, the results of Härmälä (2008: 128) imply that more instruction was required on competence test arrangements. The importance of proper rater training was also argued for in other previous studies. Orr (2002: 153) maintained that proper rater training should instruct raters to focus on relevant criteria and on the level of sufficient performance. Eckes (2008: 179) suggested that rater training could be the tool to help balance raters' assessment performance and to promote test validity. In addition, a useful observation on rater training was introduced by McNamara (1996:122-124) in a previous study in section 2.1.2. He pointed out that not all variation originating from raters could be eliminated by rater training and suggested that such elimination might not even be necessary. Nevertheless, he continued by emphasising attentiveness to such variation and implementation of relevant compensatory measures.

Along with the need for training, the assessors in this study brought up the requirement for better standardisation. Based on the findings, it can be suggested that some degree of standardisation of the foreign language proficiency criteria in CBQs might be necessary in the future. This could then have a positive effect on the precision of testing criteria as well. The assessors were very critical about the current standards. They claimed that due to a lack of proper standardisation, the required skill levels in qualification requirements and within testing criteria were not precise enough, leading to too much unwanted variation in the assessments.

Together with training, improved precision would promote all participants' understanding of the assessment procedures and criteria application in CBQs. At present, new qualification requirements are being applied in the Further Qualification in Tourism Activities, only requiring the candidates to make themselves understood in the languages they need, with no mention of the required foreign language proficiency level at all. The requirements of the Further Qualification for Hotel Receptionists have not changed. In her study on foreign language assessment in CBQs, Härmälä (2008: 131) remarked that a higher degree of standardisation could increase the fairness of assessments. However, if stricter qualification requirements or standards are created, it should also be considered what Haltia and Hämäläinen (1999: 60, 73) have pointed out in their previous research (see section 3.1.2). They argued that definitions of vocational competence are needed up to the point when all parties involved have reached a shared understanding, but even if some improvement in precision were needed, there would be a danger that qualification requirements become too restricting and outdated. A further concern arises from my own experience as assessor. It should be carefully examined how far the language proficiency assessment process can be standardised without damaging each student's individual plan for completing a competence-based qualification. Drafting and implementing such a plan is an integral and compulsory part of a vocational qualification.

Yet another finding on the future needs of assessors indicated that the involvement of language teachers in foreign language proficiency assessment in CBQs should not be cut down. It was suggested by the interviewed assessors that there should be more cooperation between vocational trainers and language teachers starting from the drafting of qualification requirements, so that a shared understanding of the language proficiency requirements can be reached. It was also proposed that language teachers should be engaged in opening up the assessment criteria into testing criteria and participate in the actual assessment as well. Härmälä's (2008: 156, 163-164) observations were similar. She found that language proficiency assessment in CBQs had increasingly become the duty of other than language teachers. The language teachers interviewed in Härmälä's study were worried about the uniformity of performance tests if language teachers do not participate in the assessment process. However, Härmälä (2008: 249) claimed that specific purpose language skills could be assessed by other CBQ assessors as well if the aim were not to assess the level of a candidate's language proficiency but to assess his or her language skills in performing a specific task. At present, foreign language competence tests are no longer held as independent parts of vocational skills testing the way they used to be when the interviews for this study were conducted. What this means in practice is that the team responsible for assessing language proficiency in a competence test no longer necessarily includes a language teacher.

To sum up, the general usability of the testing criteria for English proficiency in the two CBQs in this study was perceived to be relatively good by the interviewed assessors, but not without certain amendments and some degree of standardisation in the future. Issues such as loose qualification criteria and varying level requirements in competence tests were already detected by earlier researchers of CBQs not long after the CBQ system had been launched (see e.g. Turpeinen 1998). Regarding language proficiency, Härmälä noted (2008: 127) that it represented only five to thirty per cent of the overall professional skills in the CBQs on the basic level. However, taking into account the different CBQ levels (see section 3.1.1), the findings of this study on the assessors' future concerns seem to indicate that language proficiency assessment should receive more attention as an element of CBQs at least in the hospitality sector.

6.2 Evaluation of the study and suggestions for further research

The objective of this study was to examine assessor perceptions on the testing criteria of English proficiency in two CBQs. The information content provided by the assessors was important in order to bring out the assessors' insights, which is why a qualitative approach with semi-structured interviews was used. To obtain the data, it was most realistic at the time to conduct a case-study at one adult education institute. Content analysis was applied to the analysis of the data, since the aim was to find and understand the connections of meaning behind the assessor perceptions.

Concerning previous research, no other studies have been conducted on how the usability of foreign language testing criteria in CBQs is perceived by assessors. As reviewed in chapter 2, related studies on the assessment of foreign language proficiency and assessor perceptions have been carried out, for example, by Härmälä (2008), Lumley (2002), McNamara (1996) and Tarnanen (2002). In one of the related studies, Eckes (2008: 173) suggests that raters' perceptions of rating criteria should be studied in more detail, using personalised methods such as interviews.

In reference to the suggestion by Eckes (2008), I noticed that interviewing was a good way to allow an interviewee to elaborate on his or her experiences to formulate their perceptions. The semi-structured interview method worked well because it was possible to present the interview questions in a varying order following the interviewee's narrative. The unintelligible periods on the tapes were not long so that they did not cause major problems for the data analysis. However, the interview questions may not have been focused well enough, since even though the assessors were told that the interviews were about using the testing criteria, they also wanted to talk about other experiences with CBQs, for instance preparatory training and its arrangements. This made me think that perhaps the interviews served as a channel for the assessors to voice their concerns regarding the CBQs. It is also possible that the lack of reference to concrete assessment situations had an effect on the assessors' choice of issues, so that, for example, general problems were brought up more easily than specific ones.

The reliability of the present study (Titscher et al. 2001: 65) would be considerably enhanced by subjecting the coding of the data to different coders. It must be taken into account that the findings of this study are only representative of the case study in question, since the data are formed of only five interviews. Furthermore, my own role as interviewer was perhaps too subjective, as being a qualified CBQ assessor made me look at the assessor perceptions from an insider’s point of view. However, familiarity with the CBQ framework may have been useful for a deeper understanding of the assessors’ comments (see Tarnanen 2002: 275).

Regarding further research, studies on forming a system of model performances for foreign language proficiency in CBQs would be necessary, since the spreading of best practices in foreign language assessment at least concerning the two CBQs in this study appear to be inadequate. There is a test data-base and support service called ALVAR (see www.alvar.fi) for technical CBQs but with many other qualifications the training providers decide how to open up the assessment criteria from the qualification requirements. While the present study focused on the second level of CBQs, recent research on the basic level of CBQs has already shown that the testing criteria of foreign language proficiency vary considerably across the country (see Härmälä 2008). In order to find out what the actual form of possible standardisation might be, a set of studies covering all three competence-based qualification levels might be helpful.

At the time this study was carried out, it was not possible to include assessment discussions from actual competence test situations as background material. Therefore, further researchers might find it useful to conduct a study involving those discussions in order to examine assessors’ use of the testing criteria for English proficiency in actual situations. This could be particularly timely now that there seems to be a tendency to assign the assessment of foreign language proficiency a subordinate status compared with other targets of assessment in competence tests.

From the standpoint of language test development, it would also be valuable to conduct a comparative study involving members of qualification committees and language teachers on the definition of English proficiency required in CBQs, because construct definition for language tests along with the issues of test reliability and fairness (see e.g. Bachman and Palmer 1996, Douglas 2000, Elder et al. 2007) should be crucial concerns for test developers as well as assessors. Such a study could include a questionnaire to find out how the members of a particular qualification committee define the required foreign language proficiency expressed, for example, as fluent customer service in English on the intermediate level. For comparison, the questionnaire could also be presented to language teachers with experience of CBQs. It would then be interesting to examine the potential differences between the definitions by the two groups and the reasons behind them. Interviewing the committee members and assessors would provide a deeper understanding of the research problem. A related study referred to in section 2.2.2 of the present study was conducted by Jacoby and McNamara (1999) from the perspectives of Australian and American professional settings. They compared the assessment criteria for English as a second language received from physicists in the United States with the criteria in use in the Australian Occupational English Test. One of the findings was (Jacoby and McNamara 1999: 235-236) that linguistic performance could not be separated from professional performance in indigenous assessments. It was further argued that there could be a clear difference between the criteria formulated by linguists and those formulated by subject professionals.

As for the training necessary for foreign language assessment in CBQs, research could also explore what the training should consist of from the language teachers' point of view. In particular, if English proficiency is not assessed by English teachers but by other CBQ assessors, a study investigating the teachers' views of the training content would be beneficial. A training model could then be developed and piloted by having English teachers and the other assessors assess some performances together and then examine the results from the point of view of inter-rater reliability (see e.g. Bachman 1990).

BIBLIOGRAPHY

Bachman, L. 1990. *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. 2000. Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing* 17 (1), 1-42.

Bachman, L. and A. Palmer 1996. *Language testing in practice. Designing and developing useful language tests.* Oxford: Oxford University Press.

Blue, G. and M. Harun 2003. Hospitality language as a professional skill. *English for Specific Purposes* 22 (1), 73-91.

Canale, M. 1983. On some dimensions of language proficiency. In J. Oller (ed.), *Issues in language testing research.* Rowley, MA: Newbury House, 333-342.

Canale, M. and M. Swain 1980: Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1(1), 1–47.

Chalhoub-Deville, M. 1997. Theoretical models, assessment frameworks and test construction. *Language Testing* 14 (1), 3-22.

Chambers, F. and B. Richards 1993. Oral assessment: the views of language teachers. *Language Learning Journal* 7 (1), 22-26.

Clapham, C. 2000. Assessment and testing. *Annual Review of Applied Linguistics* 20, 147-161.

Common European framework of reference for languages: learning, teaching, assessment 2010. Council of Europe [online]. (2 Aug 2010) http://www.coe.int/T/DG4/Linguistic/Source/Framework_EN.pdf.

Competence-based qualifications 1st January 2002. Helsinki: Finnish National Board of Education.

Competence-based qualifications 5/2008. Finnish National Board of Education [online]. (1 Aug 2010) http://www.oph.fi/download/47642_competence_based_qualifications_2008.pdf

Cooper, C. 1977. Holistic evaluation of writing. In C. Cooper and L. Odell (eds.), *Evaluating writing: Describing, measuring, judging*. Urbana, IL: National Council of Teachers of English, 3-31.

Cumming, A. 2001. ESL/EFL instructors' practices for writing assessment: specific purposes or general purposes? *Language Testing* 18 (2), 207-224.

Cumming, A., L. Grant, P. Mulcahy-Ernt, and D. Powers 2004. A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing* [online], 21 (2), 107-145. (12 Nov 2007)
http://ltj.sagepub.com/cgi/reprint/21/2/107

Cummins, J. 1979. Cognitive academic language proficiency, linguistic interdependence, the optimum age question, and some other matters. *Working Papers on Bilingualism* 19, 197 - 205.

Cummins, J. 1983. Language proficiency and academic achievement. In J. Oller Jr. (ed.), *Issues in language testing research.* Rowley, MA: Newbury House, 108-129.

Davies, A. 1997. Demands of being professional in language testing. *Language Testing* 14 (3), 328–339.

Davies, A. 2001. The logic of testing Languages for Specific Purposes. *Language Testing* 18 (2), 133–147.

Douglas, D. 2000 (2002). *Assessing languages for specific purposes.* Cambridge: Cambridge University Press.

Douglas , D. 2001. Language for Specific Purposes assessment criteria: where do they come from? *Language Testing* 18 (2), 171-185.

Duckworth, R., L. Walker Levy and J. Levy 2005. Present and future teachers of the world’s children: How internationally-minded are they? *Journal of Research in International Education* [online], 4 (3), 279-311. (12 Nov 2007)
http://jri.sagepub.com/cgi/reprint/4/3/279

Eckes, T. 2008. Rater types in writing performance assessments: a classification approach to rater variability. *Language Testing* 25 (2) 155–185.

Elder, C. 2001. Assessing the language proficiency of teachers: are there any border controls? *Language Testing* 18 (2), 149-170.

Elder, C., N. Iwashita and T. McNamara 2002. Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer? *Language Testing* 19 (4), 347–368.

Elder, C., G. Barkhuizen, U. Knoch and J. von Randow 2007. Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing* [online], 24 (1), 37-64. (12 Nov 2007)
http://ltj.sagepub.com/cgi/reprint/24/1/37

Finnish National Board of Education 2008: Frequently asked questions [online]. (4 Aug 2008)
http://www.oph.fi/english/pageLast.asp?path=447,55149,5373,71321

Finnish National Board of Education 2010: Vocational Education and Training System in Finland. (2 Aug 2010)
http://www.oph.fi/english/mobility/europass/finnish_education_system/vocational_education_and_training

Fulcher, G. 2000. The ‘communicative’ legacy in language testing. *System* 28 (4), 483-497.

Fulcher, G. 2003. *Testing Second Language Speaking.* London: Longman.

Gass, S. 2001. Innovations in second language research methods. *Annual Review of Applied Linguistics* 21, 221-232.

Grant, S. 2000. Teachers and tests: exploring teachers' perceptions of changes in the New York state testing program. *Education Policy Analysis Archives* [online], 8, 14. (16 Nov 2007)
http://epaa.asu.edu/epaa/v8n14.html

Haltia, P. and V. Hämäläinen 1999. *Näyttötutkinnoissa vaadittava pätevyys.* Työelämän tutkinnot 4/99. Helsinki: Opetushallitus.

Hirsjärvi, S., P. Remes and P. Sajavaara 1997. *Tutki ja kirjoita.* 3.-4.painos 1998. Helsinki: Kirjayhtymä.

Hirsjärvi, S. and H. Hurme 2001. *Tutkimushaastattelu. Teemahaastattelun teoria ja käytäntö.* Helsinki: Yliopistopaino.

Holsti, O. 1969. *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.

Hotellivirkailijan ammattitutkinto. Tutkinnon perusteet 2001. Helsinki: Finnish National Board of Education.

Huhta, A. 1993. Teorioita kielitaidosta - Onko niistä hyötyä testaukselle? In S. Takala (ed.), *Suullinen kielitaito ja sen arviointi.* [Proficiency in speaking and its assessment] Julkaisusarja B: Teoriaa ja käytäntöä 77. [Publication series B: Theory and Practice 77]. Jyväskylä: University of Jyväskylä, Institute for Educational Research, 77-142.

Härmälä, M. 2008. *Riittääkö* Ett ögonblick *näytöksi merkonomilta edellytetystä kielitaidosta? Kielitaidon arviointi aikuisten näyttötutkinnoissa.* Jyväskylä Studies in Humanities 101. Jyväskylä: University of Jyväskylä.

Jacoby, S. 1998. Science as performance: socializing scientific discourse through conference talk rehearsals. Unpublished doctoral dissertation. Los Angeles: University of California.

Jacoby, S. and T. McNamara 1999. Locating competence. *English for Specific Purposes* 18 (3), 213-241.

Jones, R. 1985. Second language performance testing: an overview. In P. Hauptman, R. LeBlanc and M. Wesche. (eds.), *Second language performance testing.* Ottawa: University of Ottawa Press, 15-24.

Koulutusnetti 2010: Ammatilliset näyttötutkinnot. (1 Aug 2010)
http://www.koulutusnetti.fi/?path=ammatilliset_nytttutkinnot

Kunnan, A. 1997. Connecting fairness with validation in language assessment. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (eds.), 1997. *Current developments and alternatives in language assessment. Proceedings of LTRC 96*. Jyväskylä: University of Jyväskylä, 85-106.

Leung, C. and A. Teasdale 1997. What do teachers mean by speaking and listening? A contextualised study of assessment in multilingual classrooms in the English national curriculum. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (eds.), *Current developments and alternatives in language assessment. Proceedings of LTRC 96*. Jyväskylä: University of Jyväskylä, 291-323.

Lewkowicz, J. 1997. Authentic for whom? Does authenticity really matter? In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (eds.), *Current developments and alternatives in language assessment. Proceedings of LTRC 96*. Jyväskylä: University of Jyväskylä, 165-184.

Lewkowicz, J. 2000. Authenticity in language testing: some outstanding questions. *Language Testing* [online], 17 (1), 43-64. (2 Jan 2009)
http://ltj.sagepub.com/cgi/reprint/17/1/43

Lumley, T. 1998. Perceptions of Language-trained Raters and Occupational Experts in a Test of Occupational English Language Proficiency. *English for Specific Purposes*, 17 (4), 347–367.

Lumley, T. 2002. Assessment criteria in a large scale writing test: what do they really mean to the raters? *Language Testing* 19 (3), 246-276.

Matkailun ohjelmapalvelujen ammattitutkinto. Tutkinnon perusteet 2001. Helsinki: Finnish National Board of Education.

McNamara, T. 1996: *Measuring second language performance*. London: Longman.

McNamara, T. 2000 (2008): *Language Testing*. Oxford: Oxford University Press.

Messick, S. 1994. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 23 (2), 13-23.

Messick, S. 1996. Validity and washback in language testing. *Language Testing* 13 (3), 241-256.

Näyttötutkinto-opas. Käsikirja tutkintojen järjestäjille ja tutkintotoimikunnille 2003. Helsinki: Finnish National Board of Education.

Näyttötutkintosanasto 2009. Koulutuskeskus Salpaus [online]. (2 Aug 2010) http://www.salpaus.fi/material/nayttotutkintosanasto.pdf

Opetushallinnon sanasto 2009. Finnish National Board of Education [online]. (16 Sept 2009)
http://db3.oph.fi/sanasto/listaakaikki_s.asp

Opetushallitus – näyttötutkinnot 2010. Finnish National Board of Education [online]. (1 Aug 2010)
http://www.oph.fi/nayttotutkinnot/

Orr, M. 2002. The FCE speaking test: using rater reports to help interpret test scores. *System* 30 (2), 143-154.

Slater, S. 1980. Introduction to performance testing. In J. Spirer (ed.), *Performance testing: issues facing vocational education.* Columbus OH: National Center for Research of Vocational Education, 3-17.

Spolsky, B. 1985. The limits of authenticity in language testing. *Language Testing* 2 (1), 31–40.

Swales, J. 2000. Languages for specific purposes. *Annual Review of Applied Linguistics* 20, 59-76.

Taalas, M. 1995. *Ammattitutkinto ammattitaidon näyttönä. Ammatillisten aikuistutkintojen kehittäminen.* Kasvatustieteiden tutkimuslaitoksen julkaisusarja A. Tutkimuksia 62. Jyväskylä: University of Jyväskylä.

Tarnanen, M. 2002. *Arvioija valokeilassa: suomi toisena kielenä –kirjoittamisen arviointia.* Jyväskylä: Jyväskylän yliopistopaino.

Titscher, S., M. Meyer, R. Wodak and E. Vetter 2000. *Methods of text and discourse analysis.* London: Sage.

Tuomi, J. and A. Sarajärvi 2003. *Laadullinen tutkimus ja sisällönanalyysi.* 1.-2. painos. Helsinki: Tammi.

Turpeinen, R. 1998. *Ammattitaito ja sen arviointi näyttökokeissa.* Työelämän tutkinnot 4/1998. Helsinki: Opetushallitus.

Työss' on sun mittas' – ammatillisia tutkintoja koskevista käsitteistä 1994. Kehittyvä ammatillinen koulutus 10/94. Helsinki: Finnish National Board of Education.

Vihervaara, M. 2001. *Näyttötutkinnot, työvoiman rekrytointi ja koulutus. Näyttötutkintojen merkitys työelämän edustajien näkökulmasta*. Turun yliopiston kasvatustieteiden tiedekunnan julkaisuja A: 194. Turku: Turun yliopiston kasvatustieteiden laitos.

Weigle S.C. 2007. Teaching writing teachers about assessment. *Journal of Second Language Writing* [online], 16 (3), 194–209. (2 Jan 2009) http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6W5F-4PXNHV5-1&_user=949111&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_acct=C000049116&_version=1&_urlVersion=0&_userid=949111&md5=d0f86380af2c2a27b3aa32570ae7bd0c

Weir, C. 1988. *Communicative language testing with special reference to English as a foreign language*. Exeter: University of Exeter.

Weir, C. 1993. *Understanding and developing language tests*. Hemel Hempstead: Prentice Hall.

Williams, M., R. Burden and S. Al-Baharna 2001. Making sense of success and failure: The role of the individual in motivation theory. In Z. Dörnyei and R. Schmidt (eds.), *Motivation and second language acquisition (Technical report #23)*. Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center, 171-184.

Yrjölä, P., J. Ansaharju, P. Haltia, R. Jaakkola, A. Järvinen, T. Lamminranta and M. Taalas 2001. *Näyttötutkintojärjestelmän kokonaisarviointi*. Arviointi 12/2000. Helsinki: Finnish National Board of Education.

Appendix 1. Interview schedule

KYSYMYKSET SUOMEKSI

1) Käy ensin kriteerit läpi. Merkitse sitten kriteereistä yksi tai useampi, joiden olet havainnut parhaiten auttavan sinua päättämään, onko näytönantajan suoritus hyväksytty vai hylätty. Kerro tarkemmin, miksi!

2) Katso kriteerit uudestaan ja merkitse punaisella ne joiden olet havainnut olevan hyödyttömiä suorituspäätöksen teossa. Kerro tarkemmin, miksi!

3) Perustuen työkokemukseesi arvioijana, mikä on yleinen näkemyksesi kriteerien toimivuudesta?

4 a) Pitäisikö kriteereihin lisätä jotain? Jos pitäisi, niin mitä ja miksi?
b) Pitäisikö kriteereistä poistaa jotain? Jos pitäisi, niin mitä ja miksi?

5 a) Milloin ja keitä olet viimeksi arvioinut?
b) Kun ajattelet tuota kertaa, oliko sinulla ongelmia kriteerien käytössä? Millaisia ongelmia? Miten ratkaisit ongelmat?

6) Onko kriteereissä näkemyksesi mukaan jotakin kehittämisen tarvetta?

THE QUESTIONS IN ENGLISH

1) First, look through the criteria. Then mark one or more of the criteria that you have perceived to be most useful in deciding whether a candidate's performance is a pass or a fail. Give reasons why!

2) Look at the criteria again and use the red marker pen to mark any criteria you have perceived as unhelpful in making decisions about passing or failing. Give reasons why!

3) Based on your experience as assessor, what is your general perception of the usability of the testing criteria?

4 a) Should something be added to the criteria? If yes, what and why?
b) Should something be removed from the criteria? If yes, what and why?

5 a) When was the last time you worked as assessor and who did you assess?
b) Did you have any problems using the criteria? If yes, what kind of problems? How did you solve the problems?

6) According to your perceptions, is there any need for development regarding the criteria?

Appendix 2. The original Finnish interview samples and their rough translations into English

(1)

Assessor 1: elikä siinä on nyt ide- ideana se että hän ***pystyy huolehtimaan koko siitä tilanteesta elikä se on itse asiassa se kattaa koko sen idean*** siinä mielessä että et hänen pitää pystyä huolehtimaan jos ei hän tiedä sanoja hänen pitää myöskin öö pystyä kiertään asiat

Assessor 1: so there's the idea that they can ***take care of the whole situation meaning that it actually covers the whole idea*** in the sense that they must be able to take care if they don't know the words they must also be able to go around things

(2)

Assessor 3: no tää että pystyy suunnittelemaan ohjelmakokonaisuuden englannin kielellä ni sehän ***viittaa suoraan siihen kielitaitoon jota me me ollaa niinku totuttu muutenki arvioimaan***
Interviewer: mmh
Assessor 3: siis se että arvioidaan ***kielellistä kokonaisuutta***

Assessor 3: well this being able to design a full programme in English well it ***refers directly to the language proficiency that we are used to assessing anyway***
Interviewer: mmh
Assessor 3: so it is the assessment of ***a linguistic entity***

(3)

Assessor 1: ja ***ei virhei- kielivirheitä*** ku se on esite siinä ei sen pit- kieli pitää olla ymmärrettävä ei siinä saa olla olla näitä näitä tota liian vaikeita sanoja ja ennen muuta termejä et ***tämmösissä virallisissa asiakirjoissahan*** on juuri se että että jos ei sulle ihan selkeetä oo se suomenkaan asiakirjatermistö ni ***sä sit haksahdat***

Assessor 1: and ***no grammatical mistakes*** because it is a brochure and the language should be understandable it mustn't contain these too difficult words and especially terminology the thing ***with these kinds of official documents*** is that if the finnish terminology used in documents isn't really clear to you either then ***you make mistakes***

(4)

Assessor 5: ja näissä pystyy kattoo sitte että että tota myöskin ton kohteliaisuusfraseologian ja ja sen että tota öö moni asiakaspalvelutyössä oleva ihminen hallitsee niinkun sanotaan semmosen niinku puhekielen aika kivasti
Interviewer: mm
Assessor 5: mut sit se että miten miten tota ***miten tää näytönantaja osaa sit muokata sen kirjalliseen muotoon*** niin se on kans aika aika olennainen

Assessor 5: and in these you can then also check the politeness phrases and that err many people working in customer service have quite a nice command of spoken language
Interviewer: mm
Assessor 5: but then ***how the candidate in a competence test can then transform it into a written form*** is also pretty essential

(6)

Assessor 5: siinä sit sanotaan et ei hallitse kielen perusrakenteita ja ammattisanastoa ni tässä onkin tavallaan niinkun annettu jo se alataso että minkä alle ei voi mennä [...] jotenkin se että että ***tää on aika karkee tää näyttöjen välinen ero osaa ja ei osaa***

Assessor 5: it then says that a person does not have the command of the basic structures of the language and professional vocabulary so that in a way it already gives you the lowest level that you can't go under […] ***somehow this difference in the test between not knowing and knowing is rather crude***

(7)

Assessor 1: et osaako tää henkilö öö erottaa erilaiset kielen ilmaisut öö sillon kun ei oo suomen kieli kyseessä elikä jos suomessa sanotaan et mee tonne ***nin englantilainen ei koskaan sano mee tonne***

Assessor 1: if this person can tell the difference between various expressions in a language when we're not talking about the finnish language so that if one says in finnish go there ***then an english person never says go there***

(8)

Assessor 3: et se tulee aika nopeesti oikeestaan siit viestimisestä läpi että millä tavalla on niinku sisäistäny ***sen kohteliaisuusnormin joka kuuluu siihen kulttuuriin*** ja joka tulee esiin suoraan sen kielen kautta

Assessor 3: it becomes clear quite quickly through communication how someone has adopted the ***politeness standards that are part of that culture*** and come out directly through the language

(12)

Assessor 1: mutta ***ei nää musta kuulu kielen osaamisalueisiin***
Interviewer: mm-h joo
Assessor 1: et se on sikäli paha jos matkailukielen käyttö hotellin asiakaspalvelussa hä- öö hylätään ***jos kaks oo o on hylätty*** okei jos siitä hylätään esimerkiks että minkälainen hän on työyhteisön jäsenenä et siihen ei oo saatu oikein kunnollista arvioijaa taikka sitten miten hän arvostaa omaa työtänsä ***ni hä- häneltä on matkailukielen näyttö hylätty***

Assessor 1: but ***in my opinion these are not part of language proficiency***
Interviewer: mm-yeah
Assessor 1: so that it is bad in the sense if the use of a tourist language in hotel customer service is a fail ***if there are two fails*** okay if someone gets a fail for example on how he or she acts as a member of the working community when no suitable assessor is available or on what kind of appreciation the person shows towards his or her own ***job then the competence test on the use of a tourist language is a fail***

(15)

Assessor 1: että se aukikirjottaminen öö siinä pitäs **ehdottomasti** olla kielenopettaja joka kirjottaa niitä auki [….] koska ***jos*** **hyvin** ***on laadittu kriteerit*** ni sillon se ***arvioija ei tarvii olla niin kauheen kauheen hyvä öö näyttökokeen arvioija*** koska et sä et sä välttämättä poimi niitä näyttökokeen arvioijia kielen arvioijia joka paikasta ei niitä **ole**

Assessor 1: the process of opening up should **definitely** involve a language teacher who opens them up […] because ***if the criteria are*** **well** ***designed*** then ***the assessor doesn't have to be terribly good as an assessor of competence tests*** because you can't necessarily get assessors of competence tests assessors of language proficiency just like that they just **aren't there**

(16)

Assessor 5: se että tota öö että ne kriteerit on niinku ***avattu tommosiks hyväksytty hylätty öö pareiks*** esimerkiks ni se on mun mielestä ***erittäin hyvä*** mä oon havainnu sen toim- niinkun hyväksi toimintaperiaatteeks
Interviewer: mm
Assessor 5: ***mut se tarkkuus sitä kaipaan***

Assessor 5: that the criteria ***are opened up like that as pairs of acceptable and unacceptable performance*** for example is ***very good*** in my opinion I have noticed that it's a good operating principle
Interviewer: mm
Assessor 5: ***but it's the precision that is lacking***

(17)

Assessor 3: no mitenköhän mä selittäsin tän ku must tuntuu et mä luen nää kriteerit useimmiten siinä vaiheessa ku mä suunnittelen sen valmen- valmistavan koulutuksen mut sit ***ku mä oon tehny tätä niin pitkään*** ni esmes ku ***mä teen yrityksille niit kokeita*** ni ni mä oon niinku tavallaan tehny sit vielä sen ***oman kriteeristöni***
Interviewer: mm
Assessor 3: joka niinku pohjaa tietysti tähän ja pohjaa sit siihen kokemukseen mikä mul on ollu […] ***mä en ehkä*** henkilökohtasesti itte ***oo seurannu näitä kriteerejä ihan just tän mukaan*** mitä täs on

Assessor 3: well how should I explain this I feel that I mostly read these criteria when I plan the preparatory training but ***since I've been doing this for such a long time*** for example when ***I design tests for companies*** I have also kind of compiled ***my own criteria***
Interviewer: mm
Assessor 3: that are naturally based on these and on the experience that I have had […] ***I haven't perhaps*** personally ***been following these criteria just as they are described here***

(18)

Assessor 1: ne pitää olla selkeet **kenelle tahansa** kielenopettajalle pitäis olla selkeet kenelle tahansa ee muun aineen kri- näyttökokeen vastaanottajalle ne pitäis olla ihan selkeet
Interviewer: joo
Assessor 1: että ei se ei se riitä että tulee toimeen se ei oo mikään kriteeri tulla toimeen eli kuka päättää kun sitten että et mitä on se toimeen tulon raja ja ja just nimenomaan se että se on kova paikka ***jos sä hylkäät millä perusteella sä perustelet että että ei tullu toimeen***

Assessor 1: they must be clear to **any** language teacher should be clear to anyone err to assessors of other subjects they should be quite clear
Interviewer: yeah
Assessor 1: that it's not enough that someone gets by to get by is no criterion so who decides then what is the borderline of getting by and it is definitely a tough decision ***if you fail someone on what grounds do you justify that he or she didn't get by***

(19)

Assessor 3 : jos aattelee näitä ääripäitä nin ne ne vaatimukset on aika erilaiset sit että
Interviewer: mm mm
Assessor 3: ni sellasta täsmällisyyttä toivos näihin kriteereihin ettei kenenkään niinku opettajan tai niinku arvioijan tarkotan ***tarviis jäädä epätietoseks*** siitä että tota että mitä mikä niinku missä kulkee se raja että ***kuka ja millä perusteella on hyväksytty ja kuka ei***

Assessor 3: if you think about these extremes the requirements are quite different
Interviewer. mm mm
Assessor 3: so that one would hope for precision into the criteria so that no one I mean a teacher or an assessor ***would have to remain uncertain*** about what you know where the borderline goes regarding ***who passes and who doesn't and on what grounds***

(21)

Assessor 1: esimerkki tästä vois olla se et mä olin mä olin tuolla [--] vastaanottamassa näyttöö joss oli kaks kaks tota kaks meijän harjottelijaa […] mä kuuntelin sen respan joka siellä oli siis ***ammatissa toimiva ihminen öö eikä mikään harjottelija*** eikä mikään just sinne tullu ja kyse oli siis [--] jota se on nyt aika arvostettu kuitenki ihan keskellä helsinkiä
Interviewer: mmh
Assessor 1: ja tota sinne tuli tähän öö puhelu asiakas sieltä jostain huoneesta siel oli jotain hässäkkää ja ja ku hän sitten vastas siihen nin hän sano jotain tähän tyyliin että ***i come wait there*** siis tyyliin [tähän näin]
Interviewer: [(nauraa)] aha joo joo
Assessor 1: ***tää on se*** mitä mitä mä niille ***mitä mä siellä haen*** ja jo usein niinku sanonkin sitte että mietipä uudestaan mitä varten mitä varten englantilainen ei sano tai **yleensä** mene suoraan asiaan vaan että […] could you wait ni joku tulee sinne pian

Assessor 1: an example of this could be when I was over at [--] assessing a competence test and there were two of our trainees there […] I listened to the receptionist who was ***a professional working there and no trainee*** nor a newcomer and we were at [--] which is pretty distinguished and right in the middle of helsinki anyway
Interviewer: mmh
Assessor 1: and well there was a phone call from a guest from some room there was something going on well then this receptionist answered the phone and said something like ***I come wait there*** you know [like that]
Interviewer: [(laughing)] aha yeah yeah
Assessor 1: ***this is what I always look for out there*** and I often already ask them to think twice why an English person doesn't say or **usually** go straight to the point but […] could you wait someone will be with you soon

(26)

Assessor 1: mutta sillon kun tulee näitä joista ***selvästi näkee et ei ei niinku sinne päinkään oo taitotaso*** nin sillai jää miettimään että ***mitä mä tästä nyt sitten pistän hylätyks***
Interviewer: mm mm
Assessor 1: et se on kun se ***kuitenki täytyy aina sille henkilölle perustella***
Interviewer: nii
Assessor 1: että esimerkiks sen takia että jos hän haluaa parantaa hän haluaa joka tapaukses tehdä sen uudestaan nin ***mitä hänen pitää harjotella***

Assessor 1: but when you get these people from whom ***you can clearly see that no way are their skill levels acceptable*** it makes you think ***what am I going to mark as a fail then***
Interviewer: mm mm
Assessor 1: because ***you'll have to give some explanation to that person anyway***
Interviewer: yeah
Assessor 1: because for example if they want to do better they'll want to do it again anyway so ***they need to know what to practise***

(27)

Assessor 5: kun sitä käytiin sitä arviointikeskustelua siinä siin oli ***kaks muuta arvioijaa*** ja minä nin tota me käytiin sitä keskustelua siinä niin öö sen niinkun huomas että me oltiin niinku aikasemmin nää kriteerit jo luettu ja omaksuttu tietysti mutta se että ei meillä ollu niinkun mun mielestä kun me käytiin läpi sitä sitä suoritusta sitä näyttöö niin ***en mä voinu olla varma että ymmärretäänks me nää kriteerit samalla tavalla***

Assessor 5: when we were having the assessment discussion there were ***two other assessors*** with me so we had this discussion and you could tell that we had read these criteria before and adopted them of course but the fact that we had no I mean when we were going through that performance ***I couldn't be sure whether we understood the criteria in the same way***

(28)

Assessor 3: siinä näyttötilanteen aikana ei tuu useimmitenkään niinku näiden kriteerien kanssa ongelmia mutta mutta sit joskus siel valmistavas koulutukses tulee […] siel oli tullu tämmöstä palautetta että ei he et ***että en* minä *joudu omassa työssäni mitään tämmösiä asiakirjoja laatimaan*** nin tota mä oisin sanonu tohon et kuule ei tää oo mikään tää ei ole kurssi joka valmentaa sinua sinun omaan tämänhetkiseen työhösi
Interviewer: mm
Assessor 3: vaan ***tässä on nou- noudatettava näitä mitä on tutkinnon perusteissa***

Assessor 3: during the competence test there are mostly no problems with these criteria but sometimes during the preparatory training there are problems […] there had been some complaints that they don't ***that* I *certainly don't have to draft any documents like these in my job*** so I would have said hey listen this is not a course that prepares you for your present job
Interviewer: mm
Assessor 3: but ***we have to follow what is stated in the qualification requirements***

(33)

Assessor 5: ja mä toivosin semmosta ***ruohonjuuritason yhteistyötä*** että ne ihmiset jotka jotka ni et olis niinku semmonen jonkinlaista kunnon yhteistyötä oltiin me sitten minkä tahansa tason toimijoita ni että ***enemmän sellasta avoimuutta tähän***

Assessor 5: and I would like some sort of ***cooperation at the grass roots level*** here so that the people who so that there would be some kind of proper cooperation no matter what kind of role one plays so ***more openness into this whole thing***

(34)

Assessor 1: ainakin sen voi öö sen öö kielitaito öö ***taitokuvauksen siihen tutkinnon perusteisiin niin tehdä kielen arvioijan kanssa***
Interviewer: joo
Assessor 1: että arvioijilla on helpompaa ja sitte se ***että molemmat ymmärtää sen asian samalla lailla***
Interviewer: joo joo
Assessor 1: et se on must semmonen suur kehittämisen paikka

Assessor 1: at least the language skills err ***skills description for the qualification requirements can be made together with an assessor of language skills***
Interviewer: yeah
Assessor 1: so that it's easier for the assessors and ***that they will both understand a certain point in the same way***
Interviewer: yeah yeah
Assessor 1: so this in my opinion is an issue for major development

(35)

Assessor 3: ja sit näis näytöis on tota tosiaan se et siinä arvioinnissa **pitäis *olla mukana jossain ominaisuudessa myös kielen opettaja***
Interviewer: aivan joo

Assessor 3: et vaikka se tapahtus sitte tällä taval että ni niitä ***nauhotetaan ja sit joku vaikka ni kuuntelee*** sitte [ja antaa oman]
Interviewer: [mm-h mm-h]
Assessor 3: lausuntonsa siitä

Assessor 3: and in these competence tests ***there*** **should** ***be a language teacher involved in some role***
Interviewer: yeah right
Assessor 3: even if it were so that ***you record them and then someone for example listens to them*** [and gives their own]
Interviewer: [mm-h mm-h]
Assessor 3: opinion about it

(36)

Assessor 1: hiukan tarkemmin kannattas ottaa ehkä huomioon mitä ne joutuu siinä tekemään että öö ***jos öö kertakaikkisesti ei koskaan joudu töissään mitään tuollasia kirjeitä kirjottelemaan*** nin miks ne on sitten sisällytetty tohon tohon näytön kriteereihin et ku siel on eri ihmiset jotka jotka sen hoitaa

Assessor 1: it would be worth taking better notice of what they have to do there so ***if someone simply never has to write any such letters*** then why are they included in the test criteria when there are other people taking care of such things