http://www.jyu.fi/library/tutkielmat/ 337/

UNIVERSITY OF JYVÄSKYLÄ

COMPARABILITY OF A TAPE-MEDIATED AND A FACE-TO-FACE TEST OF SPEAKING
A triangulation study

A Licentiate Thesis

by

Sari Luoma

Centre for Applied Language Studies
1997

HUMANISTINEN TIEDEKUNTA
SOVELTAVAN KIELENTUTKIMUKSEN KESKUS

Sari Luoma

COMPARABILITY OF A TAPE-MEDIATED AND A FACE-TO-FACE TEST OF SPEAKING. A triangulation study.

Lisensiaatintyö
Soveltava kielentutkimus
Helmikuu 1997 140 sivua + 16 liitettä

Tutkielman tarkoituksena on selvittää studiossa ja kasvokkain suoritettavan puhumistestin vertailukelpoisuutta. Tutkittavat testit oli kehitetty vaihtoehtoisiksi tavoiksi toteuttaa puhumisen osakoe yleisten kielitutkintojen keskitason testissä. Tavoitteena oli selvittää, mitä yhtäläisyyksiä ja eroja testien välillä oli; mitattiinko niissä esimerkiksi samoja taitoja, ja millaisia tulkintoja niistä saaduille testituloksille saattoi antaa. Koska aineisto kerättiin testien kehittelyvaiheessa, näitä voitiin samalla esitestata, ja arviointitietoa voitiin käyttää myös koko tutkinnon arviointijärjestelmän kehittämisessä.

Vertailukelpoisuutta tutkitaan monimenetelmätutkimuksen avulla. Tutkimus koostuu kolmesta osatutkimuksesta, joissa käsitellään testitilanteeseen osallistuvien ihmisten — sekä testattavien että arvostelijoiden — havaintoja, testidiskurssia ja arviointia. Osanottajien havaintoja tutkittiin kyselyiden ja haastattelujen avulla. Testidiskurssia analysoitiin testi-instrumenttien ja testisuoritusten osalta erikseen, testi-instrumenttien osalta lähinnä kuvaillen. Testisuorituksia luonnehdittiin aluksi sana- ja puheenvuoromäärien avulla, sitten analysoiden tuotettuja illokutionaarisia funktioita ja suorituksissa käytettyjä sanoja ja rakenteita. Arvioinnin analyysi koostui arvosanojen numeerisesta vertailusta sekä strukturoidusta haastattelusta, jonka avulla tutkittiin sitä, minkä arviointikäsitteiden avulla arvostelijat tekivät eroja suoritusten välille. Tutkimukseen osallistui 37 testattavaa ja kaksi arvostelijaa.

Tulokset osoittivat, että tutkittujen studiotestin ja kasvokkaisen testin välillä oli sekä yhtäläisyyksiä että eroja. Vertailukelpoisuuden hyväksymisen tai hylkäämisen määrää se, kuinka tärkeänä päätöksen tekijä pitää testien välisiä eroja. Yhtäläisyys testien välillä näytti perustuvan siihen, että niissä käytetään samoja sanoja ja rakenteita. Erot liittyivät testien luonteeseen. Studiotesti oli rakenteeltaan sidottu: siinä arvioitiin sitä, miten testattava osaa sanoa testaajan määräämät asiat. Kasvokkainen testi oli avoimempi, ja myös testiin osallistuja vaikutti siihen, mitä testissä mitattiin. Arviointi keskittyi selkeämmin kykyyn pukea omat ajatuksensa ymmärrettäväksi ja järkeväksi viestiksi. Arvosanojen välinen korrelaatio oli kuitenkin suhteellisen korkea: .85. Yhden testisuorituksen perusteella pystyi siis ennustamaan toisen arvosanan 72% tarkkuudella. Jos lopputulosten yhdenmukaisuus on tärkein päätökseen vaikuttava tekijä, testien vertailukelpoisuutta ei tämän tutkimuksen avulla voi kieltää. Valinnan määräävät silloin käytännöllisyys, kielipoliittiset tekijät, sekä kaupallisessa kilpailutilanteessa mahdollisesti myös osallistujien mielipiteet. Tässä tutkimuksessa testattavat kannattivat selkeästi kasvokkaista testiä.

Asiasanat: testing of speaking. triangulation. test development. assessment analysis.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

The present study started from a real question in a test development project. The test contained a subtest of speaking, and two working versions of the subtest had been developed, one tape-mediated and the other face-to-face. These were considered the two realistic alternatives between which the new test system would have to decide, because the inclusion of both was likely to prove too expensive and therefore impractical. Experience from a previously developed test, which had included both a tape-mediated and a face-to-face test of speaking, indicated that each of the tests provided some unique information on candidates' skills when both were used in tandem. The degree of overlap between the two test modes had not been investigated empirically, however. Previous experiences also indicated that different tasks suited each mode, while closely parallel tasks often worked well in one mode but less so in the other.

The cumulated experience and the demand in the new system to only choose one of the two modes resulted in the parallel development of two possible test versions. Each test on its own, independently of the other, should elicit a sample which reflected the candidates' ability to speak the language while it should also be ratable according to the criteria used and implementable considering the resources available. Because the possibilities for communication in the two modes were different — the communication in the tape-mediated test was one-way while in the face-to-face test it was two-way—, each test was developed to reach this common aim through different means. The test development aim was to make the best possible use of the opportunities for testing that the mode provided, while not losing sight of the need to develop a high quality test test which ran smoothly.

Having developed pilot versions of two different tests of speaking, the test developers were interested in finding out whether two different constructs were being measured, and what implications the choice of one over the other had for the interpretation of the scores. The developers knew that for reasons of practicality, the test would first be introduced with just a tape-mediated test of speaking. However, it was possible that a face-to-face test might replace the tape-mediated test at a later stage if this was deemed desirable.

In designing the present study, previous studies comparing tape-mediated and face-to-face tests of speaking were used to help create the design and formulate the research questions. Chapter 2 summarises the findings from

previous research, and presents the design of the present study. Chapter 3 describes the test context from which the data for the present study came, as well as the data used in the three studies. Chapters 4-6 present the results of the three studies, Chapter 4 dealing with participant perceptions, Chapter 5 with test discourse, and Chapter 6 with assessment. Each of the chapters contains separate discussions of the results and the methods used. Chapter 7 summarises the major findings from the present study, making connections between its parts, as well as considers the implications of the results. The limitations of the present study are also discussed, and directions for further research outlined.

2 PREVIOUS STUDIES ON COMPARING TAPE-MEDIATED AND FACE-TO-FACE TESTS

The research on the comparability of tape-mediated and face-to-face tests of speaking can be divided into two broad categories: score comparability and linguistic/interactional comparability. The division corresponds roughly with a quantitative/qualitative division in methodological terms. Quantitative studies focus on the interchangeability of the two test modes in the light of assessment results, and commonly conclude that numerically the two modes are very closely comparable. Qualitative studies investigate the linguistic and especially interactional differences in test performances between the two test modes, and often suggest that there are important differences between the tests as communicative events. Although the studies broadly address the same question, comparability, the two types of studies use different methods and different data, and ask different questions.

However, the differences between the two groups of studies cannot be entirely reduced into the qualitative/quantitative dichotomy, as the primarily quantitative studies use soft data such as assessor or candidate comments as supporting evidence, and the qualitative studies regularly quantify the features studied. Furthermore, although some numerical studies do seek and give yes-no answers, these studies also highlight the necessity of understanding the construct measured, and thus making accurate and truthful interpretations of test results.

2.1 Studies of score comparability

A number of studies into score comparability were conducted in connection with the development and validation of Simulated Oral Proficiency Interviews (SOPIs) at the Center for Applied Linguistics (e.g. Shohamy et al. 1989, Stansfield 1990, 1991, Stansfield and Kenyon 1988, 1989, 1992, Stansfield et al. 1990). The main research questions in these studies deal with test quality and substitutability. Both the design and the argumentation in the studies are influenced by the fact that the tape-mediated alternative, the SOPI, is a new test, whose properties are compared with those of the already established Oral Proficiency Interview (OPI), a face-to-face interview originally developed by three government agencies in the United States. The OPI consists of a systematic interview by a trained interviewer and often contains a role-play segment (Lowe and Stansfield 1988). The assessment in the OPI is conducted on live performances immediately after the interview is finished, while on the SOPI the performances are taped and the time of assessment is independent of the time of testing. In the SOPI-OPI studies, the OPI is used as a yardstick against which the SOPI is measured. The validity of the yardstick is not questioned.

Most of the SOPI-OPI studies aim to prove the acceptability of the tape-mediated tests. Stansfield (1991:202-206) begins the list of the advantages of the SOPI with higher interrater reliability. According to him, this was possibly partly due to the longer speech sample collected in the tape-mediated tests than through the live OPI procedure. The SOPI also offers an opportunity to select which raters to use as assessors, and selecting the most reliable raters improves overall reliability. Validity is increased by providing the same quality of elicitation to everyone, eliciting speech on a greater number of topics and including more role plays in potentially more credible contexts. Close attention to all aspects of the elicitation procedure when developing parallel forms is also mentioned as an advantage for validity. The practical advantages are that the test can be administered to multiple participants by someone who is not a trained interviewer, and assessed by trained raters when convenient. This may even save the costs of testing.

In spite of the long list of advantages, Stansfield (1991:207) concluded that he viewed the OPI as potentially more valid and reliable "when carefully administered by a skilled interviewer and rated by an accurate rater", and that

in its adaptability it would be more suitable for the extreme ends of the ACTFL scale than the SOPI. The reasoning behind the last argument relies on the acceptance of the ACTFL scale and procedures: since performances on the lowest skill level consist of fragments and not the kinds of sentences and connected language required by the SOPI, the SOPI is an inappropriate instrument for that level. Correspondingly, because drawing distinctions between the highest skill levels requires extensive probing and this is impossible in the SOPI format, the distinctions are more reliably made in the live version of the test. It would be interesting to investigate the suitability of the two tests for various kinds of candidates empirically.

Stansfield's arguments for the advantages of SOPI are largely pragmatic, and solid connections between the claims and empirical data are made in the cases of score reliability and score comparability only. The rest of the claims offer several alternatives for future investigations into the comparability of the two tests, some of which are outlined in Stansfield and Kenyon (1992). If future investigators adopt the current mainstream conceptualisation of construct validation in the field of educational measurement (e.g. Messick 1989), which stresses the importance of both using multiple sources of evidence and observing the consequences of test use, studies of the nature of the construct measured and of the interests of various stakeholders in test use would have to be included in the designs alongside investigations of reliability and score comparability.

In the studies comparing the SOPI and the OPI, the proof of test quality is the reliability of scores, and the proof of substitutability is the correlation coefficient between scores from the two tests. Reliability in the SOPI-OPI comparisons is operationalised through inter-rater reliability and parallel-form reliability. This is investigated through correlations and later (Kenyon 1991, Stansfield and Kenyon 1992) through generalizability studies. The reliabilities reported vary between .89 and .99. Likewise, the correlations between the test results from the OPIs and the SOPIs are impressively high, usually .90 or above. Thus, a generalizability study with data from five test languages (Kenyon 1991:7) confirms that the two test modes "are functioning similarly". The "functioning" here apparently refers to tests as score-producing machines, with the focus on the scores rather than the modes or the instruments. This goes against the intuitions of many candidates and researchers that the tests are different in many respects, particularly that the processes of taking these two

types of test and possibly the constructs behind the tests are different. If these intuitions have a firm basis in the nature of communication in the two tests, one possible explanation for the high correlations between the scores is that the scores may represent different constructs from the tests that produced them. In order to investigate this, both the tests and the scores should be investigated within one and the same design.

One score comparability study between tape-mediated and face-to-face tests of speaking has been conducted in a different setting from the SOPI-OPI one. Wigglesworth and O'Loughlin (1993) investigated the speaking part of the Australian *access:* test (the Australian Assessment of Communicative Skills in English), a four-skill English proficiency test for certain categories of intending migrants. Parallel versions of the speaking subtest for face-to-face and tape-mediated administration have had to be developed because practical circumstances often only allow the use of one rather than the other. The design and implementation of this comparability study (Wigglesworth and O'Loughlin 1993) reflects the envisioned large scale use of the test under investigation: 94 test-takers and 13 assessors took part in the trial [1] . Half the test-takers took the face-to-face version and half the tape-mediated version first. As a concurrent validation measure, the face-to-face performance of each test-taker was also assessed according to the ASLPR (the Australian Second Language Proficiency Rating, the Australian equivalent of the OPI) scale immediately after each interview was completed. In accordance with the procedures followed in the operational versions of the test, all the *access:* assessments were conducted from audio tapes.

Wigglesworth and O'Loughlin (1993) investigated the discrimination power of *access:* tests in the two modes as well as overall item difficulty, comparability of candidate scores, and concurrent validity with the ASLPR oral component. The data were analysed in an item response theory framework with five numerical variables as facets: candidate ability, rater severity, criterion difficulty, test type — tape or live —, and order of tape or live. The results indicated that from a statistical point of view, the two versions of the test were highly comparable, especially after some problematic criteria and tasks were removed. A correlation of .92 was found between the

---

[1]Due to technical difficulties in audiotaping the performances, the IRT analyses of comparability were conducted on 83 performances.

final scores from the two modes. Moreover, many of Stansfield's (1991) arguments for and against tape-mediated tests gained support from the study, with the exception that the participants overwhelmingly preferred the face-to-face format. Two cautionary notes conclude the study, however. The scores appear to be somewhat susceptible to practice effects, i.e. when there was a discrepancy between the test results with respect to a specific cutoff point, the individuals always did better in the test form they completed second. Furthermore, Wigglesworth and O'Loughlin stress that while score comparisons indicate comparability between the two modes, there may be significant differences in other aspects of test performance which are not reflected through numerical scores.

The SOPI-OPI family of studies and the *access:* study answer numerical comparability questions relevant to each test's operational reality in the affirmative. Both lay claims to reliability and numerical validity. The results are convincing if one accepts that the results from two different processes of quantification are commensurate. Both Stansfield (1991:205-206) and Wigglesworth and O'Loughlin (1993:57-58, 66) point to the necessity of qualitative studies into the discourse produced by the two test modes. Neither of the studies mention the need for investigating the assessment processes, taking for granted that identical assessment scales and labels guarantee that assessors work similarly with both kinds of performances.

2.2 Studies of linguistic/interactional comparability

In contrast to the statistical studies, studies of linguistic/interactional comparability between tape-mediated and face-to-face tests tend to find differences between the two modes. Shohamy's (1994) study comparing SOPI and OPI in Hebrew is a case in point. The starting points were the high concurrent validity coefficients (from $r = .89$ to $r = .92$) between the two modes on the one hand, and concern with the sufficiency of this data for proving test substitutability, on the other. Shohamy's (1994) study compared both test tasks and test performances; the former on the basis of test specifications and manuals, and the latter on the basis of ten transcribed performances in each mode. The rationale for selecting the ten performances was not explained. The

results indicated that the comparability of the two test modes was not straightforward. According to some criteria, e.g. the number and types of errors made, no real differences between the two modes were found. According to other criteria, most notably ones having to do with the types of discourse elicited, important differences appeared between the two modes. Consequently, Shohamy (1994:119) concluded that in some contexts, a valid assessment of oral proficiency may require the use of both types of test. She further stressed that the decision of which test to use ultimately rests with administrative bodies, who need to be provided with a thorough understanding of what the tests are measuring. Such information can only be provided through examining tests from multiple perspectives.

Shohamy's plea concurs with current views in educational measurement, especially Messick (e.g. 1988, 1989).The central idea is that the interpretation of scores cannot be based on any one kind of evidence such as scores from two or more tests, but needs to be supported by other kinds of evidence. This can be provided for instance by investigating the other elements involved in the production of scores, i.e., the tests and assessment systems, the candidates, and the assessors; or by investigating the use of scores, and comparing this to the interpretation warranted by the test process. High correlations between scores from two measures can result from the fact that they measure the same construct, but they can also possibly be explained by similarities between aspects of the procedures, such as test methods, raters or rating scales, or by the fact that the scores only reflect some aspects of the measures but not others. In the case of tape-mediated and face-to-face testing of speaking, one of the starting points for those seeking differences may be the gut feeling that there must be differences simply because the processes look and feel so different. If an explanation could be found for the high correlations between scores, more informed judgements on the credibility of the correlations, and the substitutability of the tests, could be made.

The range of perspectives that Shohamy (1994) employs in examining the Hebrew SOPI and OPI is impressive: the functions elicited by the test tasks and the topics covered, the number of certain categories of errors in morphology, syntax and lexicon, the mean number of five communicative strategies employed in each mode, lexical density, rhetorical structure, genre, speech moves, communicative properties, discourse strategies, content and topics, prosodic and paralinguistic features, speech functions, discourse

markers, and register. However, it is not always clear how the aspects were operationalised, and how some of the aspects are inter-related.

Of particular relevance to the present study is the way communicative functions of the tasks and performances were portrayed. The face-to-face test was reported to be the more varied one in the case of Superior level test-takers in this respect. The only data presented to support this result came from analysing the test manuals, not actual performances, and no attempt appears to have been made to create a common terminology for describing the functions in the two tests. Rather, a list combining the terminologies used in the two manuals was employed. The functional aspect in the part of Shohamy's study where transcribed performances were used is spread over several categories: rhetorical structure, speech moves, discourse strategies and speech functions. Differences in discourse were reported in all of these areas and illustrated by sample clips of performances (ibid pp. 109-114), but none of the features was quantified and no analytical rationale was presented.

Shohamy's (1994) finding that the face-to-face test elicited a different and possibly more varied range of functions than the tape-mediated one was striking because it ran counter to the hypotheses made during the development of the Finnish test which provided the data for the present study. During the development of the test it had been assumed that the tape-mediated test with its array of situations and topics would have an advantage over the face-to-face one in this area. Study Two in the present study investigated whether the assumption of wider coverage of the tape-mediated test in functions and topics was true, and what kinds of similarities and differences in communicative functions could be found in the performances on the tests.

The discrepancy between the expectations of the National Certificate test developers and Shohamy's results may be partially explained by the difference in target audiences of the tests. While particularly the OPI is intended for all proficiency levels from beginner to near native, both of the National Certificate tests investigated were intended for intermediate level candidates only. Extensive probing was consequently not one of the aims in either version of the National Certificate tests.

The second study comparing discourse in tape-mediated and face-to-face tests of speaking summarised here, O'Loughlin (1995), took its material from the Australian *access:* test and focused on one aspect of test discourse, which was lexical density. O'Loughlin's analysis was based on 20 transcribed

performances, a stratified random sample from a pool of 94 performances. Ten of the selected candidates had completed the tape-based test first, ten the live one. The working hypotheses in the study were that test format (tape-based or live), task type, and the interaction between test format and task type have an effect on the lexical density of the elicited performance. The tasks investigated were description, narration, discussion and role play, all of which were included in both formats. The role plays were necessarily different: the live version consisted of a two-way exchange and the tape-based test of a unilateral message being left on an answering-machine.

O'Loughlin (1995:227-228) carefully reported which items were counted as lexical and which as grammatical, and how high-frequency lexical items were distinguished from low-frequency ones in order to also produce weighted counts of lexical density where high-frequency lexical items are assigned half the weight of low-frequency lexical items. His results indicated that there was indeed a difference in lexical density between the two modes in the expected direction, but the effect was considerably less dramatic in the case of the *access:* test than in the case of SOPI and OPI. Moreover, the difference was most prominent where test format and task type interacted, the key variable being the degree of interactiveness required by the tasks and allowed for by the test format. The clearly lower values for lexical density in the face-to-face format appeared in the role play task where the interlocutor had to take an active role in the interaction, while on the tape-mediated test the lexical density figures remained comparable to those in the other task types. O'Loughlin's conclusion was that if the tests remain as they were, i.e. interlocutor involvement is kept at a minimum in the face-to-face version of the test, there was no conclusive evidence that the tests were not tapping a common underlying construct.

Two important issues for the comparability of face-to-face and tape-mediated tests of speaking emerged from this study. One was that task types may introduce variation in language elicited but they do not necessarily do so, at least on all indicators. This was illustrated by the result that the lexical density values hardly varied at all in the four different tape-mediated tasks that O'Loughlin studied. The variation was equally small between three of the four tasks on the face-to-face test, but larger between these and the fourth task. The other important issue arising from the study was that the degree of interactiveness in the face-to-face test, i.e. the extent of involvement of the

interlocutor in the test discourse, may have a strong influence on the kind of language elicited in the test.

The third study on interactional differences between tape-mediated and face-to-face tests of speaking to be summarised here was Shohamy, Donitsa-Schmidt and Waizer (1993 / no date). They studied the effect of five 'elicitation modes' on the discourse elicited in the tests: tester face-to-face, candidate face-to-face, tester through telephone, tester through videotape and tester through audiotape. There were three tasks in all the 'modes': telling about oneself, complaining, and making a request. In contrast to the SOPI-OPI study (Shohamy 1994) in which the tasks in the different tests were different, the tasks in the Shohamy, Donitsa-Schmidt and Waizer study were very similar. In fact, there were only two variants of the complaint and request tasks across the five 'modes', and telling about oneself was the same task in all the 'modes'. The design thus had the participants perform the same tasks two or three or even five times, but it did offer a clear focus on the effect of elicitation modes as the tasks were held constant. The performances of ten candidates on all the five 'modes' were analysed.

The patterns in the results of Shohamy et al (1993 / no date) were often clearer when the live interactions - tester face-to-face or through telephone, and peer face-to-face - were grouped together as one set, and the tape-mediated interactions - video and audio - as another. The differences in the scores were inconsistent, with sometimes the taped tests and sometimes the live ones yielding higher scores. It must be mentioned, however, that the ratings on accuracy and fluency were made from transcripts, not from tapes. This decision obviously affected the ratings by excluding all or, depending on the type of transcription used, most properties of the sound of the candidates' speech, such as speed, rhythm and intonation. The assessment results might have been very different had the ratings been made from audio tapes. No discussion of the auditory properties of the candidates' speech was included in the report. The justification for rating from transcripts was that the assessors were then less able to tell which mode each speech sample originally came from. The drawbacks of this decision were not discussed.

In the analysis, some linguistic differences between the modes were found. Polite and softening expressions tended to be used more frequently in live interactions than in taped ones, but on the other linguistic criteria used (e.g. sharing personal information, degree of directness, I/You ratio) the

differences between the modes were not significant. Instead, variation on these features was found between the tasks 'complaint' and 'request' but not between the modes. Candidate preference was clearly for the live interactions with the tester, while reactions towards interacting with peers as well as towards tape-mediated tests were negative.

The finding that language samples vary according to test task: telling about oneself, complaining, and making a request, concurs with O'Loughlin's (1995), though the discourse features investigated in the Shohamy, Donitsa-Schmidt and Waizer study were politeness- and contextualisation-oriented rather than lexical density-based. The study also introduced an empirically-developed discourse analytic instrument, which can be used as a model for developing contextually sensitive instruments for other test comparison contexts. The instrument elicits subjective assessments on a five-point scale on several task-derived variables such as "way of addressing: formal ... personal", "rhetorical structure: listing .. narrative" and "genre: written ... spoken" as well as counts of specific linguistic features such as softening expressions and I/You ratio. The instrument is particularly useful for semi-structured tasks where specific expectations of candidate output can be formulated in advance, but could conceivably be developed for analysing less structured tasks and performances as well. The fact that the analyses in the Shohamy et al. study were made on transcripts rather than audio recordings probably facilitated the use of the fairly extended instrument.

In order to gain a comprehensive understanding of the comparability between tape-mediated and face-to-face tests of speaking, O'Loughlin (1996) undertook an ethnographic study on the issue, again looking at the *access:* test. He gathered data on the development of the two test versions, the training of interlocutors and assessors, the analysis of test results, interlocutor, candidate and rater feedback, and the revisions made on the test afterwards, as well as followed two candidates through the testing process by observing their performances and interviewing the two candidates and their interlocutors. The study revealed the complexity of the issues involved in comparing tape-mediated and face-to-face tests when not only the potentially conflicting aspects of reliability, validity, practicality and acceptability but also the views of different participant groups are taken into account.

The multiple perspectives of O'Loughlin (1996) are reminiscent of Shohamy (1994), and while the focus is on the test process rather than

discourse produced, results indicate a similar clonclusion: comparability looks different depending on which features are observed and which roles taken into account. One of the results was a .80 correlation between the ability estimates from the two test modes in a cohort of 94 candidates, the lowest such correlation yet reported in comparability studies. Supported by the results from other parts of the study which suggested differences between the two modes, O'Loughlin's conclusion was that neither the constructs tested in the two test versions nor the scores obtained were equivalent. Notably, O'Loughlin's study also introduced the notion of test quality as a concern for comparability on an empirical rather than anecdotal basis.

2.3 Summary of previous research and implications for the present study

The evidence on the comparability of face-to-face and tape-mediated tests is inconclusive and even conflicting. Most comparisons of scores yield the result that tests in the two modes are closely comparable (though see O'Loughlin 1996), the correlations ranging between .80 and .96. Interestingly, Shohamy et al. (1993 / no date) reported that there were statistically significant differences in scores between the elicitation modes in different tasks, but that the direction of the differences varied so that higher scores were sometimes achieved in the tape-mediated and at other times in the face-to-face test. When the different influences are combined into an overall score, some of the differences that may exist in analytic scores if such are given may be obliterated. Furthermore, scores have been investigated quantitatively but not qualitatively, and the investigation of scoring procedures might shed more light on which aspects of assessments are similar between the two modes and which are different.

The studies on test discourse, in contrast, indicate that there are important differences in the characteristics of the discourse on which the comparable assessments are made. The features where the differences are most apparent are related to the fact that one test contains real-time human-to-human interaction while the other does not: the number of softeners, the length of turns, the amount of elaboration, sharing of personal information. Features which have not been found to distinguish between the modes include types and frequencies of errors and hesitation.

Previous studies on the comparability of tape-mediated and face-to-face tests of speaking have focused on scores on the one hand, and the language elicited in the tests on the other. These are important dimensions, but other dimensions along which the test may differ also exist. Kenyon (1991) identified the test, the assessment system, the participants, and the assessors as key elements in the testing process in the communicative era. The nature of all of these elements, and their influence on the scores, need to be investigated. Milanovic and Saville (1995) reported on similar views among the Association of Language Testers in Europe (ALTE), and McNamara (1996:9) used Kenyon's model as an illustration of the various influences on scores in performance assessment. The implication for comparability studies is that all of these elements should be included in the design. This was attempted in the present study.

The goal with the present study was to develop a comprehensive understanding of a tape-mediated and a face-to-face test of speaking and investigate their comparability. To serve this aim, a design consisting of three related studies was drawn up. Study One looked at the perceptions of the human participants in the testing process, i.e. the candidates and the assessors. Study Two investigated test discourse, concentrating on the elicitation instruments on the one hand and test performances on the other. Study Three focused on the assessment system, looking into both the comparability of the scores and the features of test performance that the assessors paid attention to.

The data for the present study came from a pilot administration of the tests being investigated. In December 1993 and January 1994, a total of 37 voluntary candidates participated in pilot versions of the two tests. For reasons of practicality, all the participants first took the tape-mediated test and then the face-to-face one. After taking each test they filled in a questionnaire. Three participants also volunteered for a post-test interview, during which they viewed their performances on both tests. The face-to-face test was conducted with one of two interlocutors, and the performances were video-taped. The tape-mediated test performances were audio-taped. Both performances were assessed from the tapes by the two interlocutor-assessors. One week after the tests were administered, the assessors participated in a structured interview where their assessment constructs were investigated on the basis of a sub-sample of the performances. These data were investigated through three studies, the first concentrating on participant perceptions, the second on test

discourse, and the third on assessment. The aims and scope of each of the studies are described below.

The employment of multiple methods and data sources in the three studies briefly outlined above makes the present study a triangulation. This strategy potentially gives more depth to the understanding achieved at the end, though as Mathison (1988:13) warns, the result of triangulation is divergence or conflict as often as convergence. All of these results should be explained in the end with reference to both the specific situation which provided the data and a general understanding of that and other situations in which the phenomenon investigated can occur.

2.4 Aims of Study One

Study One investigated the ways in which the 37 candidates and the two assessors perceived each of the two tests of speaking. This was done through questionnaires and interviews. The respondents reported on their experiences during the test and assessment, and compared their perceptions of the two tests with spoken language use in non-test conditions. All the candidates filled in a post-test questionnaire after each test, while three volunteers participated in the post-test interviews. The post-test questionnaires focused on perceived fairness of the tasks and their effectiveness in bringing out the candidates' ability to speak English. The three post-test interviews focused on the respondents' impressions of the differences between the processes of taking the two tests, and the comparability between these processes and speaking outside the test situation. The assessors' opinions on the fairness and effectiveness of the two tests as well as their perception of the usefulness of the elicited performance for making a reliable rating were investigated through open-ended questions, which the assessors answered directly after completing each candidate's assessment sheet. After all the assessments had been completed, the assessors had an evaluative discussion on their perceptions. The data were analysed according to favourable and disfavourable dispositions, and grouped to clarify the reasons for these opinions.

The following questions were posed in Study One:

1a. What is the candidates' perception of the two tests of speaking, particularly with respect to fairness?

1b. What is the assessors' perception of the fairness and effectiveness of the two tests and of the assessability of the performances as samples of proficiency in speaking?

2.5 Aims of Study Two

Study Two investigated the discourse in the two tests. The analyses focused on the elicitation tasks as well as the discourse elicited. The task analysis was mainly based on Bachman and Palmer's (1996) framework of Task Characteristics, and aimed at describing the similarities and differences of the two tests as contexts for language use. The analysis of test performances had two aims. One was analysing the communicative functions elicited by the two tests, the other was exploring other ways of analysing discourse data from tests of speaking. The communicative functions were analysed because the test system was supposedly based on a functional view of language and because this aspect had been assumed to be one of the advantages of the tape-mediated test over the face-to-face one. Other methods of analysis were sought to balance this view of what was elicited in the two tests.

Study Two was based on the transcripts of ten candidates' performances in both modes. The research questions in Study Two were:

2a. How do the two test modes compare as contexts for discourse elicitation?

2b. Which communicative functions are elicited by each mode?

2c. Which other features of performance than functions should be analysed in oral test discourse in order to account for similarities and differences between test modes?

2.6 Aims of Study Three

Study Three investigated the two tests from the point of view of assessment. Both quantitative and qualitative methods were used. The quantitative assessment analyses concentrated on the candidates' scores from the two tests, and consisted of descriptive statistics and correlations. The aim was to assess the degree of association between the scores and to identify the differentially

assessed performances. These performances were then analysed to explain what might have caused the difference in assessment. The qualitative assessment analysis focused on how the scores should be interpreted. An attempt was made to specify which features the assessors paid attention to in each test mode while making their assessments. This was done through employing a version of Kelly's (1955) Repertory Grid technique, which made it possible for the assessors to consider potential assessment criteria one by one across several performances, and explore the connections between the potential criteria. The research questions were as follows:

3a. Is there a difference in the candidates' scores on the two test modes?

3b. Which features of performance do the assessors pay attention to when making their assessments in the two modes, and are the features the same in both of them?

3 THE DATA IN CONTEXT

As a key feature in the present study is triangulation and as the researcher's task in such a study is to make sense of the results through a holistic understanding of the problem, some treatment of the context from which the data for the present study came is necessary. In the context of a project evaluation, Mathison (1988:16) identifies the nature of the project, its history, the intentions of the developers, and the ongoing relationships within the project, among others, as possible sources of explanations for convergence, divergence and contradiction in the data. The present study constituted one part in a test development project, and a general understanding of the main project is needed not only as source for explanations for results but also in order to explain some features in the data and the methods used. The presentation of the data for the present study therefore begins with a description of the project where the data came from, followed by a brief description of previous stages in the development of the tests. This will be followed by a description of the tests investigated in the present study, as well as a description of the participants in the study.

3.1 The test context

The project that the present study constituted a part of was the development of the National Certificates of Language Proficiency, a system of general purpose tests of second or foreign language designed for adult learners in Finland. The target group distinguishes this test from other foreign language tests in Finland, which have traditionally been aimed at children or young people, and have usually been bound up with specific educational settings. There are three test levels in the Certificate system: Basic, Intermediate and Advanced, and eight test languages: Finnish, Swedish, English, French, German, Spanish, Russian and, recently, Italian. The development of the tests is a joint project between the National Board of Education and the University of Jyväskylä. The tests are aimed at adults who wish to work towards specific goals in their language learning or who need proof of their language skills for their prospective employers[2]. The Certificate system is independent of any educational settings.

The primary purpose in developing the examination was to provide a national test system which would produce comparable certificates in several different languages and be generally available all across the country often enough for the system to be useful for the purposes intended. Before this test system, no national general purpose language proficiency examinations existed for adult examinees; the only equivalent was the Matriculation Examination, an academically oriented achievement test at the end of the 12th year of school. A number of international tests are administered in a few test centres in the country, but they are mostly taken by young adults wishing to study or work abroad. Furthermore, international tests in only three languages are being offered on more than one location. A decision was made in favour of developing a general purpose test rather than a system of specific-purpose tests because of the size of the market on the one hand and the large proportion of common purposes for which languages are used by people in widely different areas of speciality on the other.

The development of all the tests in the Certificate system was based on a common set of test specifications, which in turn build on three core

---

[2] The test in Finnish is also sometimes used when admitting immigrants to secondary level study, but never as the sole criterion for admission.

definitions: a scale of eight proficiency levels (Appendix 1), a taxonomy of topic areas (Appendix 2), and a taxonomy of language functions (Appendix 3). The taxonomies were created with adult language users in mind, and the cultural background of Western adults is one of the presuppositions of the test. The National Certificate tests at all three test levels consist of five subtests: Listening, Speaking, Reading, Writing, and Structures and Vocabulary, and specifications have been written separately for each of the subtests on each of the test levels. Figure 1 illustrates the framework of the National Certificate test system.

3.2 The development of the Certificates prior to the present study

At the beginning of the present study, the development of the Certificates had been going on for nearly two years. The key definitions of the system – the core scale, the functional taxonomy, and the test specifications – had been written and re-written a few times, and the development of these will be briefly described below. Previous work on the test of speaking will also be relevant. The report covers the definition of the construct of speaking, previous developments on assessment criteria, and some findings from the piloting of previous versions of the oral test.

The national predecessor of the Certificate system was an advanced-level test for business and administration, the Finnish Foreign Language Diploma for Professional Purposes (Sajavaara 1992; Huhta *et al.*, 1993). The test was developed in English, German and Swedish, and it was offered on a demand-supply basis on two locations. The idea for developing a lower-level certificate was born within that test system, also developed at the University of Jyväskylä, and when the Certificate project started, the six-level *core scale* (see Sajavaara 1992:142) from that project was adopted for further development. The scale was expanded to comprise nine levels for being able to show progress throughout the learning process and especially in the beginning (North 1993:48-52 summarizes the arguments on the number of bands to be used). In the interest of promoting transparency of certificates within Europe, the English-Speaking Union's 9-level scale (Carroll and West 1989) was selected as one of the most influential models, with the exception of their Level 1, which was considered too low for formal certification purposes. The National Certificate scale was developed by the Language Testing Group at the

Figure 1 The Certificate framework

Core definitions of language	Test specifications	Tests
View of proficiency: 8 skill levels (see Appendix 1) Description of language use: - topic categories (see Appendix 2) - language functions (see Appendix 3)	Basic level Intermediate level Advanced level Specifications for Reading, Writing, Listening and Speaking common to all test languages; Structures and Vocabulary language-specific	Basic level 5 subtests, 2.5 hours speaking tape-mediated Intermediate level 5 subtests, 3 hours speaking tape-mediated Advanced level 5 subtests, 4.5 hours speaking both tape-mediated and face-to-face

University of Jyväskylä, but a record of the development work has not been published.

When developing the *functional taxonomy* used in the Certificates, several alternatives had been considered. Because the system had to be comprehensible to teachers as well as to prospective candidates, the Threshold Level categorisation of language functions (van Ek and Alexander, 1980) was selected as a basis. This approach concentrates on what the language user "does" with the language, the 'interpersonal' metafunction in Halliday's (1994:xiii) view of language. Functions under the interpersonal metafunction are active, involving the language user in producing language, as opposed to 'ideational' metafunctions, which are reflective and do not necessarily involve overt production. The same choice of concentrating on active functions had been made in the language textbooks which were consulted during the development of the functional categorisation.

The types of functions included in both the Threshold level and textbook inventories of language functions are typically individual speech acts like 'promising' or 'correcting'. They denote a 'micro' level of action, but the actions can nevertheless be verbalised in several different ways. Illocutionary acts such as those which Austin (1962) and Searle (1969) raised for focal attention in interpersonal communication form the core of these lists. This kind of linguistic analysis has intuitive appeal, because it seems to enable the analyst to summarise various kinds of language use in a meaningful way, and the existence of such inventories in present-day language learning materials proves that textbook writers and possibly users think they have some practical

value.

There are a few problems in developing functional taxonomies, however. The functions which are most easily categorisable tend to be those which are frequently used in social contacts under events which are easy to name and schematize, such as introducing people to each other, or having a meal at a restaurant. The most clearly identifiable functions in these contexts have a single, clear purpose, involve a limited amount of linguistic variation, and are often expressed through formulaic expressions. The opposite cases, where the situational context is more vague and speaker intentions are more difficult to predict, the functions are much more difficult to define or even name. Yet if all language use is to be covered, both types of functions must be included in the inventory.

Similar difficulties are met in creating superordinate categories for the system, and in dividing individual functions into the superordinate categories. The taxonomy will thus contain functions and categories whose informational value and explanatory power will differ. That is, various functions will help to various degrees in explaining exactly what a language user can or cannot do. Other descriptors apart from functions, such as topic, the number of words, the exact lexical items used, level of formality, number of errors, intonation pattern, or rate of speech, will be needed to various degrees depending on which function is being considered. "Thanking, three words", specifies the content of a turn to a much higher degree than "stating an opinion, five words". Very few functions other than perhaps the most formal of performatives are clear enough to specify exactly what was said and so other descriptors than functions will always be needed, but in some cases functions can be very helpful in specifying language use.

When a functional taxonomy is developed for characterising learner language, the developer faces the task of not only categorising the functions of language but also describing how learners proceed in learning to express the functions. The OPI way of describing how language learners proceed in their command of a language appears to be that they learn language function by function, and the further advanced the learners are, the more functions they can use. In the test system where the data for the present study came from, a different hypothesis was advanced. According to this hypothesis, language learners can, if necessary, produce various kinds of functions, but their language ability determines the degree of linguistic sophistication with which

they can express them. Beginning learners are unlikely to actively seek out similar demanding language use situations as advanced users, such as for instance negotiating business deals, but should they meet such situations they would try to cope with the linguistic resources they had. Thus, the functional taxonomy developed for the National Certificates defined a basic core of functions, and described the extent to which mastery was expected on various levels.

When the functional taxonomy for the National Certificates was being developed, no learner data were directly used. Instead, the Threshold level, Munby's (1978) inventory of micro-functions, and several language textbooks were used as materials. In terms of superordinate categories, the system came to resemble the Threshold Level inventory the most. However, the categories of "expressing and finding out emotional attitudes" and "expressing and finding out moral attitudes" were merged, and communication strategies was introduced as a category. As the objective was to make the taxonomy as comprehensible as possible, some names of categories and individual functions were changed. The resulting categorisation (Appendix 3) has six categories: giving and asking for factual information, expressing one's point of view, expressing emotions and attitudes, acting through words (performatives), acting according to social norms and customs, and communication strategies. Each main category has several sub-categories, all of which are not expected to appear in every test.

The first two categories, giving and asking for factual information, and expressing one's point of view, contain functions with wide scope and high frequency, such as stating, describing, asking for factual information, and answering in the positive or the negative. The performative and social functions, in contrast, are relatively restricted in scope. Individual functions in the first two groups are on a high level of abstraction and apply to very different kinds of language use. Functions in these categories can thus be expected to be most frequent in test performances as well as in language use outside test situations.

The National Certificate was planned as a multi-language, multi-level system from the outset, so the need for a *specification* for test writers was evident. The idea of a written specification also suited the criterion-referenced ideal of the project. The development of a detailed test specification was strongly emphasized in the early days of the criterion-referenced measurement

movement (e.g. Osburn 1968, Baker 1974, Millman 1974, Popham 1978). The goal then was to define the domain, or item pool, precisely enough to allow generalisations from test performance on random items. While all test specifications fail to fully reach this ideal goal, the benefits of specifying the construct or the skill being tested and the task of the item writers cannot be over-emphasised. More recently, the importance of test specifications has been pointed out by Weir (1988), and a concrete, practice-oriented development of Popham's rubric has been published by Davidson and Lynch (1993). Their model was followed in the development of the Certificate specification because it was found to be useful in detailing the purposes which a specification should serve.

The specifications are the same for all the languages of the certificates, and have been written separately for the five subtests. Each specification contains a definition of what is to be tested in what time and through approximately how many tasks. It also specifies possible sources for task materials, lists the kinds of task types which have been found to suit the testing of the skills specified, and presents sample items. Finally, the specification describes the assessment procedures for the subtest and gives instructions on how to write the scoring guide.

Alderson et al. (1995:11-21) point out that different groups of people need different kinds of information, and thus separate specifications for at least test writers, test validators and test users will be needed. In the Certificate system, writer and validator specifications have to date not been separate documents, while user specifications do exist and contain much less detail than the test writer specifications. The test writer specifications are used throughout the test construction process and revised annually. The documents are confidential. The user specifications have been published in Finnish, Swedish and English (National Board of Education, 1995), and are revised only as major changes occur.

The Certificates' working definition of the construct of speaking is presented in Figure 2. The definition is an attempt to describe, on a fairly general level, the kinds of skills to be tested in the Speaking subtest. In the test writer specifications, this definition would be followed by suggestions for possible operationalisations in various types of test tasks as well as examples of good and bad test tasks from previous pilots and operational tests. In the speaking tasks on the Intermediate level test, the candidates are expected to

Figure 2 The Certificates' working definition of the construct of speaking

In the context of the Intermediate Certificate in English, speaking refers to the candidates' predicted ability to communicate orally* in familiar tasks and situations related to work and freetime. This means that

a) they understand and make themselves understood in oral interactions in English where basic communicative needs are fulfilled and/or common topics not requiring specialist knowledge are discussed,
b) they can, with occasional support from a sympathetic listener or given some time to prepare, present their views and opinions on a common topic clearly and comprehensibly,
c) when meeting with linguistic difficulties, they can use compensatory strategies such as circumlocution in familiar situations and on relatively concrete topics, while on more abstract topics may have to give up the topic or limit what they can say,
d) they can use and are familiar with the usage of the most common conventions of English such as forms of address, greeting and politeness, and some ways of softening the effect of potentially face-threatening speech acts, and
e) they have comprehensible pronunciation (while non-native features not impeding comprehension are fully acceptable), and speak at a speed approaching normal careful speech rate.

*"Communicating orally" refers to the candidates' ability to use English interactively in communication. The object of measurement in the test is the test-takers' ability to express themselves in English. Their use of communication strategies to compensate for gaps in linguistic knowledge is recognised as a positive factor.

interpret and produce language with content and form roughly appropriate to the different aspects of the situations presented to them.

The working definition of speaking concerns oral production for interactive purposes. As comprehension is an integral aspect of interaction, some comprehension criteria are also included in the definition, most clearly in point (a) understanding and making oneself understood, but also in point (d) familiarity with conventions, because the choice of forms to be used often depends on comprehension of the communicative context. The definition of speaking is written so that it would, in a sufficiently abstract, non-language-specific way, define what the test-takers are expected to be able to *do* with language on this test level. The definition is thus primarily functional and focuses on discourse: points (a) and (b) on types of discourse and the limits of topical and functional ability, point (c) on what happens in the discourse when a participant has insufficient language knowledge, point (d) on some sociolinguistic aspects of discourse, and point (e) on clarity of presentation.

In relation to the Canale and Swain (1980, Canale 1983) model, the definition is intended to cover all the four major parts presented: grammatical, sociolinguistic, discourse and strategic competence. The more structurally

oriented parts of the model, grammatical and discourse competence, can be found, somewhat mixed, in the definition in points (a), (b) and (e). The more performance- and function-oriented parts of the model, sociolinguistic and strategic competence, are found as separate points ((a) and (d) respectively) in the Certificates definition of speaking. The emphasis is clearly on Canale's (1983) dimension of Communicative language proficiency.

With reference to the Bachman and Palmer (1996) model, all the parts of their language knowledge, plus strategic competence, are included in the definition. Points (a) and (b) are most concerned with grammatical, textual and propositional knowledge, point (c) with strategic competence/metacognitive strategies, point (d) with functional and sociolinguistic knowledge and point (e) again with grammatical knowledge. The greater specificity of the Bachman and Palmer model compared to its primary source, Canale and Swain (1980), makes it more useful for explicating what different tests test. Particularly useful for the present study are the definition of task characteristics and the refined (Bachman 1991) definition of strategic competence.

Along with the overall scale, the Certificate system also inherited an analytic *assessment scale for speaking* (see Huhta et al., 1993: 145-150; 155), but doubts were raised over how well the six-criterion system would fit a low-level test. The criteria were thus taken over with the intention to simplify the assessment procedure in some way, to be specified through assessment practice in conjunction with pilot tests. At the beginning of the present study, the original criterion set of Pronunciation, Speech flow[3], Grammatical accuracy, Vocabulary, Appropriacy, and Discourse management, had been developed into a system where only one holistic assessment was given, with instructions to pay specific attention to task fulfilment (content coverage and comprehensibility), linguistic criteria (level of vocabulary, speech flow, accuracy) and situational appropriacy (structural/phrasal selection, intonation). Content coverage focuses on the *informative content* of the utterance(s), and how much of it was comprehensible. Linguistic criteria are for assessing how well the test-taker managed this with reference to the *accuracy* norms of the language tested, while the situational appropriacy criterion was intended assessing the conformity of the performance to the

[3] As in the Diploma test, the Certificate developers wanted to stress the narrow definition of this feature of spoken language and not use the extremely fuzzy term 'fluency' (cf. Huhta *et al.* 1993:147).

appropriacy norms of the language tested with respect to the situational context provided in the test task.

The clearest change from the Diploma system of assessment in the Certificates was the abolishment of the analytic marks. A notable change in the bases of assessment was the introduction of task fulfilment. This reflects the expectation that not all candidates would be able to fulfill all the tasks, at least not equally fully. Four of the Diploma criteria, Pronunciation, Speech flow, Grammatical accuracy, and Vocabulary, were collapsed into one, whereas Appropriacy remained a separate class. Discourse management disappeared as a criterion simply because there was no real discourse in the tape-mediated test. In the form where only an overall score was given for speaking, none of these aspects was clearly distinguished in the Certificates, however. Through the present study and other similar ones, the aim was to re-introduce analytic marking in some form.

As part of the quality control aspect of test development, the Certificate tests begun to be piloted as soon as versions became available. This resulted in modifications of task types and changes on emphasis. On the tape-mediated version in the Intermediate level test of English, the Simulated Conversation tasks were modified to include some elements of comprehension, the Reactions in Situations prompts were developed to require longer responses, and the longer elicitation task was changed from a picture-based to a linguistic prompt-based form. The length of the warm-up section was reduced, and preparation and answering times were adjusted. The format of the test also changed in that the master tape became monolingually English while the instructions in the test booklet continued to be given in Finnish or Swedish. In the face-to-face version, a prompt-based discussion task replaced a simulated interaction task. This was because some candidates found it hard to act in the simulations, and also because the interpretation of the results was difficult. Power of imagination and acting ability appeared to play a considerable part in succeeding on the task. These developments were made as a result of piloting, and similar changes are made periodically during the operational use of the test as well.

The few completed and available studies of the National Certificate tests by the end of 1996 are by Halvari (1994, 1996). Halvari's study was designed as a preliminary step towards the possible future recognition of the Certificates within Europe. It was a comparability study between the Intermediate level

National Certificate and the German International Certificate Conference Certificate in English. Halvari compared the content of the tests, the candidates' marks, and the candidates' reactions to the tests. One of the most salient results was that as a large proportion of the tasks in ICC tests were multiple choice, the test takers found the test boring, while the National Certificate test was considered more interesting and stimulating. The reverse side of the coin was that the ICC test was perceived to be more reliable and more test-like, while the open-ended tasks in the National Certificates was less favourably received, because the candidates found it difficult to assess whether their answers were correct and/or complete enough.

Halvari's (1996) results of the comparison between the two oral tests were mixed. While the ICC test of speaking was performed face to face, the rubric was highly constrained and the test very short. This baffled some candidates, though the majority preferred the face-to-face test. Content-wise the tests covered a similar range of language, with informational functions being more stressed on the ICC and social functions on the National Certificate. The correlation between the results was very low, .33. Halvari did not compare the performances elicited in the tests.

There is a recent collection of articles on the National Certificates, which has appeared in Finnish (Opetushallitus, 1997). The collection of articles covers issues such as the development of the assessment criteria, the relationship between the tests and language teaching, and the use of the certificates. One article gives voice to those involved in the test system: test writers, test administrators, and language teachers.

3.3 The tests investigated

The tests used in the present study had different task structures. An overview of the two tests is given in Table 1. The tests are described briefly in turn below. A detailed analysis of the tests is presented in Chapter 5.

3.3.1 The tape-mediated test

The tape-mediated test had four tasks. The first was a warm-up task of reading aloud, and this was followed by three assessed tasks: Simulated Conversations,

	Tape-mediated test		Face-to-face test	
Warm-up	Reading aloud	4 min	Getting acquainted	1-4 min
Task 1	Simulated Conversations	7 min	Peer Discussion	5-15 min
Task 2	Reacting in Situations	9 min	Individual Discussion	3-7 min
Task 3	Presenting Views and Opinions	10 min		
Table 1	**The tests investigated**			

Reacting in Situations, and Presenting Views and Opinions. The total test time was 30 minutes, of which the candidate was engaged in speaking for approximately 15 minutes. The rest of the time was spent on reading the instructions, viewing the test tasks, and planning the answers.

The candidates wore a headset, through which they heard the key instructions and the prompts. In addition, they had a test booklet, which included complete instructions and the task material. The instructions in the test booklet were given in Finnish, while the task material was written in English. Finnish was used in order to make sure that all the candidates understood with ease what they were to do. A similar bilingual procedure with respect to test booklets is followed for instance in the SOPI tests, but in the SOPIs the tape is also bilingual. The tape was monolingually English and gave key instructions and the prompts; it also contained silent pauses for the candidates to read the full instructions in the test booklet. The tape included British, American, and Australian speakers of English, and both male and female voices. The voice giving the instructions was British.

In the warm-up task, the candidates acquainted themselves with the text, and then, after being prompted to do so by the tape, read it aloud onto the tape. The first assessed task, Simulated Conversations, included three conversations: an exchange at a shop; a telephone conversation with an airline operator; and a small talk situation at a coctail party. Each context was briefly described in the test booklet, and the skeleton of the conversation, with cues as to what the candidates were to say, was printed below each description. The candidates heard their conversation partner's turns from the tape, and reacted according to the cues, or, when there were none, according to what was appropriate in the context. The turns that the candidates were expected to make were 1-2 functional units long, e.g. greeting, telling why you need to change the time of your return flight, or explaining that Finland is not a part

of Russia. Approximately half the turns in the simulated conversations were cued, the other half had to be completed solely on the basis of what was heard on the tape. The performances were assessed turn by turn on a two-point scale of fully appropriate (2), not quite appropriate or causing some difficulty in language terms (1), and entirely inappropriate or unanswered (0).

The Reacting in Situations task included ten situations, which the candidates were allowed to glance through before the recording began. They were then asked to read the descriptions to themselves one by one, and prompted by the tape to say what they might have said in such a situation. The task included situations such as greeting acquaintances, reporting the loss of a camera at a police station, and telling a tourist how to get to a flower shop. The turns required in this task were 3-6 functional units or up to 40 seconds long. The performances were assessed turn by turn on a three-point scale of fully appropriate (3), not quite fully appropriate or causing some difficulty in language terms (2), hardly appropriate or clearly difficult to comprehend from a language point of view (1), and entirely inappropriate or unanswered (0).

The presentation task included two presentations, one on the candidates' attitudes towards popular music, and the other on leaders and leading. Each topic had five support questions for the candidates to start from if they so wished. They had two minutes to prepare each mini-presentation, and two minutes to speak. The task was assessed directly on the assessment scale (see Appendix 1). To form the overall score for the tape-mediated test, the point scores were added up and transformed into a skill level. This score was then combined with the scores for the two mini-presentations to form an overall score for the tape-mediated test.

3.3.2 The face-to-face test

The face-to-face test used consisted of a short warm-up phase where the candidates and the interlocutor got acquainted, and two test tasks: a peer discussion among two or three candidates and a more interview-like discussion between the interlocutor and each individual candidate. The total test time was 20 to 35 minutes, depending on whether the group was a pair or a triad, and on the dynamics of the interview situation. The time spent actively engaged in interaction was intended to be 10-15 minutes per candidate but in practice it

varied between 9 and 22 minutes. This was because the test was being piloted, and interlocutor instructions were not clear enough on managing test time.

Before the candidates entered the examination room, they were given an information sheet, which described the phases of the test in English. The candidates had approximately three minutes to read the description before entering the interview room. The description informed them that after introducing themselves briefly, they would be given a choice of three topics, and they should decide as a group which one they wanted to discuss. They would then discuss this topic for a few minutes with minimal intervention from the interlocutor, who would only come in at the end to draw the discussion to a close. After this she would ask one or two of the candidates to go out of the room so that she could interview each candidate without the others listening in. During the test, the interlocutor checked that the candidates had understood the procedure, and gave them spoken instructions as necessary.

The task material on the face-to-face test was written on two kinds of cue cards, one with three possible topics to choose from, and the other with a set of five or six support questions on a single topic. Once the pair or triad had chosen the topic they wanted to discuss, they were given the support questions on their chosen topic. During the interview phase, the support questions were used by the interlocutor. Each candidate's test thus revolved around two topics, neither of which they knew in advance, but one of which they were able to choose from among three. All in all, there were twelve topics available as material for the face-to-face test.

The face-to-face test was recorded on video-tape, and the camera was clearly visible in the room. The performances were assessed directly on the assessment scale (see Appendix 1). Separate assessments were made on each candidate's performance in the peer discussion and in the individual interview, and the assessments were combined impressionistically to form an overall score for the face-to-face test.

3.4 The participants

A total of 37 voluntary candidates took part in the present study. Most of them had heard about the experiment from their teachers at various schools

and evening institutes; seven had not been taking courses in English for a number of years but had heard about the experiment from friends. All of the participants had had the equivalent of junior secondary school instruction in English or more, with most having taken at least some additional courses. Only two of the volunteers had spent an extended period of time in an English-speaking environment.

Before they agreed to participate in the study, the volunteers had read a brief description of the experiment. The information included the first six levels of the skill level description (see Appendix 1). The text also said that the tasks of the experiment would come from the intermediate level test, which was best suited for skill levels 3-5. No ability level check other than self-assessment was made prior to the test administration, and the language skills of eight of the candidates proved to exceed the test requirements. Where this might have had an effect on the results of the present study, separate results will be reported for the whole group and for the 29 candidates for whom the difficulty level was appropriate.

Thirty of the candidates who participated in the experiment (81%) were women and seven men. The three candidates who volunteered for the post-test interview were all women. The mean age of the candidates was 28 years, the median 25 years, and the range between 17 and 56 years. Compared with the figures from two operational rounds of tests – 71% of the participants in the operational versions were female, and the mean age was 36 years – women were slightly over-represented and older adults underrepresented in the study.

The assessors (both female) who rated all the performances in both test modes were the two most experienced assessors available in the as yet non-operational test system – the present writer was one of them. The assessors had been involved in the development of the holistic assessment scale and had used it for assessing around 200 pilot test performances on the tape-mediated test and 25 performances on the face-to-face test. They were also aware that their work in the present study was going to be used for developing the assessment scale further. Thus, the assessors had a dual role as assessors and scale constructors.

4 STUDY ONE: PARTICIPANT PERCEPTIONS

Study One investigated how the candidates and the assessors perceived the two tests. As both the candidates and the assessors are directly involved in the test process where a language sample is turned into a score, their views form an important aspect of how the tests work in reality, or what the construct being measured really is. Ultimately, the perceptions also have a bearing on how the scores can be interpreted. In their comments, the candidates compared their experience of the two tests with their experience of speaking outside test contexts. Their perceptions were thus related to the authenticity and fairness of the tests. In addition to authenticity and fairness, the assessors were also asked about assessment. More specifically, they were asked to recount their perception of the usefulness of each sample for assessing proficiency in speaking, and their confidence in the rating that they assigned.

Participant perceptions consist of affective reactions to the test, such as enjoyment, anxiety or irritation, and meta-level perceptions of the activity that the participants were engaged in. The latter perceptions covered the way that the candidates made sense of the tests, their assessments of the test tasks, and their perception of how well they did on the tasks. Such information has been considered useful for test development (e.g. Shohamy 1982, Zeidner and Bensoussan 1988), and practical applications particularly in connection with the development of tests of speaking have begun to appear in recent years (e.g. Stansfield et al., 1990, Kenyon and Stansfield, 1991, Brown 1993). Alderson (1988) reports plans in the ELTS Revision Project to use candidate perceptions not only for task improvement but also for gathering insights into aspects of test validity. Study One pursues the validation angle, but adds a new dimension in not only including candidates' but also the assessors' perceptions and evaluations of the test process. The data gathered through post-test questionnaires were also used to assist immediate test development concerns. The questionnaires thus also contained questions which were not directly relevant to Study One.

The specific research questions addressed in Study One were:

1a. What is the candidates' perception of the two tests of speaking, particularly with respect to fairness?

1b. What is the assessors' perception of the fairness and effectiveness of the two tests and of the assessability of the performances as samples of proficiency in speaking?

4.1 Methods used in the investigation of participant perceptions

The data on participant perceptions were gathered through questionnaires and interviews. The candidates filled in post-test questionnaires after each test, and three volunteers were interviewed one day after the test. The interviews were based on a loose protocol, and were conducted in between reviewing each candidate's performances on the two tests. The assessors filled in a questionnaire immediately after completing a face-to-face test assessment, and after both assessors had completed the assessment of all candidates in both modes, they had a discussion on their observations of the assessment process.

The assessor questionnaires and discussions, as well as the candidate interviews, were entirely geared towards eliciting data for Study One, while the candidate questionnaires only contained a few questions directly relating to the present study. The rest of the questions on the post-test questionnaires addressed immediate test development issues, attempting to identify remediable test characteristics which potentially caused construct-irrelevant variance in the results. Although both assessors acted as interlocutors in the face-to-face test as well, the data were collected after the assessment and mainly focused on their views as assessors.

Because of insufficient planning, the post-test questionnaires (presented in Appendix 4) only had three questions in common: whether the respondents had taken a test of speaking in the mode concerned before, whether their performance on the test gave a fair impression of their ability to speak English, and why, and whether they would like to change the test they had just taken by adding or taking away something. These questions provided comparable data for both tests.

There were three questions which provided useful data for Study One but which only appeared on one of the post-test questionnaires. Two of these questions were in the face-to-face test questionnaire: one focused on test anxiety, and the other on impressions of test length. The third unmatched question was in the tape-mediated test questionnaire, and inquired about

impressions of comprehensiveness through correspondence with real life situations. The results of these questions will be reported, but conclusions regarding comparison between the two tests cannot be drawn.

There was one question in the tape-mediated test questionnaire which did not work because its wording proved too vague. It enquired how the recent test experience relates to the way in which the candidate's spoken English should be tested in order to gain a fair impression of it. The answer itself would be interesting, but perhaps too complex to obtain through writing. Such questions could perhaps more usefully be made in an interview situation, where the interviewer can follow up on the initial, necessarily vague answers.

Apart from these questions, which were directly related to Study One, the tape-mediated test questionnaire collected task-specific feedback on all the four tasks, and the face-to-face test questionnaire on one of the tasks, the peer discussion. While these questions did not pertain directly to Study One, the open-ended answers concerning each task offered some insights into how the candidates experienced them. The answers will be briefly summarised. The complete questionnaires are given in Appendix 4.

The candidates filled in the post-test questionnaires directly after each test. All the candidates completed the questionnaire after the tape-mediated test, but seven of them did not answer the post-test questionnaire on the face-to-face test. The face-to-face test was in all cases the latter of the two tests in which the candidates participated, and the seven were probably tired or in a hurry to get to their next engagement. An attempt was made to reach these seven participants through their places of study, but unfortunately this proved unsuccessful. To make the reporting of the results as comparable as possible across the two questionnaires, the actual numbers of respondents answering in a particular way will always be reported in the text below, followed by the percentages of those who responded to the question in parentheses where relevant.

All the three post-test interviews were conducted by the present writer one day after the three volunteer interviewees had taken the test. The interviews can be characterised as exploratory or depth interviews, in the Oppenheimian (1992:67) sense, in that the aim was to develop ideas and hypotheses rather than gather facts or confirm expectations. The interviews built on a loose protocol (see Appendix 5), which was followed recursively in between listening to and viewing each interviewee's performance.

The protocol focused on how the interviewees reacted to the tests, what each task made them do, and what their impression of the fairness of the tests was. Specific features of each test type -- the role of compatibility of personalities in the face-to-face test and the role of imagery, and listening versus reading the cues on the tape-mediated test -- were of particular interest.

Because the three interviewees were volunteers, they were very cooperative and regularly provided their reasoning as well as the answers to the questions. They also took initiative in taking up new aspects of the tests where their performances gave them occasion to do so. One of the interviews lasted 40 minutes, the other two were approximately 60 minutes long.

The three candidates who volunteered for the post-test interviews were all women just under thirty. The pseudonyms Marja, Sanna, and Katja will be used to refer to them. All three had gone through the senior secondary school and had been learning English for more than ten years. Marja and Sanna achieved skill level five, the highest mark available, in both subtests, Katja got a five on the face-to-face test and a four on the tape-mediated test.

Marja spoke an average amount on both tests but claimed in the post-test interview that this was because she was tired; normally she would have been more talkative. She said that the tests were very easy for her, which surprised the interviewer, as few ceiling effects had been apparent in her performance. The peer discussion part of her face-to-face test did not go very well interactionally; both candidates said very little.

Sanna gave a voluminous and animated performance in both tests, and appeared to be a confident speaker and an experienced test taker. In the interview it transpired that this was her first tape-mediated test ever, and that she had felt fairly anxious while taking it.

Katja's tension during the tape-mediated test was easily audible from the tape, her replies were often to the point but short, and sometimes she gave up on a turn and waited for the next one in silence. Her performance on the face-to-face test was more relaxed and coherent, even though she was the least talkative candidate in her peer discussion triad. She was unhappy with her tape-mediated test performance but relatively satisfied with the face-to-face one.

Similarly to the candidate perceptions, the assessor perceptions too where gathered through questionnaires and an interview-like discussion. The questionnaire concerned the face-to-face test, as this was a specific focus in the

particular stage of test development when Study One was conducted. The tape-mediated tests had been piloted four times before, while this was the second time that the face-to-face tests were piloted, and the assessors had much less experience with this mode. The questionnaire (Appendix 6) immediately followed the assessment section on a rating sheet, and was filled in directly after the assessment of each pair or triad had been made. After all the assessments and double assessments had been made, the two assessors had an evaluative discussion on the assessment process. Since there only were two assessors, and since the present writer was one of them, the evidence is best treated as suggestive.

4.2 Candidate perceptions: post-test questionnaires

As all the candidates had an opportunity to voice their views on the tests in the post-test questionnaires, the results of these will be reported first. The results are reported separately for the questions which were the same in both post-test questionnaires, and for the ones which only appeared in one of the questionnaires. As the number of respondents was rather small and the distribution of both scores and favourable and unfavourable perceptions uneven, possible associations between the variables are more usefully reported verbally than through cross-tabulations. This is because the expected frequency of cases in many cells falls below 5 in spite of collapsing categories, and statistical significances thus make little sense. The candidate perceptions collected through questionnaires will be complemented by the results of the post-test interviews in Chapter 4.3.

4.2.1 Replies to matched questions

As shown in Table 2, the tape-mediated test situation appeared to be relatively familiar to the candidates: only five (14%) had not taken a test in the language laboratory before while 32 (86%) had. The response may be an artefact of the question, however, because it did not specifically ask about tests of *speaking*. The introduction to the questionnaire had mentioned the focus on the test of speaking, but since the question was one of the first in the whole questionnaire and since the test had also included a listening section, some of

the respondents may have answered on the basis of experience with listening tests. The corresponding figures for the face-to-face test, nine candidates (31%) had taken one previously while 20 (69%) had not, paint a very different picture. Three candidates had taken neither kind of test of speaking before. Given the answers and the early state of development in testing speaking in the Finnish educational system (see e.g. Yli-Renko 1989), it is likely that more candidates had previous experience of tape-mediated than face-to-face tests of speaking, but the size of the difference is difficult to estimate. While many will have learnt speaking skills in their language education, the testing of those skills may be new to the candidates, and this must be borne in mind when the results of the questionnaire responses are interpreted.

The question whether each test gave a fair impression of the candidates' ability to speak English and why was analysed as two questions. The open-ended answers to the "whether" question were divided into five categories:'no', 'qualified no', 'qualified yes', 'yes', and 'can't say'. The reasons given were then divided into two broad categories, positive and negative features of the tests, and grouped into classes based on the central content of the responses. In the five-category division, the difference between 'qualified no' and 'qualified yes' was the degree of negativity in the answer. The criterion of a negative beginning in combination with at least two negative features was used in determining which of the two groups a reply belonged to. "Rather fair. The only thing was that I felt I was in a 'test' and it affected my replies negatively. I think that I can do better in real life." was coded as a qualified yes, while "Not quite. The laboratory was a strange location for me (I have been in the lab only five times) and I paid too much attention to the wordings of the tasks." was coded as a qualified no. The 'can't say' classification was used only when the response explicitly said so. This only occurred twice; both comments related the face-to-face test. In binary categorisations, the 'no' and 'qualified no' were grouped together as the negative pole, 'qualified yes' and 'yes' as the positive one, and 'can't say' was coded as a missing value.

The candidates clearly felt the face-to-face test was fairer: 26 candidates (86%) thought it gave at least a relatively fair impression of their spoken English and only two candidates (6%) thought it did not (see Table 2). On the tape-mediated test 16 candidates (43%) expressed negative comments, while 21 (57%) had positive views of the fairness of the impression gathered through

Have you ever participated in this kind of a test before?

	yes	no	missing
tape-mediated	32 (86%)	5 (14%)	0
face-to-face*	9 (31%)	20 (69%)	8

Did your performance on the test give a fair impression of your ability to speak English?

	NO	qualif. NO	qualif. YES	YES	can't say	missing
tape-med.	10 (27%)	6 (16%)	11 (30%)	10 (27%)	0	0
face-to-face*	1 (3%)	1 (3%)	13 (43%)	13 (43%)	2 (7%)	7

Why was the impression fair/not fair?

	NEGATIVE properties	N	POSITIVE properties	N
tape-mediated	tests cause tension	9	wide range of situations	6
	can't negotiate/rephrase	8	had to react quickly	3
	lab unnatural	6		
	tired and not prepared	5		

	NEGATIVE properties	N	POSITIVE properties	N
face-to-face*	tests cause tension	7	normal interaction	12
	test too short	5	can decide what to say	1
	too few topics	3		
	little say over topics	1		
	depends on interviewer	1		
	tired and not prepared	1		

What would you have liked to add to or take away from the test?

		N
tape-mediated		
	nothing	24 (67%)
	lengthen pauses	4 (11%)
	add test length	3 (8%)
	add sth personal	1 (3%)
	add face-to-face part	1 (3%)
	take away simulations	1 (3%)
face-to-face*		
	nothing	21 (70%)
	add length	9 (30%)

*based on 30 questionnaires all in all

Table 2 Replies to the questions in common between the post-test questionnaires

this mode. The views on fairness were not associated with how well the candidates did on the tests: the correlations were below .10 in both test modes.

The most common reason given for reacting positively to the face-to-face test was that it consisted of normal interaction. The degree of freedom in deciding the content of what one wanted to say was mentioned by one candidate.

The most frequent negative factor related to both tests was test anxiety. Judging by the proportion of negative to positive overall responses to this question, this was more of a problem in the tape-mediated test, but it had an effect on the candidates' perceptions on the face-to-face test as well. The other negative characteristics brought up in connection with the face-to-face test were the experienced shortness of the test, the small number of topics covered, and the restricted amount of choice that the candidates had on which topics were included. Furthermore, one respondent pointed out that the effect of the compatibility of personalities between the candidate and the interviewer may have a negative effect on the score. Finally, tiredness was also included as a reason for the impression not being fair; more often so for the tape-mediated test than the face-to-face one, although the face-to-face test was in all cases the second test the candidates took. The pilot tests were arranged in the evening, and the candidates had a full day of work or school behind them. The tests took approximately 90 minutes, a considerable length of time. The lack of preparation mentioned by most of those blaming tiredness probably arises from the experimental setup, and might not figure as a cause in an operational test.

A significant proportion of the reasons given for why the impression given by the tape-mediated test was not fair were related to the lack of possibility for negotiation and the unnaturalness of the speaking situation in the language laboratory. This is in marked contrast with the face-to-face test with natural interaction as its main attraction. It confirms all the previous results of studies comparing tape-mediated and face-to-face tests, and once again points to the main difference between these two types of test: interaction versus the lack of it. The affective preference of the candidates is noted as a valid concern, and further investigation into how the difference in interactiveness shows in the language samples is carried out in Study Two, while Study Three looks at how it shows in the assessment procedures associated with each test.

There may have been some association between lack of experience with tape-mediated tests and unfavourable attitudes towards them: all the five candidates who had not taken a tape-mediated test before had the impression

that the test was unfair. However, 11 candidates who had previous experience of tape-mediated tests also thought the test did not give a fair impression of their speaking skills, so no firm conclusions can be drawn.

When the tape-mediated test was applauded, this was for the wide range of situations it included as well as for the need to react quickly, as in real life. In the light of the failure to find associations between impressions of fairness and the other variables, it appears that the candidate perceptions of fairness constituted an independent factor in their reactions to the test.

As for the face-to-face test, statistics concerning association between impressions of fairness and the other variables investigated could not be calculated. This was because there was very little variance in the candidates' impressions of the fairness of the face-to-face test, with only two candidates voicing negative opinions. This finding offers weak support to the tentative conclusion that candidate perceptions of test fairness are independent of their other reactions to the test as well as of their success on the tests.

The candidates offered few suggestions for how to change the two tests of speaking. While this does not amount to a straightforward acceptance of the two tests, it indicates that the candidates experienced few obvious problems with them. Table 2 shows that 24 candidates (67%) would not have changed anything on the tape-mediated test, while 21 candidates (70%) felt the same about the face-to-face test.

There was only one change recommended for the face-to-face test: increasing the test length. This reply was not only given by two people whose face-to-face tests truly were shorter than the recommended lower bound of ten minutes (see Chapter 3.3, p. 36 for an explanation on why test length varied) but also by others, among them the one whose test lasted the longest of all, 22 minutes. The response was given by nine respondents, eight of whom received skill level five on the test (see Figure 3). The shortness of the face-to-face test was thus experienced as more of a problem by some of the more proficient candidates. While it would be likely that a longer sample would create a more valid test, it is often a practical impossibility. Ensuring that all tests fulfill the length recommendations in the future should be possible, however.

Three respondents (8%) suggested lengthening the tape-mediated test as well, and four (11%) recommended lengthening the response times. The low numbers of these responses suggest that timing was not a major problem in the tape-mediated test. Other desired changes to the tape-mediated test had to

Figure 3 **Relationship between impressions of test length and success on face-to-face test**

FTFTOT by ADD add to f-t-f test?

Count / FTFTOT	ADD nothing (0)	ADD length (1)	Row Total
2.00	1		1
			3.3
3.00	2		2
			6.7
4.00	6	1	7
			23.3
5.00	12	8	20
			66.7
Column Total	21	9	30
	70.0	30.0	100.0

Number of Missing Observations: 7

do with the characeristics of the tasks: one respondent suggested the introduction of a personal, creative element rather than always having to perform in general situations and according to instructions, another reacted to the artificiality of the simulated interaction task and suggested that fewer of these would do, and a third one reacted to the monologue-like speaking situation, wanting to include live interaction.

All of these comments have to do with the inbuilt neutrality and generality of the tape-mediated test. The questions are exactly the same for all the candidates, and they must be suitable for all of them. Free, general prompts are avoided firstly because not all candidates can be expected to give a long enough performance without support and there is no other safety net to fall back on than the tasks, and secondly because in giving candidates the freedom of deciding the focus and structure of their performance they would also require organization and imagination skills to an extent which the developers of the present test at least felt uncomfortable with. Only small improvements in this respect seem to be possible if the mode is retained.

It is apparent from the recommendations that the candidates gave that they reacted more positively towards the face-to-face test, and did so for reasons similar to those already reported in the literature. The responses also show, however, that the candidates were not entirely unhappy with the tape-mediated test either.

4.2.2 Replies to unmatched questions

There were three questions which only appeared on one of the post-test questionnaires: two on the face-to-face test questionnaire, and one on the tape-mediated one. The unmatched face-to-face test questions dealt with test anxiety and test length, the unmatched tape-mediated question with the correspondence of the situations in the test with real world situations. In addition, the tape-mediated test questionnaire contained one question which did not work, and task-specific questions which were not directly related to Study One. The relevant aspects of the responses to these questions will be reported below.

After the face-to-face test, the candidates indicated how anxious the test made them on a five-point scale: very, rather, middling, not very, and not at all anxious. The middle point may have been difficult to interpret, and only three respondents chose it. While no-one reported having felt very anxious, there seemed to be two groups of respondents: those who experienced fair amounts of anxiety during the test, and those who did not (see Table 3). The former group was smaller, with seven candidates (23%) reporting to have felt somewhat anxious and three (10%) middling anxious, while the majority (14 candidates, or 47%) reported not having felt very anxious and a further six respondents (20%) stated they felt no test anxiety at all.

It is possible that test anxiety is in some way related to success on the test, as 15 of the 20 candidates reporting low degrees of anxiety achieved the highest skill level on the test, while the remaining five non-anxious candidates were placed on skill level four, the second highest level. However, five of the top-achieving candidates reported having felt anxious, so the relationship is not simple (see Figure 4). It is very unfortunate that no comparable information exists for the tape-mediated test. Two of the three post-test interviewees reported having felt more anxious during the tape-mediated test, but no generalizations can be made.

The response scale for test length had five points, from far too short through just right to far too long. Here, the scale appears truly to have had five points, and 14 respondents (47%) chose the middle one, feeling that the test length was appropriate. Seven respondents (23%) thought that the test was a bit too short, and the remaining five (17%) felt it was far too short. This matches the favourite addition to the face-to-face test, increased length, and

The face-to-face test questionnaire

How anxious did the face-to-face test make you feel?

very	0 (0%)
rather	7 (23%)
middling	3 (10%)
not very	14 (47%)
not at all	6 (20%)

Valid cases 30, missing cases 7

What did you think about the length of the face-to-face test?

far too short	5 (17%)
a bit too short	7 (23%)
just right	18 (60%)
a bit too long	0 (0%)
far too long	0 (0%)

Valid cases 30, missing cases 7

The tape-mediated questionnaire

How well did the situations in the test correspond with real life situations?

badly	0 (0%)
fairly well but still unreal	5 (15%)
well	28 (85%)

Valid cases 33, missing cases 4

Table 3 Replies to the questionnaire-specific questions

indicates that this is an important concern for the face-to-face test, both for test development and for participant information. It would not be practical to lengthen the test very much, but the inclusion of more topics might be possible, and clear information on the length and structure of the test well in advance of the test date could alleviate some experiences of unfairness due to shortness.

Similar information on test length for the tape-mediated test was not collected. While such information might be interesting for test development purposes and for covering one facet of the test-taking experience, it would have added little to the comparison of the two tests. The question focused on the test as it was and measured it against a comparable test of ideal length, and as the two tests were of different length to begin with, a direct comparison of answers would have been impossible.

The tape-mediated test questionnaire inquired about the correspondence of the situations in the test with real world situations. The candidates indicated

Figure 4 **Relationship between test anxiety and success on face-to-face test**

FTFTOT by TENSEYN ftf made tense, binary

FTFTOT	TENSEYN no 0	TENSEYN yes 1	Row Total
2.00		1	1
			3.3
3.00		2	2
			6.7
4.00	5	2	7
			23.3
5.00	15	5	20
			66.7
Column Total	20	10	30
	66.7	33.3	100.0

Number of Missing Observations: 7

that they were fairly happy with the test in this respect: 28 of them (85%) thought the correspondence was good. Five respondents pointed out that although the correspondence was good, the test nevertheless was very different from real life language use because it was not interactive. None of the respondents indicated that the correspondence was bad, and the remaining four did not answer the question at all. This question was not included in the face-to-face questionnaire. Had the question been included in exactly the same form, though, it would probably have worked somewhat differently. As the face-to-face test had no simulation or role play tasks, the question would have requested comparison of the abstract events of face-to-face test interaction and test-external oral interaction. While this aspect is interesting and was covered to an extent in the post-test interviews, it does not lend itself well for written elicitation. Its proper place would probably be the post-test interview, where further questions based on the answers can be made.

The tape-mediated test questionnaire further included a vague question on the characteristics of an ideal test of speaking for a fair assessment of the respondent's skills and the relationship of the tape-mediated test to that test. The question elicited a wide range of replies, usually answering either the former or the latter of the two-in-one question. The replies to the former were very vague. "As wide and varied as possible" is a representative example of

these. The replies to the latter usually compared tape-mediated and face-to-face tests with the majority favouring face-to-face ones, but some also supporting tape-mediated tests like the one they had taken. Due to the varied interpretation of the question, a result of its elusiveness, the replies cannot be classified to any greater extent. It might make sense to present the latter part of this question to the candidates, but only after they have completed both tests, so they do not have to invent the basis of comparison.

The tape-mediated test questionnaire contained some questions which were not directly related to Study One. These collected task-specific feedback from the candidates (see Appendix 4). However, the further comments sections of these questions illuminated some of the references to the unnaturalness of the tape-mediated test reported above. It is thus worth reporting these results.

The Simulated Conversations task seemed unnatural in three respects: timing, lack of non-linguistic context, and specific demands introduced by the test situation. The problem with set response times had been identified in the responses dealt with before, but in the task-specific questions it was brought up more often. If simulated conversations are used, however, there is no solution to this problem, since one can hardly expect all candidates to be able or willing to control their own tape-recorders. Voice activation of the recording equipment would be impractical simply because such equipment does not exist in all test centres. Long pauses within responses might also cause problems with this function. Furthermore, in this particular task, the "conversation partner's" turns would be cut out from the recording and assessment would become difficult.

Several respondents were disturbed by the strong focus on linguistic production in the tasks. Some found the requirement to invent the non-linguistic context demanding and time-consuming, others pointed out that they would have done better in real life because they could see expressions or point at things, or win time by gesturing and producing some sort of noises until they came up with some appropriate words. Several were aware that all of their performance that counted had to be linguistic. Lastly, a few respondents felt that combining reading and listening under time pressure in the way the task demanded was artificial. It must be admitted that reading the instructions, following the cues in the test booklet, and listening to the prompts on the tape is a demanding task seldom matched by anything in real

life. The assessors may well compensate for the unavoidable effects of the test tasks, but the majority of the candidates may not realise this.

The Reactions in Situations task exhibited the same problems as those listed above and added one element, the intrusion of translation. The situations were described in the test booklet in Finnish, and although anything said there was meant to guide the answer rather than dictate it, many candidates treated the task as free translation. The wordings of some of the situations made it a very difficult task for these candidates. This could be improved by writing the tasks in English, something that has been tried in subsequent pilots.

Some of the comments on the Extended Speaking task matched those on the interview: the number of topics was limited and the topics were given rather than introduced by the candidates, making them adapt to the test and not giving them much room to develop the task. The lack of support from an interlocutor and the missing indicators of comprehension were also noted as contributing to the unnaturalness of the test.

4.3 Candidate perceptions: post-test interviews

The post-test interviews largely confirmed the results of the questionnaire feedback. The interviewees favoured the face-to-face test, although they would have liked to increase test length and the number of topics covered. The best point of the tape-mediated test was considered to be its wide coverage of everyday situations. The specifics of the confirmatory results will not be reported below; only new insights into the ways in which the candidates perceived the tests will be brought up.

Most of the new comments on the tape-mediated test explained different aspects of why the reactions of the candidates towards it were negative. The aspects identified caused anxiety, irritation, and frustration. An acute awareness of answering times and the programmed nature of the test was brought up by all three interviewees. On the one hand, they had felt pressured to start speaking as soon as possible in order to get any sort of an answer completed; on the other, they had felt a need to fill as much of the answering pause as possible and reported feeling both anxious and frustrated if the silence

between the end of their turn and the next turn on tape seemed too long to them.

Secondly, all three interviewees reported feeling aware of the need to produce a grammatically correct answer, not necessarily while answering but certainly after having completed an answer. This sense was heightened because the performance was being taped, the error was irreparably on the tape and a correction would only draw more attention to it. All three contrasted the tape-mediated test in this respect to the face-to-face one, where their main concern had been to be comprehensible to their interlocutors.

Thirdly, the tasks were seen to guide the content of their replies to a very high degree. When the candidates followed the guidelines as best they could, their performances sometimes resembled translation, and when there was a word they did not know in English, they felt their performance was deficient, which caused anxiety. Cues which made the candidates say things they would not have said naturally were found irritating. If some cues were unnatural or partly incomprehensible, the awareness that there was no way of negotiating the task caused both frustration and anxiety.

One interpretation for why the above features caused anxiety, frustration and irritation in the candidates is that while taking the test, they seemed to be forming implicit perceptions of what the test required for a good performance. The candidates seemed to apply the red ink on their performance themselves if and when it did not result in a continuous flow of speech with the tape as a live interaction would. They would mark themselves down for failing to fill a pre-programmed answering slot, producing spoken grammar complete with false starts and disjunctive clause structures, or for not finding the exact words they felt that the tasks required for perfect performance. This tallies with Manninen's (1984:73, 89) suggestion that a speaker's beliefs about the demands of the situation and the expectations of the communication partner have a strong influence on the amount of anxiety that the speaker feels. In the present study it appeared that the interviewees perceived the demands to be different for the tape-mediated and face-to-face tests, and more strict and artificial in the case of the tape-mediated test.

All the three interviewees had had some sort of visual images of the situations portrayed on the tape. These came as a natural part of the test performance and assisted performance to a degree, but the experience was still far from a real life situation, because the images did not respond to each

candidate's unique performance. The headphones and the microphone constantly reminded the candidates that the performance was being taped. Moreover, they could hear the performance of fellow candidates; sometimes they could even distinguish individual words. If they themselves had nothing to say, this assisted them; if they had already completed their answer, it increased anxiety because others apparently had more to say on the question. The comments were not all negative though. Sanna reported that in spite of a certain level of test anxiety the tasks made her perform as she would in her mother tongue, focusing on content and not planning form before saying anything. Marja felt appropriately challenged by the tasks and overall felt that she did not do too badly on the tape-mediated test.

The great advantage of the face-to-face test was reported to be that it consists of real interaction with real people. In such a situation, the candidates reported concentrating more on content and comprehensibility and less on form. Furthermore, the experience that others listened to you while you spoke, and you listened to them and built on what they said, contributed to the experience of authenticity and the feeling of successful performance. The timing problem, so overt in the tape-mediated test, was not relevant at all. When listening to her performances, Katja remarked that the linguistic features in her face-to-face performance were quite similar to the tape-mediated one, but her impression both during the tests and afterwards was that she did much better in her live performance. The explanations that she gave were that the interview was video-taped, so she, and presumably the assessor as well, could see what she was doing during her pauses, and also that the others were waiting and allowing her to pause when she needed to, even helping her out sometimes. The ability to negotiate any tasks and decide exactly what they wanted to say increased the attractiveness of the face-to-face test to all the interviewees.

The interviewees also talked about two features of the face-to-face test which seemed unnatural or unfair to them. Firstly, the possibility of a difference in skill levels between the candidates was seen to complicate the peer discussion, as the level of the discussion would be accommodated to the lowest performer – the higher performer might not get a chance to show what she can do. Secondly, the more talkative candidates, Marja and Sanna, reported that turn-taking in the peer discussion was unnatural. Both had held themselves back and consciously tried to allow the other participant(s) time to

speak and distribute the time justly. They had also tried to regulate the length of their own speaking turns. The explicitness in turn-taking was clearly perceivable from the tapes: transitions were regularly marked by short pauses as well as visual cues. The whole situation was of course somewhat contrived, as there was no other genuine reason for the conversations than test performance. One of the expected characteristics of such an occasion seems to be the relative anonymity and distance between the participants.

All three interviewees, independently and unguidedly, developed the idea during the interview that the best topics for such an interview might not be ones of their own speciality, since they would have to express deep thoughts and create chains of argument, which they preferred not to do on a language test. This idea may partly have been influenced by a feeling that a unilateral delivery of considered and in-depth views of a topic of importance would not have suited the context of talking to an uninitiated stranger.

All the three interviewees recommended combining the two types of test for a comprehensive test of speaking. If they had to choose one, all would have chosen the face-to-face test. Katja summarised the views of the interviewees on the deficiencies of the tape-mediated test well: there is no real interaction, the need to say anything is highly contrived as you are talking to a tape, there is less tendency to focus on the content of the message, and there are few means available for communicating the message other than lexical. However, all three also realised that the restrictions on test length in face-to-face tests were high, and that only few topics could ever be covered in them, which would reduce their value as representative samples of candidate speech.

4.4 Assessor perceptions: post-assessment questionnaires and discussions

Directly after the face-to-face test assessment, the assessors commented on the fairness of the two test modes and on their degree of confidence in the correctness of the rating they assigned. They did this through answering a short questionnaire (Appendix 6), and after having completed the double assessment, through evaluative discussions.

The assessors considered the tape-mediated test reasonably fair. They ascribed the fairness to the variety of situations, topics, and taped interlocutors involved. Compared with the few topics and interlocutors of the face-to-face

test, the concreteness of the variety in the tape-mediated test gave the assessors a sense of some sampling taking place. Furthermore, the number of tasks offered the candidates several opportunities to remedy eventual flaws in any one bit of their performance. Aspects which were seen to reduce the fairness of the tape-mediated test were the additional skill of imagination (for instance inventing rather than translating or reading from a map the way to a flower shop), specific test-taking skills demanded by the tasks such as reading cues in the test booklet while listening to the taped interlocutor, the anxiety caused by the impersonal setting, and the impossibility of negotiation of meaning. These aspects were considered to cause unfairness because they were seen to affect performanse although they were not directly related to language ability. The test was considered to yield a rather useful sample of language for assessment purposes, although it clearly did not include interaction. The assessors considered the tape-mediated test fairly demanding, since the responsibility of producing appropriate language was on the candidate alone, unsupported by interlocutors.

The face-to-face test was equally seen to be reasonably fair. The fairness was seen to build on the flexibility and dynamicity of the interaction and the relatively relaxed atmosphere, while the variability of the tasks to different candidates, the inclusion of only two topics, the invariance of interlocutors, and the personality factors influencing the assessment were seen to present threats to the fairness of the test. The assessors were highly aware of their own performance both as interlocutors and as assessors influencing the process and outcome of the test. They were particularly concerned about the influence of the interlocutor on the usefulness of the sample for assessment purposes. If the interlocutor ended the interview too early, or if she did not seem to pose challenging enough questions to the candidate, there was nothing the assessor could do about this afterwards. Such flaws in elicitation may have been more common in this pilot test than in a future operational test situation, because the interlocutors were not thoroughly familiar with the test as yet, and only some of the possible caveats with the test formats chosen had been discovered.

The assessors' confidence in the correctness of the rating they assigned was fairly good on the tape-mediated test: the sample was sufficiently long and varied, its structure and overall characteristics were predictable, and the assessing of it seemed quite straightforward. Similar assessor attitudes have been reported before by Stansfield (1991:202). Furthermore, the assessors'

confidence was boosted by the feeling that unlike face-to-face tests, an individual assessment is here made on an individual performance and in response to standard prompts.

When there were problems in the assessment, it was mostly due to characteristics which would have appeared in any form of performance. There sometimes appeared to be a difference between a candidate's performance on short turns of speech and his or her performance on the longer presentations task. The short turns were assessed turn by turn through giving points, while the longer presentations were assessed directly on the scale. This accentuated the difference, though it was not the only cause for it: some candidates really did well on the short turns of speech but did not reach the same standard in the extended performance or, more rarely, vice versa. The difficulty was deciding which overall assessment to give. There were also some cases where the candidate's anxiety was clearly audible in the performance, and this reduced the assessor's confidence in the assessment she gave.

An interesting observation by the assessors was that on the predictable tape-mediated test they had formed their own expectations for each answer. The expected answers tended to be short, factual, and commonplace. When a candidate surprised the assessor by a creative or imaginative answer, she was positively surprised and rewarded this with high points. Thus, the assessors felt that at least with the highest achievers, they were not only assessing language but also imagination. This aspect might merit attention in studies of what assessors do when making their ratings.

With regard to the face-to-face test, the assessors' confidence in the rating they assigned was probably affected by their relative unfamiliarity with the technique: they felt that double assessment was absolutely necessary to achieve reasonable confidence, and that the possibility to go back to the performance through video before giving the first rating was highly beneficial. Even with double assessment, however, difficulties arose when a candidate's language knowledge and a combination of personality factors and ability for use were in dissonance. The differences went both ways: there were those who were talkative and communicated well but were sometimes difficult to comprehend because of problems in grammar or vocabulary, as well as those who had the grammar and the vocabulary but whose presentation skills were less well developed and comprehending them was difficult because of that. The status of personality factors influencing the rating was a cause for concern. The

assessment guidelines as they stood were not helpful in solving the dilemma of assigning a final grade in these cases, a fact which certainly calls for amendment.

4.5 Summary and discussion of results

The research questions in Study One addressed the candidates' and the assessors' perceptions of the fairness of tape-mediated and face-to-face tests of speaking investigated, and the assessors' perception of the usefulness of each test for assessing proficiency in speaking. The candidates' perceptions will be dealt with first, followed by the assessors' perceptions. The discussion concludes with a consideration of the usability of the results.

The candidates' perceptions of fairness were closely related to their view on the naturalness of the discourse in the tests. In questionnaires and in post-test interviews, the candidates listed more points which made the tape-mediated test less natural and less fair than the face-to-face test. These seemed to result from two main features of the tape-mediated test: that the test situation did not adapt to the individual, and that the performance was entirely constituted by one actor, the participant.

The lack of room for variation in performing the tasks on the tape-mediated test can also be seen to have caused some of the negative reactions. This feature may have become particularly salient to the participants because of the contrast with the loosely structured face-to-face test. Other face-to-face tests, such as the ICC Certificate in English (as reported in Halvari 1994) or, to an extent, the *access:* test (O'Loughlin, 1996), regulate the structure of the face-to-face tests much more strictly, and in that case the candidates may have an equally negative reaction towards the constraining of communication in these tests as in tape-mediated ones.

Lack of experience with tape-mediated tests may have been associated with the candidates' impressions of unfairness of the test. However, some candidates who had previous experience of tape-mediated tests also reacted unfavourably to this test. Moreover, not all candidates reacted negatively to the tape-mediated test: 57% of them felt that the impression given by their performance on the tape-mediated test was at least reasonably fair. The candidates considered the best points of the tape-mediated test to be its wide

coverage of everyday situations and the need to react fast. Predictably, the test was more readily accepted by those who did well on it.

The candidates' perceptions of the face-to-face test were very favourable: 86% felt that their performance on it gave at least a reasonably fair impression of their ability to speak English. The positive reaction was most often ascribed to the feeling that the test consisted of normal interaction with real people. When the candidates felt less than satisfied, the most frequent reasons given were test anxiety, the shortness of the test, and the low number of topics covered.

The candidates also noted some features in the face-to-face interaction which were unnatural when compared to speaking in general. This particularly concerned turn-taking and the distribution of speaking time between the participants. They felt that since it was a test, everyone should have an equal share of speaking time, and everyone should have a chance to finish their speaking turns before a new one was started. They said that in normal conversation, they would not have been equally attentive to these aspects of communication. Compatibility of personalities between the candidate and the interviewer, and proficiency differences between candidates, were seen as possible threats to the fairness of the face-to-face test.

The assessors were reasonably happy with the fairness of both the tape-mediated and the face-to-face test, and for similar reasons as the candidates. Their opinion of the face-to-face test was more negative than the candidates', however. The difference can be explained by the different roles of the two groups in the process: the candidates only saw the tasks and their performance on them, while the assessors worked with the criteria and used the performances to complete their task, the assessment. Where the candidates were most concerned about the characteristics of test tasks, and the closeness of the behaviour in the test to non-test language use, the assessors were most concerned about what was being assessed in the performances, and how much of that was language ability and how much something else. This something else was creativity and imagination in the case of the tape-mediated test, and interactivity, personality and personal compatibility with interlocutors in the face-to-face test. Interactivity was viewed to be problematic, because it meant that an individual assessment had to be given for a performance which was not individual.

When the assessors considered the usefulness of the samples elicited by the two test for making a reliable rating, interactivity came up again. The tape-mediated test did not test interaction, and the assessors could not be sure that the assessment given on the basis of the tape-mediated performance reflected each candidate's ability to interact in spoken English. The individual assessment given, however, satisfactorily reflected the characteristics of individual performance. The sample was useful for making an assessment, but whether it was a truthful assessment of spoken interaction, the assessors were not entirely sure. The usefulness of the face-to-face sample depended partly on the skill of the interlocutors and the quality of the interaction during the test, which was observed by the assessors to be somewhat variable from one occasion to the next. This problem was undoubtedly compounded by the relatively early stage of development of the test, and could conceivably be alleviated later. The strength of the face-to-face test was that performances consisted of interaction, which was thus definitely measured in the test.

Neither the candidate nor the assessor questionnaires in Study One included a direct question of preference for either the tape-mediated test or the face-to-face one. The results of Study One on the candidates' part can nevertheless be seen to confirm previous studies where such a question had been asked (e.g. Shohamy et al. 1993, Stansfield 1991 and Wigglesworth and O'Loughlin 1993), the preference was in favour of live interaction. The assessors' preference was more ambiguous: while they recognized the artificiality of the tape-mediated test, they were troubled by the subjective elements of the face-to-face test, both in terms of assessment and in terms of having to give individual assessments for a joint performance.

The significance of Study One in terms of test development was that it provided one type of validation data, stakeholder perceptions. These data were valuable in the test development stage, when even large adjustments to the tests were still possible. The results indicated that each test had its own strengths in the eyes of the respondents, although there were ways in which both tests could be improved. In the case of the tape-mediated test, the suggestions for improvement were specific because feedback was gathered task by task. In the case of the face-to-face test, similar task-specific feedback was not sought. The area identified for possible improvement was on the ideational level and concerned the possibilities for constraining variability in the process of the test. Such use of stakeholder perceptions in both validation

and test development has been reported for instance in Zeidner and Bensoussan (1988), Kenyon and Stansfield (1991), Brown (1993), and Alderson (1988).

The connection between test development and validation is strong when participant perceptions are investigated, as is evident in the above report on the results of Study One. If validation is understood in the wide Messickian (e.g. 1988, 1989) sense of clarifying the interpretation of test scores and the justification for their use, this by no means undermines the use of participant perceptions as a stage in the validation exercise, as they provide an angle on the test and the scores that is difficult to gain through any other means.

4.6 Discussion of methods

The research questions in Study One were broad, and lengthy descriptive answers were provided on the basis of the data gathered. However, had the data collection instruments been better designed, some more informative comparisons could have been made. With larger numbers, some quantitative analyses could also be conducted once the instruments are well designed. It also appeared that Study One had attempted to address three interlinked areas of inquiry — post-test questionnaires, post-test interview procedures, and investigations of assessor processes — each of which could be specified further in its own right. Suggestions will be made for improved designs in all three areas.

In addition to mending outright flaws in the questionnaires of Study One, it was possible to use the results of this exploratory phase for specifying the working hypotheses and writing selected-response questions rather than open-ended ones. In order to serve the potential need to collapse categories into binary responses, the decision was made to provide four response categories: two positive, and two negative. These were labelled 'agree', 'tend to agree', 'tend to disagree', and 'disagree'. A category of 'no opinion' was included as well to avoid forcing the respondents to make a choice they were unwilling to make. Some of the questions were worded negatively in order to avoid inducing a set response.

The proposed questionnaire is presented in Appendix 7. The 14 questions cover perceptions of more tangible aspects of the test such as

comprehensibility of questions, sufficiency of answering times, and appropriacy of total test length, as well as less tangible aspects, namely test difficulty, validity, and authenticity. In addition, perceptions of own performance with respect to test anxiety, success, and enjoyment, are also investigated.

The main flaw in the post-test questionnaires in Study One was that each of them addressed different questions, while few parallel questions were asked. Even if Study One was exploratory, the two questionnaires should have been similar, since the aim was to compare the respondents' perceptions of the two tests. In the new proposal, this aspect is mended through making the two questionnaires as similar as possible. Only two questions are worded slightly differently: questions 7 and 12. Question 7 makes reference to the mode-specific restrictions: interacting with the tape, and interacting with one person on two topics. The questions were more comprehensible when the wording was as concrete as this, and thus the difference was felt to be justified. These shortcomings were found to be salient to the candidates in Study One, and were expected to be meaningful for other candidates as well. Question 12 named the mode of the test investigated.

The inclusion of each question was carefully considered in the proposal. The justifications for including each question, as well as working hypotheses for responses, are provided in Appendix 8. The hypotheses were formulated in order to ensure that the results would be analysable (see for instance Oppenheim 1992:61-62). The main tendencies expected are that positive reactions to one aspect of the test coincide with positive reactions to other aspects, and that the relationship between test performance and affective reactions will be complex. In addition to Study One, the working hypotheses stem from two previous studies on candidate perceptions: Scott (1986) and Fulcher (1996). These studies provided the impetus for investigating the factors of enjoyment and difficulty in addition to fairness.

The post-test interviews, as conducted in the pilot study, yielded in-depth information, which both helped the interpretation of the questionnaire responses and highlighted the individuality of each candidate's test experience. The match between the questionnaire and the interview responses was fairly close, but the difference between the quality of information was clear. Only the interview gave the candidates enough time and opportunities to reflect on their language use and test performance, and thus helped the present writer in

beginning to understand the construct of the test from the candidates' point of view.

It was equally clear, however, that the type of data yielded by the interviews was less clearly categorisable, and the group of interviewees would have to be considerably larger than three before any patterns could be discerned. This would require time and careful planning of the procedure and interview protocol, for instance with respect to how long time after the test the interviews should be conducted, how many times the performances should be reviewed, and who should be in charge of stopping the tape for comments. The data gathered would be rich and extensive. This would be a study on its own right, the goal being to describe the construction of the test by the candidates.

The assessor perceptions in Study One were gathered through a post-assessment questionnaire on one of the test modes, and through an evaluative discussion, loosely based on the questionnaire responses. The status of the assessor perceptions in Study One was difficult because the present writer was one of the assessors, and because there were only two assessors altogether. The information was interesting because it both supported some perceptions offered by the candidates and contrasted them as regards assessment. This was understandable because the roles and responsibilities of the candidates and assessors were different when assessment is concerned. The results were also highly specific to the current situation of test development and may not be generalisable to any other situation. Furthermore, the assessor viewpoints reported in Study One may be more rationalised and polished than assessor perceptions from a different design, both because the perceptions were gathered after the assessment had been completed and because the present writer was one of the informants.

A new proposal for investigating the concerns that the assessor attends to when making a rating on speaking is collecting both think-aloud data and conducting post-assessment interviews. The think-aloud data would provide information on which sections of candidate speech invoke reactions in the assessors, and the post-assessment interview would offer the assessors an opportunity to explain more thoroughly what they were attending to and why they judged the performances the way they did. The aspects of candidate performance that the assessors attended to, as well as the justifications for their assessments, would be the focus of the post-assessment interviews. The

interviews should be conducted immediately after the assessment. This would provide assistance for interpreting assessment results, and enable the investigation between assessor perceptions and properties of the performance, as well as assessor perceptions and scale descriptors.

5 STUDY TWO: TEST DISCOURSE

Study Two examined the two tests as discourse events. From the discourse point of view, tests of speaking can be defined as pre-planned language use situations in which a representative sample of candidate speech should be elicited. Study Two investigated the two aspects of this definition separately, firstly examining the tests as pre-planned language use situations, and secondly concentrating on the sample of speech elicited in the two tests.

The first analysis consisted of an examination of the tests as discourse environments. As the environment concerned was the test, which consisted of test tasks, this part of Study Two is called Task Analysis below. This is a more appropriate name for the part than test analysis, which normally refers to the numerical analysis of scores. The aim in Task Analysis was to help specify the ways in which the discourse environments of the tape-mediated and the face-to-face test differed from each other. The Task Analysis also had a practical aim, which was to help improve the test. The method used was an application of Bachman and Palmer's (1996) framework of Task Characteristics.

The second part of Study Two concentrated on the actual language samples elicited from the candidates in the two tests. This part of Study Two is called Performance Analysis below. The aim was to see what kinds of similarities and differences there were in the language samples elicited in the tests. The analyses included an initial characterisation of the discourse through simple counts of words and turns, an analysis of the functions elicited, and an automatic part-of-speech and lexical/phrasal analysis of language use.

The research questions in Study Two were:

2a. How do the two test modes compare as contexts for discourse elicitation?

2b. Which communicative functions are elicited by each mode?

2c. Which other features of performance than functions should be analysed in oral test discourse in order to account for similarities and differences between test modes?

As with Study One, the materials and methods used in Study Two will be reported first. The results are then reported analysis by analysis according to the conceptual division into task and performance analyses, while the discussion of the results is arranged to correspond to the research questions. Study Two conlcudes with a discussion of the methods used.

5.1 Methods used in Task Analysis

Task Analysis was conducted in Study Two because of the high likelihood that the task setting will influence the language elicited in the test the same way the language use context in a non-test situation will influence the language used in it. Previous studies which have analysed tests as discourse environments have used analysis of test specifications (Shohamy 1994), ethnographic investigation of test development (O'Loughlin, 1996), and analysis of the test instrument through Bachman's (1990, 1991) framework of Test Method Facets (Hoekje and Linnell 1994, Bachman et al., 1995). In Study Two, an approach based on the last method was adopted. Relying solely on comparing test specifications was unsuitable in Study Two because the two tests under investigation were based on essentially the same set of specifications, and an ethnographic investigation was impossible because the development data that were available were not collected systematically enough to allow a thorough analysis.

Task Analysis was based on an updated version of the Test Method Facets (TMF) framework (Bachman 1990, Bachman and Palmer 1996), now called framework of Task Characteristics (TC). The TC framework is intended for closely describing target language use situations and test tasks, with the purpose of serving test development. In the present study, the framework will be used for describing test tasks only. Like its predecessor, the TC framework has five groups of characteristics: environment, rubric, input, expected response, and relationship between input and expected response. There are several sub-categories under each main heading, offering a detailed grid for describing the test instrument and its use.

Bachman et al. (1995:100) point out that task analysis as a part of test content analysis serves three purposes: it provides useful information to prospective candidates, it can be used by test developers to check how well the intentions as stated in the test specifications have been realised, and it serves as a useful complement to analyses of test performance. The last two of these were the purposes to which TC analysis was put in Study Two. The TC framework as it appears in Bachman and Palmer (1996) and Bachman et al. (1995) was slightly modified to suit these purposes and the test context to which the system was applied.

Two previous studies which used TC analysis in describing tests were mentioned above. Hoekje and Linnell (1994) used the framework to aid their description of the three tests of speaking that they contrasted in their article: the SPEAK test, the OPI, and an in-house performance test. The writers apparently employed their knowledge of the three tests to assist them in identifying some of the interactional differences between the tests, and used some aspects of the TMF to make their point. Hoekje and Linnell did not state clearly what they used as data, nor did they specify who made the judgements and whether some kind of an instrument of analysis was used. The reporting of the results concentrated on interactional characteristics of the three tests, and no close linguistic analyses of performance were reported. Transcripts of one candidate's performance in the three modes were used as illustrative material. The article makes a case for considering authenticity in selecting a test for the assessment of the speaking skills of prospective International Teaching Assistants.

The Cambridge-TOEFL comparability study (Bachman et al. 1988, 1995) employed three of the five TC groups: testing environment, test rubric and test input, for analysing the similarities and differences between the two tests. Expected response and relationship between input and expected response were not investigated due to lack of time. The results of analysing the facets of the testing environment and test rubric were published in the 1988 report, while those for the facets of input were included in the 1995 report. The instruments of analysis used in the Cambridge-TOEFL comparability study evolved during the progress of the project. Unlike in Hoekje and Linnell's (1994) study, several experts were eventually employed to analyse the two tests. The experts used a very detailed rating instrument directly based on the TMF framework. However, only listening, reading, vocabulary and structure tests were

analysed. The analysis of tests of writing and speaking was mentioned as a fruitful direction for further research (Bachman et al., 1995:124), but due to lack of time this was not pursued.

The method employed in Task Analysis in the present study was a formalised version of Hoekje and Linnell (1994). This meant that the TC framework (Bachman and Palmer 1996:49-50) provided the descriptors used for describing the tests, but the tests were described rather than assessed numerically through a data collection instrument. All the aspects of the TC framework were covered in the description as far as was feasible. However, the description of the format of language in input and expected response was modified, because the classifications offered in Bachman et al. (1995) and Bachman and Palmer (1996) were not fully appropriate for analysing spoken language. Furthermore, the linguistic analysis of the test instruments in isolation from realised performances — an implicit assumption in the TC framework — appeared rather problematic particularly in the face-to-face test. It was therefore decided that the realised test discourse should be explored separately in Performance Analysis.

The Task Analysis in Study Two was conducted by the present writer alone. The resulting description thus represents a single observer's impression guided by the TC framework, and particularly where judgements are concerned, all the details of this view may not be shared by others. The reasons for not acquiring a second informant were that the process was long and time-consuming, that the data other than the transcripts were not collected systematically enough when the tests were administered to allow independent assessments of the conditions after the fact, and that the study was exploratory in the first place. Should the analysis prove useful, an instrument would be developed for further use on the basis of the experience gained in using the framework in this open-ended way.

The primary data for TC analysis is the test, and information on test administration and scoring procedures. Since the exact form of a face-to-face test only materialises during administration, transcripts of actual tests must be analysed. The decision was made to analyse both tests from transcripts to maintain comparability. For the TC analysis to be comprehensive, a "rich" transcript of each test including both all the material actually spoken and the material read silently by the candidates was created. Samples of the rich transcripts for the two tests are provided in Appendices 9 and 10. It was

possible to conduct the TC analysis of the tape-mediated test from one transcript, but because the exact form of the face-to-face test varies between administrations, five rich transcripts of the face-to-face test administrations (involving the full performances of ten candidates) were analysed.

5.2 Methods used in Performance Analysis

5.2.1 Materials

The analysis of the test performances was based on ten transcribed performances from both test modes. Random sampling was not attempted as the whole group of 37 candidates was not large to begin with. Instead, the sample of ten was chosen to represent the whole group of 37 participants on the one hand and the target group for the intermediate level test on the other.

A vague notion of "candidate type" was used in selecting different performances to represent the group, most clearly characterised by the amount of speech or willingness to talk, a bipolar scale of approach to speaking with planned at one end and spontaneous at the other, and some central auditory properties of the performances such as accent and rhythm. The speech of four of the ten candidates could be described as markedly planned, while the speech of three could be characterised as markedly spontaneous. The ten performances were fairly evenly distributed across an "accentedness"-scale, one end of which might be labelled 'heavily accented' and the other 'less clearly accented'. If the material had included advanced level performances, the high end might have been more clearly labelled 'indistinguishable accent'. With respect to rhythm, the scale ends could be labelled 'clearly un-English' and 'resembling English rhythm'. The aim was to include a range of performances within the scope of what could be expected on the intermediate level test, but the low end of the scale may have been overrepresented by the ten candidates whose performances were transcribed.

As few performances were of a very low level, only one or two of these were included in the sample to be transcribed. As the test was meant for the intermediate level, all performances clearly exceeding the test requirements were excluded. Furthermore, to make transcription easier, the performances of

both or all candidates in the face-to-face dyad or triad were included except in the case of the final candidate to make the count add up to ten.

There were nine women and one man in the group of ten candidates whose performances were transcribed. Four of the candidates were assigned the highest skill level for the Intermediate level test, level 5, for both tests. Four were assessed to be skill level 4 on both tests, and two received a 4 on the face-to-face test and a 3 on the tape-mediated test. In reporting the results, first-name pseudonyms will be used for the ten candidates.

The ten performances were transcribed into the CHILDES format (MacWhinney 1989). A rough word-level transcription was used, and in accordance with the CHILDES conventions, pauses were marked with one, two or occasionally three hashes depending on length (see example in Figure 5, page 91). The pause lengths were determined impressionistically by the present writer after a training period with a stopwatch, such that a single hash corresponded to a pause of up to approximately 0.5 seconds, a double hash to a pause of approximately 1.0 second, and a triple hash to a pause longer than 1.5 seconds. The measurements were not exact, however. Hesitations, false starts and isolated non-word-formed sounds were as a rule preceded by the ampersand character and thus not counted as words. The exception to this were backchanneling signals on the face-to-face test as well as the few cases where a non-word sound clearly fulfilled the function of a word both in terms of its place in the intonation pattern and in terms of utterance structure. These were counted as words. Contracted forms were transcribed and counted as one word.

5.2.2 Types of analysis conducted

Three types of analysis were conducted on the language of the test performances: an initial characterisation of the performances through counting words and turns; a functional analysis of the kinds of meanings that the candidates were conveying; and numerical comparisons of the actual word forms used by the candidates in performing each test.

The word and turn counts were made in order to get a quick overview of the data, and establish whether there were some salient differences between the modes on this level. The counts also served in investigating the degree of

variation between candidates within a mode. The counts were made automatically through the CHILDES program.

The functional analysis was conducted in order to investigate a working hypothesis that had guided test development, to the effect that the tape-mediated test would be more efficient in eliciting a large range of functions from the candidates. Interestingly, this hypothesis partially conflicted with the results of Shohamy's (1994) comparison of Hebrew SOPI and OPI, which concluded that for other than the lowest skill levels, the face-to-face version was more efficient in eliciting a large range of functions. The functional taxonomy used in the analysis was a development of the one developed for the National Certificates (Appendix 3). The performances were coded manually, while the CHILDES program was used to tally the frequences.

The actual forms of words used by the candidates on the two modes were investigated to identify key differences in language use between the two tests, and help raise hypotheses for further analysis. The analysis was based on parts of speech, in accordance with previous research (Shohamy 1994). The utility of semantic field analysis for investigating differences between the kinds of meanings expressed in the two tests was also explored in the present study. Two types of comparisons were made with the data: some between one candidate's performances on the two tests; and some between the ten candidates' performances on each test. This would help judge whether any of the differences found were likely to be associated with an individual speaker's performance, or whether it was likely that the difference was associated with the test modes. The analyses were run on ACAMRIT (Automatic Content Analysis of Marketing Research Interview Texts), a suite of programs for investigating spoken and written English developed at Lancaster University (see e.g. Thomas and Wilson 1995, Wilson and Leech 1993). The central function of the program is content analysis, but because it is versatile, including syntactic and semantic taggers, a lemmatisation program and a concordance and statistics package, it can be used as an aid for conversational or discourse analysis.

5.2.3 Variables

The variables investigated in the Performance Analysis were words, turns, language functions, and parts of speech. The choice of the word as the basic

variable was simply made because of its frequency and its interpretability. It should be noted, however, that word refers to the spoken word, such that contracted forms were counted as one word.

The unit of analysis above the word was chosen to be the turn. Of other alternatives, the sentence would have been a very difficult unit indeed to employ, because of the frequency of coordinating conjunctions in the performances. The division of the stream of talk into sentences would have had to be done by the transcriber. The clause would have been a slightly more plausible unit of analysis, but as there were many anacolutha and verbless constructions, even this would have led into frequent trouble over unit boundaries. The turn seemed an easier unit to define than the utterance because it was seen to be more strongly connected to the interplay of speakers than to the clause structure of what was said, though both of these criteria were eventually used in the definition. The core definition for the turn was "anything produced by one speaker in one stretch, uninterrupted by other speakers". On the tape-mediated test the turn was simply determined to be anything that a candidate produced within the time provided for answering each prompt. Each candidate's performance thus consisted of 31 turns. On the face-to-face test, the turn required some further specificatopm. In a few cases a candidate paused, often at a clause juncture, for a clearly noticable time and with a potentially final intonation contour, thus giving an opportunity for the interlocutor to step in. If nothing was forthcoming, however, the candidate went on to say something more. This was counted as two consecutive turns by the same speaker. The situation was thus the same as in the majority of cases where the interlocutor did produce something, minimally a backchanneling signal. Backchanneling signals and other comparable short turns were regularly transcribed as turns, leading to the result that very long turns were rare on the face-to-face test. Only when a backchanneling signal or a short turn entirely overlapped with a part of the main speaker's turn and appeared to make no intonational or structural difference to it was the main speaker's turn transcribed as one turn, and the backchanneling signals as interjections.

Functions, as used in the present study, refer to what the language user does with the language. The functional taxonomy used in the Performance Analysis was a development of the taxonomy for the Intermediate level certificate (see Chapter 3.2 and Appendix 3). There were six main categories: Giving and asking for factual information; Expressing one's point of view;

Expressing emotions and attitudes; Acting through words (performatives); Acting according to social norms and customs; and Communication strategies. Where the original original taxonomy had a single category of "asking" under the informative function, the analytical taxonomy distinguished between asking for factual information, asking about opinions, and asking about emotions. In addition, "continuing" was added as a category into the informative, opinion, emotion and performative functions. The developments had been prompted by trial coding.

Parts of speech were chosen as a basis of the analysis of word forms because some starting point was needed for this explorative part of the study, and because parts of speech had been analysed before, and some significant differences between tape-mediated and face-to-face tests had been found (see Shohamy 1994).

5.2.4 Coding and analysis

The ten performances analysed in Study Two were transcribed by the present writer. This process involved the division of the sound stream into words and turns. The transcripts were checked three times, both from audio tape and from video tape, but only by the present writer. The rationale for dividing the sound stream into turns was described above. After the transcription, the frequencies of words and numbers of turns were counted automatically by CHILDES.

The most impressionistic coding procedure in the Performance Analysis was the functional analysis. Here, a double coding system was used, such that the performances of three candidates were coded independently by two coders and disagreements negotiated. The coders disagreed in approximately 15% of the utterances coded, and disagreements usually concerned more open-ended categories such as stating a fact versus giving an opinion. The functional coding of the remaining seven performances was performed by one coder only, though it was checked and accepted virtually unchanged by the second coder. The most difficult aspect of the functional coding according to both coders was deciding whether a second, third or fourth aspect possibly conveyed in a turn was a main or a subsidiary meaning unit.

In the functional analysis, an attempt was made to account for the main units of meaning in the discourse, not to exhaustively describe all the possible

shades of meaning that each utterance might have. The content and form of the turn under analysis, as well as the discourse context in which it was spoken, were used as basis for coding. The context preceding the turn was used more extensively than the context following it. Thus, the main attention was on the analyst's perception of illocutionary intent rather than perlocutionary effect, but the effect was sometimes needed in the coding, particularly when deciding how backchanneling signals ought to be coded.

Due to the multi-functional nature of language it was initially expected that some turns could have multiple main functions, but not more than two for a short utterance, three for a longer one. Multiple functions were expected to occur in two cases; either when a turn, however short, genuinely seemed to serve two purposes and it would be difficult for the analyst to decide which was more important, or when the turn contained more than one units of meaning which would serve different purposes in the discourse. The expectations held well for the face-to-face test, but less well for the tape-mediated one, chiefly perhaps because the non-reciprocity of the situation precluded the use of backchanneling signals. Since a turn was defined as whatever one speaker produced uninterrupted by other speakers, and interruptions were not forthcoming in the tape-mediated test, the candidates' turns became long, provided they were willing to answer with more than a few words. Furthermore, in most tasks the cues and instructions contained a number of points which the candidates were to express in each turn. Some of the replies in the second task on the tape-mediated test, Reacting in situations, easily contained five or six different functions.

The part-of-speech analysis of the transcribed performances was done automatically by ACAMRIT and checked by the present writer. One of the advantages of the program is that in addition to an extensive single-word lexicon, it includes a multi-word lexicon with phrasal verbs, multi-word proper nouns, genuine 'idioms', and other multi-word units such as compound connectors, for instance "as well as". Based on the experiences using ACAMRIT on two extensive projects where spoken English was analysed, Thomas and Wilson (1995:106) state that ACAMRIT is accurate and fast in revealing certain important aspects of language use in the material being analysed. However, as they also note, "some aspects of ... discourse (politeness, indirectness, topic control, interruptions, etc) could not be identified using ACAMRIT." The use of the program is thus best combined with more

detailed analyses of language use, such as conversation analysis or discourse analysis. According to Thomas and Wilson (1995:107), the analyses carried out by ACAMRIT are theory-neutral in being useful both for raising hypotheses and for confirming them.

The parts of the program utilised in Study Two were the stochastic part-of-speech tagger Claws (Garside, Leech and Sampson 1987), the semantic field category tagger SEMTAG (see Thomas and Wilson 1995:96) and the concordance and statistics package SEMSTAT. The aim was to identify key differences in language use between the two tests, and help raise hypotheses for further analysis. The Claws analysis would help reveal syntactic differences in language use between the two modes, and the SEMTAG analysis would investigate whether there were any interesting differences between the meanings conveyed in the two tests. The comparisons were conducted automatically through the SEMSTAT program, which takes into account differences in the size of the texts being compared.

The analyses were conducted on the language produced by each individual speaker only. For instance, the file comparing Eeva's performances on the tape-mediated and the face-to-face test contained this candidate's turns on the tape-mediated and the face-to-face test only, and none of her interlocutors' turns or the instructions she received on either test. The file comparing the tape-mediated test performances only contained the turns produced by the ten candidates on the test, nothing else. The language of the interlocutors was not investigated because the data set was too small and therefore susceptible to chance variation.

5.3 Results of Task Analysis

The method employed in the Task Analysis was a version of the framework of Task Characteristics proposed by Bachman and Palmer (1996). The results of the analysis are described below. A summary of the results is given in tabular form in Appendix 11.

5.3.1 Characteristics of the testing environment

The characteristics of the testing environment concern the nature of the physical and temporal contexts of testing, and candidate familiarity with these.

In the present study, the physical contexts of testing were different. One of the tests was administered in the language laboratory in individual booths with each candidate wearing a set of headphones, so the performances were individual even though a group of candidates always took the test simultaneously. The other test was conducted in a classroom, first with a pair or a group of three candidates first interacting with each other, and then with each candidate interacting individually with the interlocutor. The tape-mediated test was audio-taped while the face-to-face test was recorded on both audio and video. The candidates were aware that this meant that paralinguistic communication would not count in the assessment on the tape-mediated test. The role of the personnel involved in administering the tape-mediated test was that of an invigilator, while in the face-to-face test the role was that of an interlocutor. Both tests were given in the evening, after a full day of work or study. The time of testing was not particularly conducive to good performance.

Candidate familiarity with the two settings varied, in that some had visited the specific location where the tests were given while others had not. All had been in a classroom but few under oral test conditions; all but five had taken a test in a language laboratory before. Three candidates knew their face-to-face test interlocutor, and all but ten knew the peers they were interacting with in the peer discussion task. The candidates are likely to have reacted differently to various aspects of the testing environment, but these were not investigated in a systematic way. In the post-test questionnaires, more candidates expressed negative feelings towards the tape-mediated test environment than the face-to-face one.

5.3.2 Characteristics of the test rubric

The characteristics of the test rubric (Bachman and Palmer, 1996) are divided into four sub-groups: instructions, structure, duration/time allotment, and scoring method. There were similarities and differences between the two tests in each of them.

Both tests contained both written and spoken instructions. The spoken instructions were given in the target language in both tests, while the written ones differed in terms of language: target language in the face-to-face test and

native language in the tape-mediated one. This was done in the interest of guaranteeing that all candidates understood what they had to do in the tape-mediated test where they did not have an opportunity to remedy any difficulties in comprehension. The instructions were fairly specific and different for each task in both tests. In the tape-mediated test, all the instructions were given task by task. In the face-to-face test, the written instructions were given in one bulk before the test, while the spoken instructions were given task by task. The different sequencing of the instructions, as well as the variation in language, resulted in clear differences in the nature and the structure of the test discourse.

The tape-mediated test consisted of a warm-up task and three test tasks, while the face-to-face test contained a warm-up section and two test tasks. The parts were salient in each test, and transitions between the tasks clearly marked — in the tape-mediated test through instructions, in the face-to-face test through instructions and people's movement when changing from peer discussion to individual interview.

The tasks in the tape-mediated test were in an order of increasing difficulty. This was evident from a clear progression of task complexity, an increase in length of expected response, and a gradual introduction of more abstract topics. Though there was no change in task complexity in the the prepared interlocutor prompts for the face-to-face test, there was a similar orientation from more concrete and perhaps narrow questions in the beginning towards more abstract and open-ended ones in later stages of the test. The transcripts for the ten test performances showed that this was evident in the interlocutor's conduct as well. The progression of the face-to-face test, however, was partially dependent on the candidates, as a lot of the elicitation was done through backchanneling signals which mostly just encouraged the candidate to continue.

Neither test provided clear information to the candidates on the relative importance of the tasks. The candidates are thus likely to have assumed equal weighting for all tasks. This was true for the face-to-face test but not for the tape-mediated one, where the last task carried more weight than the first two. This point should be remedied in later versions of the test.

Duration and time allotment clearly distinguished between the modes. From start to finish the tape-mediated test was thirty minutes long while the length of the face-to-face test for each candidate varied between nine and

twenty-two minutes. The candidates felt that the length of the tape-mediated test was appropriate, while many of them were surprised by the shortness of the face-to-face test. The live administration of the face-to-face test made it a possible power test, whereas the pre-programmed nature of the tape-mediated test meant that not only ability to answer but also speed in answering would be measured in some cases, particularly as some of the task types involved regulating the maximum length of individual turns of speech. Attempts had been made to make the response times suit the majority of the candidates, but there would always be some candidates for whom the pauses were not long enough. In the post-test interviews it appeared that the tape-mediated test was perceived as speeded even by those who had time to answer, an observation which was not evident from the performances alone. The face-to-face test was not perceived as speeded by any of the candidates.

On a general level, the criteria for correctness on the two tests were the same: comprehensibility and approximate appropriacy to situation. On closer observation, however, differences can also be detected. The first two tasks in the tape-mediated test were scored turn by turn by giving points, the sum of which was then converted onto the scale, and only the extended speaking task was assessed directly on the scale. The face-to-face test performances were assessed through the scale only. Furthermore, for the individual turns scored through points, the tasks were rather specific and exactly the same for all candidates, which meant that the assessors were able to formulate fairly clear expectations of what the candidates were going to say. In the post-assessment discussion the assessors reported that they felt this had an effect on their marking. Similar expectations could not be formed for the face-to-face test performances because the tasks and prompts were much more open-ended. Here, however, the factual content of the candidates' speech could have had an influence on the marking according to the assessors. The criteria and procedures followed in the assessment were not explicit to the candidates in either test.

Table 4 illustrates some of the most salient differences in instructions between the two tests. These are the existence of an initial chunk of instructions in the face-to-face test, and the use of two languages and the pervasiveness of instructions in the tape-mediated test. Within each task, the procedure in the tape-mediated test was explained both in written Finnish and

Table 4 Size (number of words) and types of instructions in the two tests

	Tape-mediated test			Face-to-face test	
Initial instructions				Written English	Procedure211
Warm-up task - t-m[1]: reading aloud - f-t-f[2]: introducing oneself	Written Finnish Spoken English	Procedure Transition Introduction Procedure (78+8) Transition	53 16 43 86 28	Written English Spoken English	0 0
Task 1 - t-m: simulated conversations - f-t-f: peer discussion	Written Finnish Spoken English	Procedure Transition Procedure (64+2+2+2) Transition	45 16 70 34	Spoken English beginning: procedure end: transition and logistics il1 $\bar{x}$ il2 $\bar{x}$	 149 77
Task 2 - t-m: reacting in situations - f-t-f: interview	Written Finnish Spoken English	Procedure Transition Procedure (116+ 10*16) Transition	91 6 276 28	Spoken English end: closing and logistics il1 $\bar{x}$ il2 $\bar{x}$	 51 6
Task 3 - t-m: talking about views and opinions	Written Finnish Spoken English	Procedure (123+20) Closing Procedure (69+52+36+52+2) Closing	143 5 210 15		

[1] t-m = tape-mediated, [2] f-t-f = face-to-face, [3] il1, il2 = interlocutors 1 and 2

in spoken English. The sums in parentheses under Procedure in spoken English illustrate the fact that the instructions interspersed the performance, telling the candidates at each juncture what to do next. The transitions between the tasks were also announced both in written Finnish and in spoken English. In contrast, the face-to-face test employed few instructions after the introductory chunk, indeed none during the warm-up task. The differences in means between the two interlocutors may indicate different styles, but can be at least partially explained by the fact that four of the six candidates that interlocutor 1 interviewed did not perform particularly well and needed more support and explanation of procedures. Interlocutor 2's low average for words in closing indicates an abruptness which may not be entirely desirable.

5.3.3 Characteristics of input

The notion of 'input' in the case of tests of speaking requires careful consideration. Bachman and Palmer (1996:52) define test input as "the material contained in a given ... test task, which the ... test takers are expected to process in some way and to which they are expected to respond". 'Input' is

one of three categories for characterising language in tests, the other two being instructions and responses. Input comes in the form of items, which require limited production, or prompts, which require extended production. As for length, input can be restricted or extended, thus requiring varying amounts of interpretation. A further distinction suggested by Bachman (personal communication) can be made between verbal input and input that consists of contextual information. The last distinction helps in deciding which sections of language in the test are input and which are instructions in tasks which include simulations and which thus must include descriptions of context. Ultimately both the instructions and the input need to be comprehended by the candidates, so fine distinctions may not be necessary.

The three categories of instruction, input and response are relatively clear in written tests. Specifically regarding input, the advantage of written tests is that the input is the same for all candidates, it is formulated by the test developers well before the occasion of the test, and it can be investigated in isolation from actual performances. The invariability of the input, its plannedness and its isolation from responses also applies in the case of tape-mediated tests of speaking, but in face-to-face tests the exact form of the input is only set during the test administration. Input can thus only be investigated from transcripts, and as the input is also at least potentially variable from one administration to another, several administrations need to be investigated in order to characterise the input in a face-to-face test. Furthermore, if the test contains peer interaction, some of the input will be from them, and what is input for one candidate is another candidate's performance. The linguistic analysis of input thus overlaps with the analysis of actual responses.

The conceptual problems in the analysis of the characteristics of input as it appears in the TC framework result from the implicit assumption that test input can be analysed before the test is administered. The idea of needing to know and control the kinds of language that the candidates are going to meet in the test is understandable, and such definitions are to an extent desirable. However, in the case of the face-to-face test, detailed analysis of the format of input before the test is administered is impossible. The fact that the analysis of input is possible for one of the two tests but not the other is a clear indication of the differences in input between these two test modes.

As some description of the input in the two tests was deemed desirable in the present study, the definition of 'input' was modified to make the analysis

possible. Thus, input is defined to be test developer- or interlocutor-originated language intended to elicit performance from the participants. The analysis of the linguistic characteristics of language produced by the candidates in the test was considered to be part of Performance Analysis.

In the TC framework (Bachman and Palmer 1996), the characterisation of input is divided into format of input and nature of language in input. The framework and its previous applications to analysing tests offered considerable support to analysing the format of input, but as far as the nature of the language in the input was concerned, some of the old categories were more difficult to apply. The Cambridge-TOEFL comparability study only focused on written tests, and the categories used in analysing organizational characteristics of the input language, i.e. sentence structure, cohesion and rhetorical organization, were naturally suited for analysing written language. These were less appropriate for analysing spoken input.

The characterisation of the nature of language of input in the Cambridge-TOEFL comparability study included two types of descriptors, counts and judgements. In the present study, only the judgements were included. This was because it was not easy to decide what to focus on and what to count when investigating the language spoken by one participant in spoken interaction, as the case would be if only the input and not the whole interaction is analysed. Units appropriate to be counted in the present data were sought in Performance Analysis.

As for format of input, there were both similarities and differences between the two tests. Most of the input in both tests was spoken, but both also provided some written input. In the tape-mediated test this concerned the descriptions of the situations where the candidates supposedly were, the discourse structure and cues in the simulated conversations, and the support questions to the two topics in the extended speaking task. In the face-to-face test this concerned the three topics from which the peers had to choose one, and the support questions to their chosen topic. This input was given in writing because it was thought that it would be easier and faster to process in that form. As evident from Table 5, there was more input in written form in the tape-mediated than in the face-to-face test.

Neither test contained figures or pictures, but the visual channel was naturally used in communicating on the face-to-face test as the interlocutors shared the same physical space. All of the input in the face-to-face test was in

the target language, while on the tape-mediated test some of the written input was given in the mother tongue. The spoken input in both tests was in the target language. The difference regarding spoken input, apart from the taped versus live distinction, was that the input was spoken by native speakers in the tape-mediated test while in the face-to-face test all the spoken input was produced by non-native speaker interlocutors.

The length of each individual turn in the spoken input varied in both tests, between a few words and three or four clauses in the tape-mediated test and between single-syllable backchanneling signals and multi-clause turns in the face-to-face test. The input was in the form of prompts rather than items in both tests, and extended responses were usually expected. Two tasks in the tape-mediated test contained written input in the form of contextual information. This type of input was completely absent in the face-to-face test because no role plays or other simulations were employed. Furthermore, the constant changes between the native and target language marked the discourse in the tape-mediated test, distinguishing it clearly from the entirely monolingual face-to-face test.

Table 5 summarises some characteristics of the format of input in the two tests. A notable feature of the tests, and consequently of the table below, was that different tasks within each test contained different kinds of input. For instance, where Task 1 in the tape-mediated test built heavily on spoken input in the target language, Task 2 mostly contained written input in the native language, and Task 3 written input in the target language. In the face-to-face test, written input was only used in Task 1. The task also contained some spoken input by the interlocutor, while the performance was a spoken interaction between candidates. Task 2 was an individual interview, where the input was spoken by the interlocutor.

The clearest differences between the two modes visible from Table 5 are that the tape-mediated test provides some input in the native language while the face-to-face one does not, and that the figures for the spoken input in the face-to-face test are means, since this input is variable between administrations. Furthermore, separate means for the two interlocutors are reported because their language use was slightly different. When looking at Table 5 it has to be remembered that this table only covers characteristics of input, not instructions.

Table 5 Amount and types of input in the two tests

	Tape-mediated test			Face-to-face test		
Warm-up task - t-m[1]: reading aloud - f-t-f[2]: introducing oneself	Words written	Finnish	17	Words written	Finnish	0
		English	111		English	0
	Words spoken	English	0	Words spoken	English / il1[3] $\bar{x}$	33
					English / il2[3] $\bar{x}$	74
Task 1 - t-m: simulated conversations - f-t-f: peer discussion	Words written	Finnish	85	Words written	Finnish	0
		English	37		English	60-99
	Words spoken	English	326	Words spoken	English / il1 $\bar{x}$	33
					English / il2 $\bar{x}$	72
Task 2 - t-m: reacting in situations - f-t-f: interview	Words written	Finnish	256	Words written	Finnish	0
		English	0		English	0
	Words spoken	English	0	Words spoken	English / il1 $\bar{x}$	133
					English / il2 $\bar{x}$	144
Task 3 - t-m: talking about views and opinions	Words written	Finnish	0			
		English	66			
	Words spoken	English	0			

[1] t-m = tape-mediated, [2] f-t-f = face-to-face, [3] il1, il2 = interlocutors 1 and 2

The linguistic characteristics of the input were assessed because they are likely to influence the content and form of language elicited, i.e. performance. Bachman and Palmer (1996:53) suggest describing grammatical (vocabulary, morphology, syntax, phonology and graphology) and textual characteristics (cohesion, rhetorical and conversational organisation), functional and sociolinguistic characteristics, and topical characteristics of the input. These will be briefly covered below. Where relevant, separate descriptions are provided for written and spoken input, and for input in the native and the target language.

The written input in the target language was very similar in both tests: it consisted of topic titles and four to eight questions or statements which were intended to help the candidates in thinking of something to say on the topics. The present researcher's subjective assessment of the level of vocabulary in the written input was that some fairly difficult words for intermediate level learners were included in each prompt, resulting in a difficulty assessment of 3 on a 4-level scale for both tests. The idea of using subjective assessments of difficulty in the present study was taken from the Fortus, Coriat and Fund (1995) study where such assessments had been found to work in comparing estimated and realised levels of difficulty in tests of reading comprehension.

During the development of the tests investigated in the present study, attempts had been made to use vocabulary that the candidates would be able to

comprehend at least in the context where the words were being used. As it was possible to clarify uncertainties in the face-to-face test, some vocabulary items which could cause problems to some candidates had been included there and the interlocutors had been sensitised to making sure that the candidates understood them. In the 37 candidates' performances, problems with lexical items in the input occurred three times in the face-to-face test and once in the tape-mediated test. The problems in the face-to-face tests were clarified during the tests.

As for syntax in the written input in the face-to-face test, sentence length varied between three and 25 words, and each set of questions contained at least one complex sentence, usually one with an embedded question. However, for the intermediate level, the syntax should not present problems, resulting in a difficulty estimate of 2 on a 4-level scale. The syntactic structures in the input did not appear to cause any comprehension problems to the participants.

Only the tape-mediated test contained written input in the native language. This concerned Tasks 1 and 2, and consisted of the descriptions of the situations where the candidates supposedly were. In Task 1 of the tape-mediated test, Simulated conversations, all cues were given in the native language. The contextual descriptions were written out in full sentences, while the cues for the simulated conversations were very short, two to five words long. They had been designed to be comprehended at a glance, and were not written in full sentences. Since the input was in the native language, comprehensibility of the input was not as critical as in the case of input in the target language. However, it was possible that the lexical selections in the native language input made the candidates search for close correspondences in the target language in their response, thus causing a clear difference in the type of language use required in the two modes.

The syntax of the spoken input in the two tests differed slightly in accordance with the differences in format. In getting the interaction on a topic started, the interlocutors on the face-to-face test used multi-clause turns similar to all the spoken input in the tape-mediated test, which consisted of the prepared turns in the simulated interactions. The only real syntactic difference between the modes in these longer turns of input was that in the case of the face-to-face test there were some rephrasings and anacolutha, which were absent from the simulated interactions on the tape-mediated test. Once the candidates had gotten started on a topic, the spoken input by the interlocutors

on the face-to-face test largely consisted of single-word backchanneling signals and very simple questions or statements, such as *What about you?* or *Ten per cent.* The latter was a repetition of the figure that the candidate had given, and as intended, it caused the candidate to specify the information and motivate the argument. This type of elicitation devices did not occur on the tape-mediated test, because making them fit the candidates' responses would have been impossible.

The overall difficulty estimate of the syntax in the spoken input in both tests was 1 on a 4-level scale, which means that the syntax was very simple. The spoken input consisted mostly of simple clauses, and the few embedded clauses that were used were appropriately signalled with pauses, so that the candidates only had to process the input one clause at a time. The syntactic structures in the spoken input did not appear to cause any comprehension problems to the candidates.

The textual characteristics of the input in the two tests were different because of the different task structures and because of the possibility for real-time interaction in one but not the other. The tape-mediated test consisted of three tasks, the first containing three simulated conversations, the second consisting of ten individual turns of speech within the framework of three situations, and the third requiring mini-presentations on two topics. There was some cohesion in the discourse within each task, but even so the rhetorical structure of the input was fragmented. The face-to-face test, on the other hand, had two tasks, each with a single topic. The shared physical and temporal context as well as the interactive nature of the situation provided a firm foundation for creating cohesion. The discourse structures of peer discussion and conversational interview were collaboratively developed by the participants, and the textual characteristics of the input contributed to the creation of the discourse. The interlocutor had the most power to control the direction and stages of the discourse, but fragmentation in the sense it was present in the tape-mediated test was only there at the transition between the two tasks.

Sociolinguistically, there was more variation in the tape-mediated test than the face-to-face one. There were more interlocutors in various (imagined) social roles, and both British and American accents were included. However, there was very little personal interaction, because it was felt that simulating this in the laboratory context would be difficult. The intimate register was

thus not included in the input on the tape-mediated test. In the face-to-face test each candidate had two or three interlocutors, all of whom were non-native speakers of English, and there was only one context, that of the examination room. The roles were set; candidate and interlocutor. Because of the test context the communication was not intimate nor would it have been entirely appropriate for it to be so, but the social dimension of interaction — the maintenance of acceptable social relations between the interactants — was there in the live interaction of the face-to-face test while it was missing in the tape-mediated test. The linguistic features of the spoken input did not appear to reflect the presence or absence of relationship-building as such, but the post-test interviewees all paid attention to this distinction between the two tests. This aspect of language use was possibly shown through extra-linguistic means as well as through backchanneling signals and the timing of reactions, turns interlocking appropriately to construct a conversation.

The TC framework employs Halliday's (1976) categorisation of functions into ideational, manipulative, heuristic and imaginative ones. The Cambridge-TOEFL comparability study (Bachman et al. 1995:123), however, showed that there was very little variance in the functional characteristics of those two tests when investigated through this system. It was suggested that this could be because of the restricted range of language sampled in the Cambridge tests and the TOEFL. While this may well be true, it is also fairly understandable, as great care is usually taken in test preparation against possible bias as well as against testing variables other than language ability. Control of manipulative functions could easily be seen to measure features of personality, and control of imaginative functions creativity as well as aspects of language ability. Situations where heuristic functions would be used naturally could well be difficult to simulate under test conditions.

The functional analysis in Study Two did not employ Hallidayan functions but used a functional categorisation specifically developed for the test where the data was taken from. It was considered that the functional analysis was more appropriately conducted on the whole of the test interaction and not just the input. The results of the functional analysis are reported in Chapter 5.4.2.

The topical characteristics of input in the two tests were somewhat different. The tape-mediated test contained simulations of discussions in altogether six contexts, as well as mini-presentations on two topics; the face-to-

face test discussion was structured around two discussion topics. In a sense, the tape-mediated test thus included more topics, while the face-to-face test dealt with fewer topics but in greater depth. On both tests the test developer and the interlocutor were ultimately in control of the choice and development of the topics. The role of the candidates was to make the best of the topics they had been given.

5.3.4 Characteristics of expected response

The format of the expected response was spoken target language on both tests, but some differences in format were inevitable due to differences in the nature of the tests. The tape-mediated test responses were not freely constructed but guided: the expected content of most of the responses was given in two of the three tasks. Furthermore, as the performances were recorded on audio cassette, all the characteristics to be rated had to be vocal. The face-to-face test responses were more freely constructed by the candidates especially regarding content; only the topic was given by the interlocutor. The performances were recorded on video cassette, thus enabling the use of extralinguistic as well as linguistic means of communication.

The expectations for the linguistic characteristics of the responses corresponded with the types of tasks employed in the two tests. Both short and extended turns were expected as responses on the tape-mediated test, varying from single utterances to mini-presentations up to two minutes long. The face-to-face test responses were expected to be lengthy but not to be given without support from the interlocutors, minimally in the form of backchanneling signals. Willingness to talk was expected on both tests. Lexically, the minimum expectation was ability to discuss everyday topics, but any evidence of extended knowledge of vocabulary was welcomed and thought possible within the constraints of both tests. The candidates were also expected to produce grammatically and rhetorically comprehensible language which would be approximately appropriate for the situations that arose. It was thought that this posed higher expectations on the tape-mediated test since various situations were included where the candidates were expected to take on different roles. There were no strictly formal situations in either of the tests, but the tape-mediated test included some semi-formal situations such as representing one's place of work on social occasions during a visit abroad. On

the face-to-face test the candidates represented themselves, and the tone of the conversations was that of friendly acquaintances.

Both tests placed some restrictions on the candidates' responses, the tape-mediated test more so than the face-to-face one. On many tasks in the tape-mediated test, the content and/or the main function of the response was given in the task, and in one task — Simulated Conversations — the turn structure was pre-determined. Turn length was always pre-determined, but combined with the expectation of willingness to perform, it was considered during the planning of the test that the response times given on the tape-mediated test would be sufficient for most candidates. On the face-to-face test, the selection of the discussion topic was restricted, as was the overall length of the test, but otherwise the discussion was considered free.

The characteristics of the expected response that can be specified for each test, and the effects that the expectations have on the assessment of the actual responses, were different in the two tests. The tape-mediated test offered a high degree of control in elicitation, and thus the linguistic characteristics of the expected response were more predictable in the tape-mediated test than in the face-to-face one. On the one hand this meant that there might be less variation between candidates on certain features of performance on the tape-mediated test. On the other, it could result in assessors forming very specific expectations as to the content and form of responses. They might then mark candidates down for not using at least some of the expected phrases or functions while marking them up for linguistic creativity in expressing the expected content. The face-to-face test did not offer as high a degree of control over elicitation proper to the test situation itself, and thus similar expectations could not be formed.

5.3.5 Characteristics of relationship between input and expected response

In the TC framework the relationship between input and expected response is characterised in terms of reactivity, scope of relationship, and directness of relationship. Reactivity, according to Bachman and Palmer (1996:55), is the extent to which the input or response directly affect subsequent input and responses. In this respect the difference between the two modes is clear: In the face-to-face test the input and responses affect subsequent input and responses

in a very direct fashion whereas in the tape-mediated test the input affects the response but not vice versa.

As for scope of relationship, which refers to the amount or range of input that must be processed in order for the candidate to respond as expected (Bachman and Palmer, 1996:55), the tests can be characterised as fairly similar. The length of each individual section of input was more easily controlled in the tape-mediated test, but in neither test were the sections of input very long. The transcripts showed that the longest continuous sections of input were the written support questions for Task 3 on the tape-mediated test and Task 1 on the face-to-face test. Some of the peers' turns of speech in Task 1 on the face-to-face test were also several clauses long.

Directness of relationship is defined by Bachman and Palmer (1996:56) as the degree to which the response deals primarily with the information in the input or whether candidates have to rely on information in the context or their own topical knowledge. In both tests, as in most tests where language production is involved, the candidates had to rely on their own topical knowledge. This was more pronounced in the face-to-face test; in the tape-mediated test some responses were guided.

5.4 Results of Performance Analysis

5.4.1 Initial characterisation of performances

To give an initial characterisation of the performances in the two modes, Table 6 reports some word counts. The statistical significance of the differences was estimated through a two-tailed t-test. As there were ten pairs of observations, the values have nine degrees of freedom. The tape-mediated test clearly elicited more language, the mean number of words spoken by the candidates was 989 whereas for the face-to-face test it was 578. The difference is significant at the .001 level. The explanation for the difference is that the tape-mediated test was longer than the face-to-face one: thirty minutes as opposed to a mean of slightly more than fifteen minutes. When words per test minute values were counted, the differences between the mean values in the two modes were not significant. This applied for words-per-test-minute values calculated both for the total test time in each mode and for the time that each individual candidate

Table 6 Comparison of the number of words spoken on each test

	tape-mediated		face-to-face		t	p
	mean	std	mean	std		
Words spoken	989	312	578	175	6.20	.000
Words per test minute / total test time	33	10.4	39	5.8	-1.64	.135
Words per test minute / modified	65.9	20.8	64.6	12	.301	.770
Turns spoken	31	0	45.8	17.6	2.65	.026
Words per turn*	23.6	6.8	13.1	2.5	5.41	.000

* The last task of the tape-mediated test, the two-minute presentations, was excluded from this count.

approximately had for speaking. The mean values for the tape-mediated test were slightly lower, but regarding individual candidates the trend was not consistent.

The number of speaking turns for each candidate on the tape-mediated test was invariably 31, while on the face-to-face test the mean number of turns spoken by the candidates was 46. The difference becomes all the more significant when remembering that the tape-mediated test was almost twice as long as the face-to-face one. This difference was significant at the .05 level. The two last turns on the tape-mediated test were two-minute mini-presentations, and in order to reach somewhat comparable results for counting the mean number of words per turn, these two were excluded from the calculations. Nevertheless, the difference between the mean number of words per turn on the two modes was 10.5 words. This difference was significant at the .001 level.

The two indicators mentioned above, the number and length of speaking turns, express in numerical form the difference that is immediately observable when looking at performances from two tests. Particularly in the case of the peer discussion, the face-to-face test discourse seems to consist of short turns with frequent changes of speaker. The interlocutors tend to show they are listening and indicate their thoughts at least through short verbal comments. This sort of interplay is impossible to imitate on a tape-mediated test simply because there is no immediate listener. Thus, the turns become much longer. This may well be one of the discourse features which make the candidates experience the tape-mediated test as different and un-life-like. The extracts in Figure 5 illustrate the difference.

In the examples in Figure 5, both tests are represented through their "most interactional" tasks. It is quite easy to imagine how the tape-mediated test example, the exchange at the shop, would have differed in discourse

Figure 5 Extracts of the two tests illustrating differences in turn structure

peer discussion triad on the face-to-face test [on why Sweden takes more refugees than Finland]		simulated discussion on the tape-mediated test, cued in the test booklet [returning a jacket to a shop for a friend]	
PIRJO:	but the economical and the political system is different.	TAPE:	do you have a receipt with you?
ANJA:	yes the system is different and the nature # the nature of the people # +...	PIRJO:	&ehh I'm afraid I don't have it # my friend have # lost it.
PIRJO:	yea.	TAPE:	hmmm ## I'm not sure I remember your friend ## what does she look like ## what time was she here?
ANJA:	+, is also different # we want to # live <alone> [>1] # and &aa <yes> [>2] how could we have # on our own and +...	PIRJO:	&mm I guess it was around four o'clock # four p m and she's ## she's tall # and redhead.
MARJA:	<yea> [<1].	TAPE:	oh yes I remember her ## I sort of know her # she lives in my neighbourhood ## and besides # she comes here quite often ## I think we can work things out!
MARJA:	<yes> [<2].	PIRJO:	well thank you very much that would be wonderful.
MARJA:	yes # yes.		
PIRJO:	I think they are more open # than we are.		
MARJA:	yea the swedes are.		

structure had the interaction been conducted in a live setting. Many short comments by the listener would have been added at or close to clausal or intonational junctures in each main speaker's turns. In the rest of the tasks on the tape-mediated test the spoken "exchange" was between instructions from the tape and performance from the candidate, and substantive discourse structure was missing. On the face-to-face test, the second task was individual interview, where the interlocutor's turns could be expected to be short while the candidate's turns would tend to be longer. However, even there the interlocutor did make comments and questions, it was just that the turns were often short. On the tape-mediated test such short turns by the "interlocutor" were virtually non-existent because of the limits of the medium on adaptivity. Quite sensibly, the test developers had accordingly decided to use the test time for eliciting longer turns by the candidates.

Figures 6a and 6b show the number of words elicited from each candidate in each test, and the number of words per turn that each candidate spoke in each mode. The higher lines, marked with black boxes, reflect the results of the tape-mediated test, while the lower lines, marked with white boxes, reflect the figures for the face-to-face test. From Figure 6a it can be seen that willingness to talk was an individual property that seemed to be rather stable across the two modes: the lines follow a very similar curve. Figure 6b shows how willingness to talk was reflected in the length and number of turns that the candidates took.

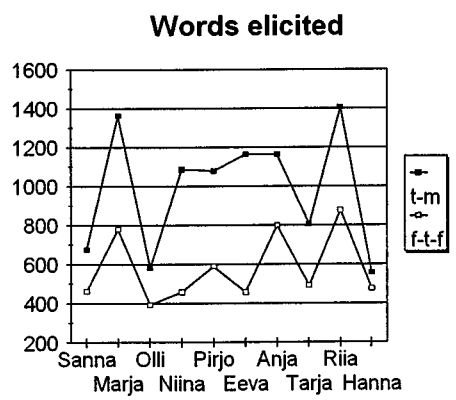


Figure 6a Total number of words spoken by the candidates on each test

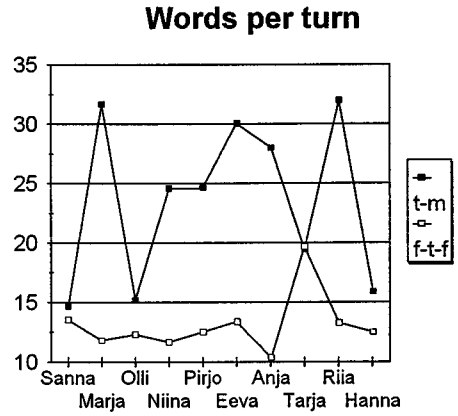


Figure 6b Number of words per turn spoken by the candidates on each test

In the case of the four candidates who clearly spoke less than the six other candidates on the tape-mediated test (the mean number of words spoken by the four was 655, as opposed to a mean of 1211 words for the remaining six candidates), the mean number of words that they spoke per turn was virtually the same regardless of mode. This can be seen in the jagged higher line in Figure 6b, where the four candidates whose turns were short are easy to identify. For the six candidates who spoke more on the tape-mediated test, turn lengths are much higher in the tape-mediated test than the face-to-face one. Because the number of speaking turns did not vary on the tape-mediated test, the only way for a candidate to speak more was to produce longer turns. On the face-to-face test, however, the candidates had another opportunity, namely taking more speaking turns, and according to the figures (6a and 6b), this was what the six candidates did.

There may be more factors influencing the results mentioned above, however. The four candidates who spoke less on the tape-mediated test were all in the lower-scoring half of the sample. Furthermore, their face-to-face tests tended to be short, but this was not a clear trend as there were short test lengths also among the candidates whose turn length varied greatly between the modes. It is difficult to say which of these factors are causes and which consequences. It could be claimed, for instance, that the low-scoring candidates' face-to-face tests were short because the interlocutor was able to establish fairly soon what the level of the candidates was, and the test could therefore be ended. An alternative explanation might be that because of their

lack of linguistic resources, the candidates were unable to handle the topics in an in-depth, more time-consuming manner, and the face-to-face test drew to a natural close as the candidates ran out of things to say. Furthermore, there might have been another variable not captured by these indicators but influencing their values, or indeed the variation may have been random. All in all, these results must thus only be treated as indicative of a possible trend, the nature of which could only be investigated with a larger sample.

5.4.2 Functional analysis

As can be seen in Table 7, there were differences between the two tests in various kinds of counts of the functions elicited. The mean number of functions per turn was 2.6 on the tape-mediated test while it was 1.5 on the face-to-face one. According to a two-tailed t-test with nine degrees of freedom, the difference was highly significant. The reason for this difference was explained above.

The significant difference between the number of functions elicited overall, 93 on the tape-mediated and 65 on the face-to-face test, can be explained by the difference in test length. This difference might not be very serious if it indicated that the same functions were being repeated on the tape-mediated test. However, the difference in function types elicited, 32 versus 24, clarifies that the face-to-face test actually seems to sample a more narrow range of functions. Furthermore, as the wider standard deviation for the face-to-face test on this count indicates, there was much more variation between candidates in the number of functions elicited on the face-to-face than the tape-mediated test. The values ranged from 29 to 35 function types for the tape-mediated and from 18 to 32 function types for the face-to-face test.

The reasons that the range of functions elicited from all candidates on the tape-mediated was stable were most likely that the test was exactly the same for all the candidates, and that the tasks were carefully cued and guided. This lead all the candidates attempting to fulfill the same functions. The skill level differences between the candidates were caused by the fact that they were not equally able to express the functions required in the test. On the face-to-face test, the range of functions expressed by the candidates depended as much on the individual candidates as it did on the interlocutors, who personified the elicitation device in this mode. The range of function types elicited on the face-

Table 7 Comparison of the number of functions elicited on each test

	tape-mediated		face-to-face		t	p
	mean	std	mean	std		
Functions per turn*	2.6	0.4	1.5	0.2	7.12	0
Functions (tokens) elicited	93	17	65	19.1	6.3	0
Function types elicited	32.4	2	24.4	4.5	4.6	0

* The last task of the tape-mediated test, the two-minute presentations, was excluded from this count as the large number of functions in its long turns would have distorted the figures.

to-face test did not correlate significantly with candidate scores. This may be because these factors are not related, but it may also be that the data set was too small to show any effects, or that other variables, such as variation in test length, obscured the relationship. This aspect could be investigated with a larger data set where different skill levels would be equally well represented.

The functional characteristics of the language elicited in the two tests are summarised through counts in Table 8. The functions are reported candidate by candidate, and separate counts are given for each of the six major categories of functions: informative, opinion-expressing, emotional, performative, social, and strategic. Table 8 makes it possible to assess the relative frequency of the function types elicited. It also offers a possibility to compare the functional characteristics across the ten performances analysed within each mode, as well as between the two modes. The two columns where the differences between the modes are the largest are highlighted.

As expected, the first two categories, Giving and asking for factual information, and Expressing and asking about opinions, were the most frequent functional categories displayed in the performances on both tests. Informative functions included categories like stating, narrating, decribing and answering in the positive and the negative, while opinion functions included arguing for or against something, agreeing and disagreeing. There appeared to be some tendency for the face-to-face test to elicit more opinions than the tape-mediated test, but the tendency does was not consistent across all candidates.

The differences between the functions sampled in the two tests were clearest in the performative and especially the social functions. The category of performatives included functions like making and cancelling appointments, offering and requesting help, recommending, and asking for permission. Social functions included greeting and leave-taking, addressing, apologizing and congratulating. The differences in occurrence of functions reflect a real

Table 8 Counts of the types of functions elicited in the two tests

	Tape-mediated						Face-to-face					
	inf	opn	emo	per	soc	str	inf	opn	emo	per	soc	str
Sanna	31	8	1	9	24	5	33	5	3	3	0	2
Marja	54	15	1	13	22	7	30	33	2	8	6	9
Olli	32	6	2	7	18	5	34	13	0	1	1	7
Niina	46	12	1	8	19	8	35	16	2	2	7	10
Pirjo	55	8	3	8	20	9	35	15	2	1	3	4
Eeva	47	7	5	11	22	7	26	14	0	1	4	7
Anja	55	10	3	10	25	7	42	33	1	5	4	8
Tarja	39	6	4	9	20	5	19	7	0	3	2	2
Riia	49	18	0	10	25	9	47	17	4	3	3	14
Hanna	26	6	3	9	17	3	28	18	1	1	2	6

difference in language use between the two tests. Situations requiring the fulfilment of performative and social functions were easily simulated on the tape-mediated test but occurred much less frequently in the face-to-face test because there were no role plays in the test.

The performative functions which did occur during the face-to-face test discourse were suggesting, responding to suggestions, requesting, and joking. The last function did not occur on the tape-mediated test. Social functions that did not come up during the face-to-face test but did occur on the tape-mediated one included opening conversations, inviting, expressly softening/mitigating a speech act, and proposing a toast. Furthermore, as mentioned above, when a function appeared on the tape-mediated test it tended to appear in all the candidates' speech, while on the face-to-face test the range of functions elicited varied from candidate to candidate.

The significance of the specific functional differences between the test modes did not lend itself to statistical testing. The reasons for this were that the process of functional characterisation was subjective, the resulting characterisations only described the main functional content of the data, and the number of candidates in relationship to the number of possible functional categories was very small. It is fairly apparent from the raw count figures in Table 8, however, that there are some differences between the two modes. On substantive grounds, this difference must be counted as potentially significant, as it shows a real difference in skills tested on the two modes.

Due to a number of difficulties, the test instruments were not analysed functionally in the present study. The first of these was defining what should be analysed, a problem particularly concerning the tape-mediated test. The discourse in this test mode could be claimed to consist of the utterances actually spoken aloud when the test was in progress — a narrow definition, but it could be claimed that this corresponds with what would be analysed regarding the face-to-face test. It could also be argued that everything that can be considered input should be analysed, thus including some written material as well. The difficulties in this case would be deciding what was input and what was instruction, and deciding how to deal with input in two languages: the target language and the mother tongue. If all the spoken and written material that had to do with the conduct of each test were analysed, the problem of two languages would be compounded by the problem of different types of text: input and instructions. Furthermore, the kinds of instructions and input varied somewhat from task to task. A different type of problem having to do with the interactiveness of the face-to-face test was the variability in the data, caused by differences in skill levels and differences in the candidates' willingness to talk, among other variables. Frequency counts of the functions in the non-candidate-originated language were run, but the variability of the aforementioned kinds made the interpretations impossible. The small size of the entire data set only made the disentangling of the various effects harder. Functional analysis of the discourse context was thus not conducted in the present study. The effects of this will be taken up in the discussion.

5.4.3 Analysis of word forms

The results of the ACAMRIT analysis indicated that there were a few statistically significant differences between the tests in terms of the forms of language language elicited. On closer examination, however, very few of the differences proved interesting from the point of view of mode-related differences in language use.

For judging the statistical significance of the differences, the SEMSTAT program used the chi-square test. When two texts are compared, chi-square values of 3.8 or higher reach the .05 level significance. When ten texts are compared, a chi-square value of 16.9 or above indicates that the difference is

significant at the .05 level. The interpretation and evaluation of the statistically significant differences was done qualitatively, with reference to the content and topics of the relevant performances as well as to the requirements of the tasks in the two tests.

In the part-of-speech analysis, the number of significant differences per candidate ranged from four to sixteen, and the most common grammatical items on the lists were various personal pronouns, coordinating or subordinating conjunctions, and articles and determiners. In addition to clearly grammatical items such as those mentioned above, the lists quite often displayed the classification 'interjection', which was assigned to words like *hello*, *oh* and *yea*. In the present context, this category is probably more helpfully labeled 'discourse particles'. Various hesitation markers, which had been transcribed as *ehh*, *emm* or *er* according to the nature of the sound on the tape, were also relatively often on the lists of statistically significant differences.

Features for which the differences between the two tests were non-significant included nouns, adjectives and adverbs, i.e. content-oriented categories. The analysis included one initially baffling exception regarding nouns, however. Singular common nouns tended to occur more often on the tape-mediated test than the face-to-face one, and the difference was statistically significant in seven candidates' performances. Plural common nouns, while overall being less commonly used in both tests than singular nouns, were used more on the face-to-face test than the tape-mediated one. The differences were statistically significant differences in six cases.

The explanation was twofold. Firstly, the tagger erroneously analysed the hesitation marker *ehh* as a singular common noun. As this particular form of hesitation was much more common on the tape-mediated test, it explained some of the difference. The other explanation was contextual: the kinds of tasks that were included contained a number of simulations where frequent reference to single entities such as a coat, a ticket or a shop were made. The topics in the face-to-face test tended to involve a more balanced mixture of plural and singular nouns. Apart from the case of the hesitation marker, the concordances showed little which would raise concern about the different tendencies in noun use in the two tests. Though statistically significant, from the point of view of language use the difference was uninteresting.

Two groups of linguistic features which were more grammatically oriented but did not differ significantly between the two tests were negation and various verb forms such as modal auxiliaries and simple and complex tenses. This indicates that the tests are unlikely to have differed in terms of this kind of grammatical complexity.

Regarding personal pronouns, when significant differences occurred, they tended to suggest that singular pronouns, particularly *my*, *you*, *he* and *she* were used more often in the tape-mediated test while *they* and *that* tended to be used more often on the face-to-face test. This was because of the different situations and topics included in the two tests: the simulated encounters between the candidate and public officials or relative strangers often required reference to the interlocutor or other individual people, while dealing with the general topics on the face-to-face test often made the candidates refer to various people and how "they" perceived the world to be. Furthermore, as there was statistically higly significant variation between individuals within a test mode regarding personal pronouns, particularly *you* and *they*, it is unlikely that this difference was singularly associated with the test modes. From an assessment point of view, additionally, this difference can hardly cause concern, since it is not likely that plural pronouns would be more difficult to master than singular ones, or vice versa.

The frequent use of the second person singular pronoun might indicate an interactive orientation in language use, a you-focus on the interlocutor. The concordances of the pronoun *you* were thus investigated in order to find out if this was the case. It appeared that the second person singular pronoun was used rather differently between the modes, but not quite as was expected.

The second person singular pronoun was used for two purposes, generic and personal reference. The use of *you* in personal reference was much more common than its use in generic reference on the tape-mediated test, while the situation was the reverse on the face-to-face test.

This difference was clearly related to task type. The instances of personal reference on the tape-mediated test came from the simulated interactions, while the instances of generic reference appeared in the last task, two mini-presentations on topics of general interest. The last task was the closest parallel to the topic-oriented discussions of the face-to-face test, and the use of *you* in this task and the face-to-face test was practically identical.

Table 9 Frequencies of conjunctions in the two tests					
Coordinators, tape-mediated test			Coordinators, face-to-face test		
conjunction	N	%*	conjunction	N	%*
and	364	44	and	247	41
but	90	11	but	63	11
or	86	10	or	29	5
or_anything	2	0.2	and_stuff, and_everything	1 each	0.2 each
as_well_as	1	0.1			
Subordinators, tape-mediated test			Subordinators, face-to-face test		
that	119	14	that	104	17
so	36	4	because	44	7
if	32	4	if	30	5
because	28	3	so	18	3
if (whether)	15	2	but	13	2
when	14	2	if (whether)	10	2
but	12	1	when	7	1
than	7	1	so_that	6	1
as	4	0.5	than	4	0.7
so_that	4	0.5	as	3	0.5
where	3	0.4	before	3	0.5
for	3	0.4	where	3	0.5
before	3	0.4	after	2	0.3
now	2	0.2	once	2	0.3
whether	2	0.2	as_I_said	2	0.3
except, now_that, as_you_have_said, even_if	1 each	0.1 each	now_that, while, though, for, as_it_says, as_it_is	1 each	0.2 each

* Percentage of all conjunctions in this test mode. This figure is reported for ease of comparison between the modes.

The few instances of *you* in personal reference on the face-to-face test appeared as a rule at the beginning and/or end of the one-on-one interview and referred to the interviewer. Only one candidate used *you* to address her peers in an interactively-oriented way during the group discussion. None of the candidates had problems with either use of *you* reference, and thus the difference in emphasis of elicitation seemed to make little difference.

The most frequent conjunctions on both tests were *and*, *that*, and *but*, and the order of frequency of these conjunctions was the same in both tests. Furthermore, the range of conjunctions that appeared in the two tests was almost identical (see Table 9). Coordinating conjunctions were more frequent than subordinating ones on both tests, and the range of different subordinators was much larger on both tests than the range of different coordinators.

The only notable difference in the use of conjunctions between the two modes was the ratio between coordination and subordination: there were approximately twice as many coordinators as subordinators on the tape-mediated test performances, while the ratio was approximately 4:3 on the face-to-face test. Shohamy (1994:114, 117) suggested that there was a difference between the use of connectors in test performances on the SOPI and the OPI, but did not specify what this difference was. The results of the present study suggest that there is no difference in the range of forms used, but that there is a possible difference in a tendency towards more coordination on the tape-mediated test.

In all the test performances in both modes, discourse particles, such as *oh* or the filler *emm*, tended to appear in the beginning of turns, or in connection with pauses within turns. The differences in their use were statistically significant in six candidates' performances out of the ten analysed. Five of these six candidates used discourse particles more frequently in the face-to-face test, while one candidate used discourse particles more often in the tape-mediated test. The difference in this one candidate's performance was that she was particularly well able to adjust to the simulations in the tape-mediated test, and used discourse particles to fill what little thinking time she needed before making a reply. As this candidate had lived abroad, the public and semi-official encounters in English were probably also familiar to her.

For the most part, the use of discourse particles was similar in both test modes, as the examples in Table 10 illustrate. The differences that appeared seemed to be based on the use of a few interactive and/or backchanneling signals, which very rarely appeared in the tape-mediated test: *aha*, *ok* and *yea*. By using these particles the candidates indicated attention, comprehension, and agreement in the discussion. Such use of discourse particles is virtually impossible on a tape-mediated test, and their absence shows that the test developers had realised this. The key feature in backchanneling is that it

Table 10 Examples of use of discourse particles in the two tests	
tape-mediated	face-to-face
and I go to the concerts too but ***emm*** *# sometimes they play too loud and...*	*But now we have had a #* ***emm*** *# so many people out of work.*
oh *that's very nice # thank you very much for your help*	***oh*** *I didn't see that # that it is so late # I should be home now*

accompanies and adapts to the main speaker's turn, and such adaptability is impossible in the kind of tape-mediated test investigated in the present study.

As regards voiced hesitation, there were significant differences in the performances of five of the ten candidates. The hesitations were transcribed in a semi-formalised way, but some attempt was made to preserve the original character of the sound. This was done because at the transcription stage the informational value of the various non-word-formed hesitation sounds was unclear. It turned out that voiced hesitations were very often of the form *ehh* on the tape-mediated test, whereas this form very rarely appeared on the face-to-face test. However, rather than being a meaningful difference in language use, this was probably due to greater accuracy of recording on the tape-mediated test where the microphone was within two inches of the speaker's mouth. In future analyses, while some differentiation between non-word-formed hesitation sounds may be relevant, it may be more beneficial to use one uniform transcription in the texts submitted to ACAMRIT analysis. Even if this were done with the present data set, however, the result would remain the same: the five candidates whose performances differed between the modes with respect to hesitation produced vocal hesitation markers more often on the tape-mediated test than on the face-to-face one.

The concordances of the hesitations in the performances of two candidates are presented in Appendix 12. The markers *ehh* and *er* are both shown as *er* to reduce the variation which appears to be more related to quality of recording than to the candidates' language use. From these examples it can be seen that the voiced hesitations in either mode functioned in a very similar way. It is possible to interpret the hesitation markers as lexical searches or as thinking pauses, or as a combination of both. It would be extremely difficult to find proof for which, or what sort of combination of both, the reason for the hesitation marker was in each case. The only conclusion that can be drawn from the results is that for some candidates, there seems to be a tendency for the tape-mediated test to cause more vocal hesitation.

A possible interpretation for the higher frequency of voiced hesitation in the tape-mediated test is that this test form caused more anxiety or encouraged more self-monitoring in some candidates than the face-to-face test. This interpretation was expressly supported by one of the three post-test interviews conducted in Study One. Further investigation of this connection would require a more extensive and systematic focus on monitoring and anxiety through gathering information from candidates after the test and comparing this with features in the test performances. It would be important to confirm the interpretation of possible questionnaire-based data on self-monitoring through interviews to make sure that the researcher's interpretation coincides with what the candidates wanted to express.

With the present data, it is impossible to judge the effect of voiced hestitation markers on the grades awarded. None of the candidates whose vocalised hesitation differed significantly between the modes received a different grade from the two tests. However, the data set was small, and the skill levels covered by the tests limited. The possible effect of voiced hesitation on grades awarded cannot be ruled out on these results either.

The concordances of voiced hesitation in Appendix 12 show that the category of voiced hesitation is problematic. It only illuminates part of the hesitation phenomenon, excluding both silent hesitation/pausing and various word-formed ways of winning time, such as circumlocution or repetition of previous words in the utterance. Unless it can be shown that non-word-formed hesitation markers as a distinct category have an effect on grades awarded, it may be more useful to analyse the whole phenomenon of hesitation. This would require tagging all the relevant features of language use by hand. An obstacle in tackling this question is that on the basis of the present data set, hesitation appears to be at least as strongly related to interlocutor, assessor or tagger perception as it is to features of candidate speech.

Similarly to the part-of-speech analysis, the semantic field analysis of the ten performances produced a number of statistically significant differences in language use between the two modes. However, all the differences were explained by the different topics covered in the two tests. For instance, there were seven sematic categories which differed significantly in Candidate 1's performances in the two tests. Four of these had to do with music, sound, liking, and good/bad evaluations. This was because one of the two mini-

presentations had to do with pop music. One category had to do with power and organising, and resulted from the second mini-presentation topic, leaders and leading. Two categories had to do with obligation and necessity, and exclusivisers. These reflect the attitude of Candidate 1 to recycling, the topic of her interview.

When individual words within the semantic categorisations were analysed, the results reflected two tendencies: the topics covered in each test, and the syntactic differences covered in the part-of-speech analysis. Although grouped under different categories, the syntactic features highlighted were the same, as was the fact that few differences were tied to anything but lexical and grammatical choices of individual candidates.

5.5 Summary and discussion of results

The first research question in Study Two inquired how the two test modes compare as contexts for discourse elicitation. In addition to obvious differences in physical context, the analysis of the two tests as discourse environments revealed an array of features which reflected the central difference between the two tests: the presence of real-time human interaction and the absence of it. This affected the instructions, the input, the expected response and the relationship between input and expected response.

The language, structuring and content of the instructions on the two tests were affected because various means had to be used on the tape-mediated test to make all the instructions fully comprehensible to all participants without online explanation, while on the face-to-face test the interlocutor was able to clarify misunderstandings if any occurred. Two strategies were employed in the tape-mediated test to thelp the test run smoothly: the instructions were delivered task by task, and where processing might be too difficult or consume too much time, the native language was used instead of the target language. On the face-to-face test, in contrast, the greater part of the instructions was given in one written bulk before the interaction began, and valuable test time was only used for making sure that the instructions had been understood and in making the transitions between the two tasks. Negotiation was always possible during the transitions if something remained unclear, but was only entered into if the situation demanded it. Such tailoring

of instructions was impossible on the tape-mediated test, hence the difference in format and sequencing.

The presence and absence of real-time interaction also showed in every aspect characterising the format and language of the input, including the observation that input was more difficult to define for the face-to-face test and impossible to investigate a priori, while it was much more easily identified and investigated on the tape-mediated test. There was more written input on the tape-mediated test, and both the target and the native language were used. The formulations of the input were planned on the tape-mediated test and exactly the same for all the participants, while the formulations varied a great deal on the face-to-face test.

Finally, the difference in interactivity between the two modes had an effect on expected response. In the interest of making the simulated conversations on the tape-mediated test work, the conversations had to be predictable and therefore guided. The guidance meant that the central content of the responses was predictable, only the way in which the candidates expressed this would vary. Few restrictions of this sort were present in the face-to-face test although the interactions were directed by the interlocutor. The difference also showed in the relationship between input and expected response, where on the tape-mediated test the input was not influenced by previous responses, while this, at least to an extent, was the case on the face-to-face test.

The overall conclusion on the comparison of the tests as contexts for discourse elicitation was that the discourse environments were different; the discourse in the tape-mediated test was much more structured and guided than in the face-to-face test. This was because each test had been developed to make use of the potentials of the mode in which it was conducted while observing the restrictions and avoiding to emulate the other mode at the cost of harming what was perceived to be the nature of the test. The differences in task structuring are likely to have affected the discourse elicited from the participants, and some support for this hypothesis in lent by the candidates' comments in Study One. The effect seems to have concerned the process of the test, particularly the perceived room for negotiability and perhaps involvement in real social interaction. When the language elicited was analysed in terms of words, turns and functions, few differences were detected between the modes.

The second research question in Study Two related to the communicative functions elicited in the two modes. The results indicated that in accordance with the expectations of the test developers, the tape-mediated test elicited a slightly wider range of functions than the face-to-face test. Furthermore, the range of functions elicited from all candidates was rather stable on the tape-mediated test regardless of skill level, indicating that the skill level differences between the candidates appeared in how successfully they were able to express the functions elicited. On the face-to-face test, the range of functions expressed varied between candidates. However, there appeared to be little association between range of functions expressed and skill level achieved.

The differences between the functions sampled in the two tests were clearest in the social and performative functions, the expression of which was overtly elicited in the tape-mediated test. These functions only appeared in the face-to-face test if they came up naturally in the course of the test, not through direct elicitation. The only social function which appeared on some face-to-face test performances but in none of the tape-mediated ones was joking, an interpersonal strategy for managing social interaction.

The third research question concentrated on features to be analysed in oral test discourse in order to describe similarities and differences in language use between the modes. The answer provided to this question in Study Two was partial: the features analysed in the present study offered a start and showed similarities, but other analyses are needed as well, particularly to capture the nature of the differences between the tests.

The initial characterisation of the performances through counts of words and turns was useful for quickly describing some basic features in the data. The differences found here related to test length and turn structure, while the similarities showed that the two modes can elicit approximately the same amount of language in number of words per test minute. This superficial characterisation reveals few aspects of the nature of language used in the tests, and thus needs to be complemented.

The automatic content analysis of the performances indicated that there were many similarities and few differences in language use on the lexical-phrasal level between the two test modes. Features that were not significantly different between the test modes were syntactic features of open word classes such as adjectival comparison or use of verbs. The singular-plural differences in pronoun use, though statistically significant in some candidates' speech,

appeared not to be of great importance from a language use point of view. The closer investigation of the second person singular pronoun revealed a difference in use of *you* in personal and generic reference. However, this difference was more related to task type than to test mode.

Regarding mode-related differences on the lexical-phrasal level, it was possible that the tape-mediated test elicited slightly more coordination than the face-to-face test. If subordination is considered more complex, this could be interpreted to mean that the face-to-face test had a tendency to elicit slightly more complex structures. The range of conjunctions elicited by the two tests was practically identical, however.

Another difference was that discourse particles were more commonly used in the face-to-face test. This difference was clear and interpretable, being most clearly related to backchanneling signals, which the tape-mediated test did not elicit at all.

A third difference was that vocalised hesitation appeared to be significantly more common in some candidates' performances on the tape-mediated test. This could suggest higher test anxiety and/or more monitoring in the tape-mediated test, which may be related to the candidates' perception of the requirements of the test. The voiced hesitation could thus indicate that they are planning how to lexicalise their performance and observing how they have done so far. Some support for this interpretation is lent by one of the post-test interviews in Study One.

It must be pointed out that the hesitation phenomenon investigated in Study Two was voiced hesitation through non-word sounds, rather than the whole phenomenon of what could be interpreted by interlocutors or assessors as hesitation. Further investigation of the significance of this difference, as well as the perception of hesitation overall, is needed before this finding can be interpreted.

The semantic field analysis of the performances proved only to show that different topics were taken up in different tests. This type of analysis is probably more useful for material from a single context, such as doctor-patient discussion on a single medical condition (for instance Thomas and Wilson, 1995).

The shortcomings of the analysis of test discourse conducted in Study Two were that apart from the functional characterisation, the analyses were word-focused and mostly conducted through counts rather than judgements.

Such analyses are useful if the units counted are relevant and a sufficient number of them are used so that multiple dimensions of the data can be investigated. The discrepancy between the results indicating similarity and the ones suggesting differences both in task characteristics and in participant perceptions suggest that too few dimensions were covered in Performance Analysis in Study Two to describe the differences.

Perhaps the most severe shortcoming of the analyses of test discourse in Study Two was that the tests as elicitation devices and the performances elicited were analysed in isolation of each other, while the interplay was not analysed. This was partly a flaw in the original design, partly a concern of the quality of the data. The reason for the decision in the original design to divide the investigation into Task Analysis and Performance Analysis was that in other test modes than in speaking and perhaps in writing, such divisions are possible and commonplace. Furthermore, very little was known of the discourse in the two tests at the beginning of the present study, and the results of Study Two both confirmed expectations and highlighted the shortcomings of the analysis. Had the discourse data collected for the present study been representative both in terms of numbers of candidates and in terms of the quality of the test instruments, analyses of the interplay could have been conducted. As the case was, the data were assessed to be too unrepresentative to warrant further analyses. The recommendation that the interplay between elicitation and performance be analysed in future studies was allowed to suffice as a result in Study Two.

5.6 Discussion of methods

Similarly to previous studies where characteristics of test tasks have been compared through systematic analysis, Task Analysis in Study Two produced a detailed description of the two tests, from which few generalisations could be drawn. In the present case, the generalisations pertained to the difference between the tests which had been evident before the study was conducted: interactivity. In the Cambridge-TOEFL comparability study, the results were a list of features which highlighted the differences between the British and American testing traditions — tests of speaking were not examined in that study, but it is difficult to see whether this would have changed the

conclusions. The reason for the difficulty in forming generalisations may be the detailed and atomistic nature of the data analysis instrument: many aspects of the test instruments are covered, but they are covered one by one, and the compartmentalisation of the list carries over to the analyst's interpretations. The result of implementing a detailed taxonomy is a highly specific list.

The TC analysis in Study Two was able to specify ways in which interactiveness and the lack of it appeared in the test instruments. The analysis concentrated on the settings and the language, while it did not consider the structure of the situation or the roles of the participants. The reason that the TC framework does not focus on task structuring or interactant relationship may be that the system has not been extensively used for analysing tests of productive skills, or that the framework is intended for characterising all kinds of language use situations and thus has to be very general. In contrast, a system developed specifically for classifying composition assignments (Purves et al. 1984) identifies these and other characteristics of test tasks which are not specified in the TC framework. The Purves et al. (1984) system consists of fifteen dimensions, including cognitive demand, purpose, role, audience, and rhetorical specification, all of which dimensions are largely missing from the TC framework. These dimensions are specific enough for characterising individual test tasks, so much so that they provide a useful checklist for aspects which should be covered in test specifications. This was indeed one use that Purves et al (1984:400) propose for their system. The dimensions of task structuring and interactant relationship help specify the difference between the two test modes from a specifically testing-oriented point of view.

Task structuring focuses on a central difference in the nature of test discourse between the tape-mediated and the face-to-face test investigated: the highly structured tape-mediated test aims to find out what the candidates *can* do in response to specific tasks, while the face-to-face test samples what the candidates *will* do when faced with a more open-ended prompt (cf. Purves et al. 1984:400). This is one way of expressing how the communicative expectations from the two tests are different. In future studies of task characteristics in tests of speaking, a test-oriented, task-specific approach such as that of Purves et al. (1984) may prove useful. The system must be adapted to the tests and tasks being analysed, but its approach to communication as linguistic, cognitive and social activity holds promise.

The theoretical gain from the TC analysis as it was conducted in Study Two was that the category of input was shown to be problematic. This was an important discovery both in terms of the tests investigated and in terms of the TC framework. For the tests, this alerts the developers to exert what control they can over the variability of the input in the face-to-face test and observe its possible effects on test scores. For the framework, this calls for consideration of implicit assumptions in the categories, and possibly for considering limitations of applicability of the system without modifications.

The ACAMRIT analysis of the test performances was useful in three ways. Firstly, it offered a number of potentially significant differences in language use which it was then possible to investigate quite easily through concordancing. This was a great advantage both for specifying and interpreting the results of interesting differences and for rejecting other proposed differences. Secondly, it raised the issue of hesitation as a potential focus for further research, albeit through a more qualitative approach. Thirdly, the small number of mode- and test-related differences found suggested that on the lexical-phrasal level, there are many more similarities than differences in language use between the two tests.

The investigation of the use of *you* confirmed suggestions by other researchers (eg Shohamy et al. 1993, O'Loughlin 1995) that some differences in language use may be more closely related to task type than to test mode. Either for task- or for test-related variation in language elicited, the main result of Study Two is that if there are significant differences in language use, they may be found at the level of the co-text and context rather than the text. Thus, approaches such as Lazaraton's (1996) descriptive conversation analysis, or the somewhat more quantitative approach of Ross and Berwick (1992) may prove fruitful, particularly if the performances of both the candidates and their simulated or real interlocutors are analysed together.

6 STUDY THREE: ASSESSMENT

Study Three consisted of two different analyses of the scores given on the two tests. The first was an essentially numerical analysis of the scores, complemented by a closer description of those cases where assessments from

the two modes differed. This was done to search for possible explanations for the difference. The second analysis was an exploratory investigation into the ways in which the assessors construed the similarities and differences between the performances. This was done through employing a version of Kelly's (1955) Repertory Grid procedure, which uses systematical comparison to help the respondents operationally define their conceptual space. The completed grids were interpreted through the assessors' comments as well as submitted to factor analysis.

The research questions addressed in Study Three were:

3a. Is there a difference in the candidates' scores on the two test modes?

3b. Which features of performance do the assessors pay attention to when making their assessments in the two modes, and are the features the same in both of them?

The working hypothesis regarding the first question was that as in previous studies where scores from the two modes had been compared (for instance Stansfield 1991, Wigglesworth and O'Loughlin 1993), the correlation would be quite high. As no previous studies on the assessment constructs in the two modes had been conducted, the first part of the second research question simply explored the assessment constructs that were important for the assessors. Regarding the second part of the second question, two contrasting working hypotheses were posed. It was possible that, whether or not the scores were highly correlated, assessment in the two modes would be based on the same features of language. This would mean that the construct of the score would not necessarily be the same as the construct of the test. It was equally possible that the assessors paid attention to different aspects of the performances in each test, but that these features would tend to co-occur in individual candidates, at least the ones who participated in the present study, such that their performance on one test could be predicted from the other. This would mean that the constructs of the scores were rather closely linked with the constructs of the tests, and without empirical investigations of association, each score on its own should be interpreted differently.

6.1 Methods used in the analysis of assessment

The candidates' scores on the two tests were obtained through getting two independent assessments of the performances. When these did not coincide, the assessors listened to or viewed the performance again and reached a consensus mark through discussion. Interrater reliability was assessed on basis of the two sets of original scores through Cronbach's coefficient alpha. After this initial step, the strength of association between the candidates' scores from the two tests was measured through correlation. Since the data were ordinal, the Pearson coefficient was deemed appropriate (Hatch and Lazaraton, 1991:436), although the distribution of the scores was negatively skewed. In addition, the scores were cross-tabulated in order to identify the candidates whose scores differed between the two tests. These cases were then considered one by one, observing both the performances and the questionnaire and interview responses, in order to detect possible causes for the difference.

In order to investigate how the numerical scores could be interpreted, a different focus was needed. The choice was made to investigate which assessment constructs were important for the assessors. These were investigated through a version of the Repertory Grid technique, developed as an application of Personal Construct Psychology by George Kelly (1955) and extended and modified by several of his students. The variants of the Role Construct Repertory Test were first intended by Kelly to be used in clinical psychology instead of traditional psychological tests, and the approach has subsequently been applied in personality psychology as well as other research fields such as management, education, and marketing.

According to Bannister and Mair (1968), the advantage of the Repertory Grid technique is that instead of investigating a phenomenon with constructs determined by the researcher, it gives the power to define the constructs and the yardsticks used in the analysis to the respondent. This is in accordance with Kelly's theory, which builds on the conception that reality is subjective and that individuals organise their reality with each their own sets of hierarchically ordered constructs in order to make sense of their worlds.

The grid form of the Role Construct Repertory Test involves taking a host of elements -- in the present study language test performance extracts -- and investigating through systematic comparison which constructs the respondent uses to distinguish between these elements. In the elicitation phase,

the respondents may name a large number of possible constructs, and subsequent analysis of the relationships between the constructs named may result in a much smaller number of constructs actually used.

The choice of the Repertory Grid as the method for analysing the criteria that assessors employ was inspired by a study by Pollitt and Murray (1993), where the method was used in combination with a binary decision on which of the two language test performances being compared was better. For the context of the present study, the preference decision was not relevant, while the systematic elicitation of individual assessment constructs appeared to hold promise. A different variant of the method was chosen in accordance with the need to focus on individual assessment constructs across several performances.

The type of Repertory Grid used in the study is described in Bannister and Mair (1968:63-65); the specifics of the procedure are described below. The potential assessment constructs were elicited with the help of approximately five-minute-long extracts from test performances. The face-to-face test grid analysis was based on 15 video-taped extracts, and the tape-mediated test analysis on 12 audio-taped extracts. Both groups represented the whole skill level range available in the material. Two assessors took part in both analyses; the present writer was one of them. Unlike the normal applications of the technique, the assessors went through the elicitation process together rather than individually. The elicitation procedure is illustrated in flow chart form in Figure 7. A completed grid for the face-to-face test is shown in Appendix 13.

The construct elicitation was begun with an empty grid. First, three performance extracts were chosen, viewed, and held in mind. On the first round these were any three performances, on subsequent rounds attempts were made to use different combinations and make sure that all the extracts were used at least once. After viewing the three performance extracts, the criterion which most clearly distinguished between them was named and written down in the first row of the grid. In addition, the high (7) and low (1) ends and the middle point (4) of the criterion were described to establish its range. Then the chosen triad of elements was rated (1-7) according to this criterion, and after that all the rest of the elements were viewed and rated on the same criterion. After the last performance extract was assessed according to the current criterion, a new triad of performances was chosen, and the process

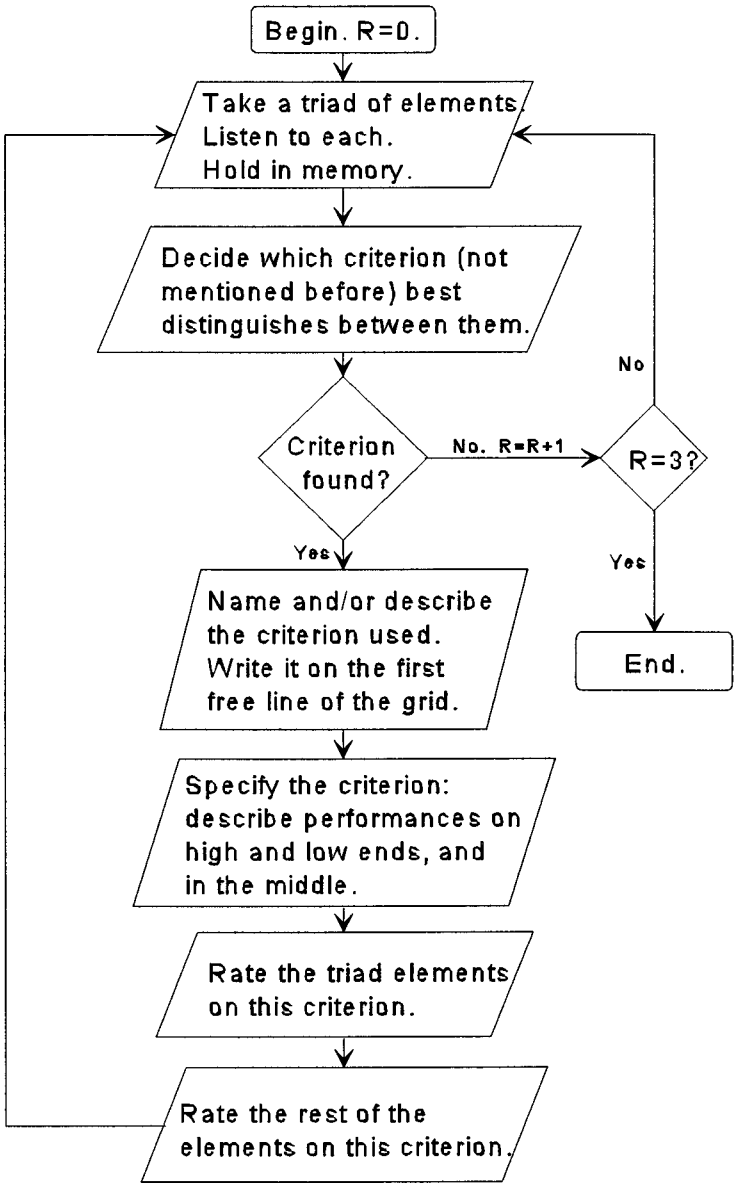


Figure 7 The grid elicitation procedure

begun again. When it became difficult to find further important criteria to distinguish between the performances, two new triads were chosen to make sure that our repertory of important assessment constructs really was exhausted. When the result after three triads was no new criteria, the elicitation was finished.

According to Fransella and Bannister (1977), the grid could best be viewed as a structured interview technique, and that is how it was used in the present study. The aim was to find out which features the assessors were paying attention to when making the holistic assessments. The procedure was

almost the opposite of normal assessment practice: instead of using several criteria for making a global assessment of one performance, the assessors compared several performances and only looked at one feature in the performances at a time. This gave them an opportunity to focus on and clarify the criteria they were using, and both of them felt that the experience was very rewarding. For instance, this process helped them realise that a vague feeling of unnaturalness they had had about some performances stemmed from an almost complete absence of idiomatic language in these candidates' speech.

The formulation of the grid elicitation question was initially a context-specific application of the most basic grid question: "Which criterion (not mentioned before) distinguishes two of these performances from the third?" However, it soon became apparent that a more appropriate formulation was "Which important assessment criterion (not mentioned before) distinguishes best between these performances?" The difficulty with the previous formulation was its strong bipolarity, the latter formulation suited the scalar assessment better. The change was in keeping both with the internal working of the ratings grid and with the general aim in the use of repertory grids to help informants express their constructs rather than artificially constrain them to the arbitrary requirements of a method.

In spite of possible problems with memory load, the choice was made to follow the standard procedure and do the construct elicitation with triads rather than pairs of elements. In retrospect, this solution was appropriate because it helped the assessors concentrate on criteria that applied to more than just two performances. It also helped them see a continuum and specify the criterion before applying it to the rest of the performances. Holding three video-taped extracts in memory was not problematic, while considering a triad of audio extracts required concentration initially, before the process and the extracts became familiar to the assessors.

Study Three investigated the assessments of experienced assessors because it is their assessments that count in test results, and because we do not know very well which aspects of the performances are important for trained assessors, let alone lay people who do not need to think of language proficiency assessment on a very detailed level. Also, as the data were going to be used for developing more analytic scales for the test system, this method could provide indications as to which criteria the assessors find useful in assessment work.

The labels that the assessors used for the important assessment criteria during the elicitation were arrived at on the basis of observing twelve or fifteen performances. Different combinations of triads might have produced slightly different labels, as might different performances. Furthermore, because the aim was to elicit *all* the important criteria, a much larger number of labels might have been elicited than the number of different criteria actually used by the assessors. It was therefore necessary to analyse the grids both qualitatively and quantitatively to find out whether the information in them could be simplified in some way. As discussed in Bannister and Mair (1968:65-66), both kinds of analysis are potentially informative. The analysis by the informants themselves reflects the degree to which they are aware of the relationships between their constructs, while the numerical analysis may reveal relationships that the informants are unaware of.

The qualitative analysis of the grid data was done directly after completing the grids, and consisted of a discussion. The questions that the assessors were addressing were how each criterion label was related to each other, and specifically whether there could be a learner whose skills were good in one respect and bad in the other. The discussion profited from the assessors' familiarity with the descriptive and operational definitions that they had just provided.

In order to triangulate the qualitative data, attempts were made to analyse the grids quantitatively. The size of the grids proved small for extensive analysis, however. Two analyses were conducted: correlations between the criterion labels, and factor analysis with both orthogonal and oblique rotation. The analyses were conducted separately for the tape-mediated and the face-to-face test.

6.2 Results of the score comparison

The quantitative analysis of scores was begun by an assessment of interrater reliability. This served as a quantitative check on the degree of confidence that could be placed on the scores. Interrater reliability was estimated by means of Cronbach's coefficient alpha. The means and standard deviations of the scores that the two assessors gave on each test are given in Table 11, followed by the

Pearson correlation coefficient between the scores and the reliability estimate, Cronbach's alpha.

Table 11 Means and standard deviations and interrater reliability on the two tests

	Mean	Standard deviation	Pearson correlation	Cronbach's Alpha
Tape-mediated test				
Assessor 1	4.38	.794		
Assessor 2	4.40	.798	0.978	0.989
Face-to-face test				
Assessor 1	4.51	.806		
Assessor 2	4.49	.753	0.906	0.951

The range of scores from both tests in both assessors' results was from 2 to 5. The alpha coefficients suggest a very high degree of interrater reliability on both tests: .99 for the tape-mediated test and .95 for the face- to-face one. The reason for such high agreement between the two assessors is probably that these were the two people who had developed the scale which they were using to report the scores.

Regardless of the near-perfect agreement between the assessors, the conclusion cannot be drawn that they would have given their assessments on the basis of their reaction to the same features in the performances. It must be remembered that the scores represented an assessment of overall skill level, not individual aspects of the performance. The agreement between the assessors was a little higher on the tape-mediated test, but the difference was very small. On the basis of these results, the case can thus not be made for the scores from one test being more reliable than those from the other. With 37 candidates in the sample, the interrater reliability figures can be considered fairly accurate. Any differences between the scores from the two modes should thus be treated seriously.

The Pearson correlation between the scores from the two test modes was .89 when all 37 candidates were included, .85 when the eight candidates whose skills exceeded the requirements of the test were excluded. The cross-tabulation of the scores is presented in Figure 8. Six candidates received different scores from the two test modes; two did better on the tape-mediated test and four on the face-to-face test. Harshness was thus not consistently associated with either test mode. All the six candidates whose scores differed

Figure 8 Cross-tabulation of the assessment results

face to face test \ tape-mediated test	2	3	4	5
2	1			
3		2	1	
4		3	8	1
5			1	12 (20*)

* includes the 8 cases whose skills exceeded the test requirements

between the modes were assessed to be in the upper ranges of the lower assessment they received, for instance a "high" 3 on the tape-mediated test and a 4 on the face-to-face one, or a "high" four on the face-to-face test and a 5 on the tape-mediated one.

Mode-related factors explained at least three of the four cases where candidates received a better assessment on the face-to-face test. Two of them had a very strong negative reaction to the mechanical testing environment of the language laboratory. Their anxiety was clearly audible in a shaky voice and long pauses, and it affected their performance. One candidate had poor test-taking strategies in that he did
not attempt the more difficult tasks on the tape-mediated test at all. The fourth case was less clear in that this candidate had a favourable reaction to the language laboratory environment but was assigned skill level 3, mainly because of her slow speaking rate and unclear pronunciation. On the face-to-face test her speech was easier to understand because her lip movements were visible. In addition, she was able to decide more freely what she wanted to say and thus show rather good command of vocabulary. Consequently, she was assessed at level 4. It is very difficult to say which of the two skill levels was more appropriate for the candidate.

Of the two candidates who did better on the tape-mediated test, one apparently had good memory for phrases and excellent communication and test-taking strategies for the tape-mediated test. He used many phrases appropriately, only occasionally leaving some of the requested content unexpressed, and kept his answers fairly short. On the face-to-face test it

proved that the candidate's command of basic grammatical structures was shaky, and simplification and lexical approximation were clearly evident in his performance all through the test. Part of the reason why this showed so clearly was that his strategies for communicating in the face-to-face situation were not very appropriate for the test context. He often seemed to expect the interlocutor to help him when he had started an expression but ran into difficulties, and instead of circumlocuting he twice just gave up trying to express an idea, saying he was unable to explain this in English. The first of these strategies should not automatically trigger a lower assessment, of course, but unfortunately the situations in which it appeared revealed just how little of the language the candidate knew, and how much his English was influenced by other languages that he knew.

The other case of candidates who performed better on the tape-mediated test was more difficult to explain. It is possible that she performed more poorly on the face-to-face test both because of lack of rapport with the interviewer and because of bad luck in the choice of topic.

In sum, the scores from the two tests correlated well. The sample was small and the conclusions therefore only tentative, but statistically, the correlation was quite impressive, the variance shared being approximately 70%. Judging by the differentially assessed candidate performances, additional variance could be explained at least partly by method-related variance on the tape-mediated test and possibly personality-related factors on the face-to-face test.

6.3 Results of the analysis of assessment constructs

Through the Repertory Grid procedure, the assessors named nine assessment constructs for the tape-mediated test and ten for the face-to-face test (see Figure 9). Six of these were common to both test forms: speed, or rate of speech, confidence, vocabulary, use of collocations and idioms, prosodic features, and grammar. Aspects of pronunciation other than prosody were considered important in the assessment as well, but in slightly different ways in the two test modes: on the tape-mediated test the criterion was called accent, or the amount of effort needed to understand the speaker's pronunciation, while on the face-to-face test the criterion was called pronunciation of individual

Figure 9 Potential assessment constructs elicited	
tape-mediated test	face-to-face test
Rate of speech Confidence Vocabulary Use of collocations and idioms Pronunciation / prosodic features Grammar	
Accent / amount of listener attention needed	Pronunciation / Individual sounds
Creativity (in using the test rubric)	Transfer / Mother tongue interference
Phrasal knowledge	Comprehensibility / Ability to convey core and exact meanings
	Length of turn

sounds. The former construct probably encompassed a larger number of features than the latter. Rather than a difference in assessment between the test modes, however, the most likely reason for this difference was the characteristics of the triads which were used to elicit the constructs. In the tape-mediated test, it was accent which differentiated between the three, while on the face-to-face test it was the micro-level accuracy of pronunciation.

Content and meaning were focused on in the assessment as well, and it was here that the differences between the test modes showed very clearly. On the tape-mediated test, where the assessors knew what the content of the expected response was, the criterion was called degree of creativity in expressing the content, whereas on the face-to-face test, the criterion was called comprehensibility, or ability to convey core and exact meanings. Another aspect of expressing meaning was suggested as a criterion on the face-to-face test as well, namely transfer, or interference from other languages in conveying meaning.

Some of the criteria mentioned were not only related to test modes but to task types as well. On the Simulated Conversations and Reacting in Situations tasks in the tape-mediated test, phrasal knowledge was clearly the focus of assessment, while this criterion had no direct equivalent on the face-to-face assessment. Similarly, length of turn, or willingness to talk, was reported to influence assessment throughout the face-to-face test, but was not mentioned among the tape-mediated test criteria. This was probably related to the degree of structuring in the tasks of the two tests. In the tape-mediated test, the expected content of the replies was most often clearly specified, and this guided the candidates to say quite a lot, whether they would have been

inclined to do so without prompting or not. The face-to-face test was not structured in this fashion at all. Thus, if a candidate wanted to answer briefly, the interlocutor was obliged to ask many more questions than for candidates who offered further information out of their own initiative. This was an important criterion to the assessors, indicating that one possible interpretation for unwillingness to talk was difficulty to express oneself.

After the elicitation of the assessment constructs was completed, the assessors discussed the relationships between the criteria they had mentioned. Both of them recognised that several of the labels were related, some hierarchically. For instance, mother tongue interference (face-to-face test) was in fact one aspect of comprehensibility. Interference in turn was seen to be the product of the influence of the pronunciation, grammar and vocabulary of the mother tongue on the target language performance. However, mother tongue influence was only a rather marginal aspect of the criteria of grammar and vocabulary, or even pronunciation. The main content of these criteria was the amount of knowledge of the target language that the candidates were able to display.

During the elicitation of potential assessment constructs, the assessors had described the high and low ends of each scale as well as a middling performance. During the discussion on the relationships between the criteria, the assessors noticed that several of their descriptions of the low end resembled each other in one feature: there was little performance to be assessed. When the criterion was grammar, the small amount of performance meant that there were few grammatical structures in it in the first place, and most of the structures were used inappropriately. When it was vocabulary, the small amount of performance reflected the lack of lexical resources. When the criterion was willingness to talk, the description of the low end was "extremely unwilling to talk". This was considered to be a correct description of language skills at the low end of the scale, however. Similar conflation did not occur in the descriptors for the high end of the scale, though plenty of positive adjectives were used.

Another immediate observation by the assessors was that there were two fundamentally different seven-point scales used in the grid procedure: those assessing features of personality — confidence and, to a lesser extent, creativity — and those relating to linguistic features of the performances, the rest of the criteria. A further, tentative division was made between criteria that focused

on language knowledge (most clearly, vocabulary, grammar and pronunciation, hence also transfer), and those that focused on language use (rate of speech, collocations and idioms, and phrases). Of course, language knowledge can only be inferred from language use, and thus a strong interaction between the two would be natural. In judging the degree of association between the criterion labels, the assessors identified the pronunciation criteria, i.e. individual sounds and prosodic features, as fairly independent of the rest of the criteria.

These qualitative interpretations gained some support from the numerical analysis of the grid data, but unfortunately the size of the grids proved small for extensive analysis. In fact, many statisticians would say that the nine by twelve and ten by fifteen grids were far too small for statistical analysis; the error terms would be too large and interpretations shaky, because there are too few observations in relation to the number of variables. However, two analyses were conducted in order to investigate the most salient tendencies in the data: correlations between the criterion labels, and principal components analysis of the grids.

The correlations between the criterion labels were mostly in the .7 to .9 range (lowest .37, highest .96). For the tape-mediated test, the correlations were the lowest between confidence and grammar, accent and rate of speech, and accent and overall assessment. For the face-to-face test, the correlations were the lowest between the pronunciation of individual sounds and all the other criteria except grammar and prosodic features. The correlation tables are displayed in Appendix 14.

A Principal Components analysis of the grids yielded a first factor which explained 82% of the variance for each test. In the tape-mediated case the second factor accounted for 6% of the variance and the third factor for 4%, while in the case of the face-to-face test the second factor accounted for 8% of the variance and the third for 4%. This would strongly suggest a single-factor solution, but the alternative of multiple correlated factors was also possible. Both oblique and orthogonal rotations were performed on the SPSS for Windows. The outputs are presented in Appendix 15 for the tape-mediated test and Appendix 16 for the face-to-face test.

For the tape-mediated test, the "eigenvalues over 1" rule produced a single-factor solution with communalities of .69 to .92 on the criteria. The forced two-factor solution raised the communalities to range between .80 and

.95, and the percentage of variance explained by the factors to 88%. As the third factor would only have added a further 4% into the variance explained, two factors was probably the highest number that the data warranted.

Both the orthogonal and oblique solutions grouped the criteria into the same two factors. The first was a language use-oriented group with rate of speech, confidence, knowledge of phrases, and creativity in expressing expected content as one factor. The second was a language-oriented group with accent, prosodic features, grammar, knowledge of idioms, and vocabulary. In the orthogonal solution, all the variables loaded on both factors at over .3 and five of the nine criteria at over .5, which renders the solution implausible. The oblique rotation, however, produced a more clearly interpretable factor structure with none of the criteria loading on both factors at over .5 and four loading on both factors at over .3. The correlation between the two factors in the oblique rotation was .76.

An oblique rotation for the forced extraction of three factors for the tape-mediated grid further separated confidence from the presentation-oriented group, and again none of the criteria loaded on more than one factor at over .5, but the low eigenvalue of .36 and the small amount of additional variation explained by the third factor suggest that the solution may be spurious.

For the face-to-face test, the "eigenvalues over 1" rule also produced a single-factor solution, with communalities ranging between .36 and .90. Although the variance explained by the single factor was 82%, the low communality value for the pronunciation of individual sounds suggested that the solution was not ideal. The forced extraction of two factors increased the amount of variance explained to 90%, and the communalities to a range between .87 and .95, a much more satisfactory solution. Both the orthogonal and the oblique rotation separated pronunciation of individual sounds as the second factor, with the oblique solution again being more clearly interpretable. The correlation between the two factors was .47.

An oblique rotation for the forced extraction of three factors for the face-to-face grid retained pronunciation of individual sounds as a separate factor and further separated a presentation-oriented factor of turn length, rate of speech, and confidence from the rest of the criteria. Again, the low eigenvalue of .35 and the small increase in amount of variance explained make the interpretation tentative.

In addition to an overall factor, the statistical analysis of the grids suggested that the assessors may have paid attention to pronunciation as a separate criterion, at least as far as pronunciation of individual sounds is concerned. The assessors' qualitative interpretation of the grids concurred with this tendency. Interestingly, the result also concurs with de Jong's (1991) finding that pronunciation functions differently from other assessment criteria for speaking. The other qualitative interpretations of relationships between the assessment labels received scant support from the statistical analysis. It was possible that there was a presentation or language use oriented factor, which was fairly highly correlated with the language-oriented factor, but the evidence for this was very tentative and could have been a coincidence.

The reason that the principal components analyses strongly suggested a single-factor solution was that the assessments on the various criteria were highly correlated: when candidates were good in one respect, they also tended to be rather good in others. There were small variations within candidates between various criteria, but principal components analysis, which is performed on a group of data, is not intended for detecting a multitude of different small variations in individual cases. This variation became visible through a qualitative transformation suggested by Tagg (e-mail communication 23 February 1994).

The procedure consisted in the assessors going over the grids again, considering each candidate individually. They would first identify which overall level the candidate had achieved, then consider the performance on each of the criteria, and decide whether the performance was worse than, equal to, or better than an average performance *of the candidate's overall level* on that criterion. The transformed score became -, 0, or +. For instance, if the candidate were of level 3, a subscore series of 3, 3, 5, 4, 2, 3, 2, 3, 5 would most likely become 0, 0, +, +, -, 0, -, 0, +. The transformation was not mechanical, however: the assessors viewed the performance extracts again and made the transformation decisions on the basis of the extracts.

The transformations required a kind of norm-referenced thinking in that the real performance was being compared to an imaginary, average performance. The imaginary performance was rather easy to envision in the case of clearly language-related criteria, but very difficult in the case of confidence. It was difficult to decide what an average level of confidence would be with respect to a certain skill level. The transformed judgements on

confidence were thus not related to skill level but were overall assessments of whether the candidates in the extracts clearly displayed confidence, clearly displayed lack of confidence, or appeared average, considering this was a test situation.

The result of the judgmental transformation was a matrix of minuses, zeroes and plusses. The transformed grid for the tape-mediated test is shown in Figure 10, and the transformed grid for the face-to-face test in Figure 11. The criteria are organised according to possible groupings so that the pronunciation criteria are listed first, then the presentation-oriented criteria, and lastly the criteria that focus on grammar and vocabulary.

The columns of the matrices indicated each candidate's strengths and weaknesses. The resulting patterns rang true in the assessors' minds in relation to the performances, and this type of information bears strong resemblance to diagnostic feedback that classroom teachers can give to their students on the basis of speaking performances. There were no candidates for whom the patterns would be exactly the same; a reflection of the rich variation in the data. The variation highlights how many ways there are of being on any one skill level; much of this information is lost when a single skill level is reported as a score. The reverse side of the coin is that in order to report this information to the candidates, truly individual assessment information would be needed, while vague statements describing the average performance on each level are unlikely to fit any one individual. Before assessments on individual analytic criteria can be reported to candidates, the assessment system considering this alternative must ascertain that these assessments are reliable enough to be reported. With respect to the analytic criteria investigated here, it must be remembered that the whole procedure investigated the internal assessment systems of two assessors. Whether these impressions would be shared by other assessors, or whether the features that were named "really exist", was not investigated.

The transformed matrix can also be used to observe which criteria the assessors weighted the most in assigning overall skill level, and which criteria were less important. In the transformation, the candidates' performances on each criterion were being compared to an average performance of the same overall level, and if these were approximately equal, the code assigned was zero. Consequently, the criteria with the most zeroes on them were the ones which most strongly influenced the assessors' judgement on overall skill level.

Figure 10 Transformed grid for the tape-mediated test

E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	
-	+	0	0	+	+	+	0	0	-	+	-	clarity of pronunciation
0	-	-	-	0	-	-	-	-	-	+	-	prosodic features
+	0	0	0	-	-	0	0	0	+	-	-	rate of speech
-	0	+	-	0	-	+	+	0	0	0	0	confidence
0	0	0	+	-	0	+	0	0	-	-	-	creativity (in using the test rubric)
+	+	0	+	+	0	0	0	0	0	0	0	grammar
+	+	0	+	+	0	0	0	0	0	0	0	vocabulary
0	-	0	0	+	0	+	0	-	-	-	0	collocations and idioms (extended sp.)
0	0	0	+	0	0	+	0	0	0	-	0	phrases (simulated conv. and reactions)

On both tests, these were the vocabulary- and grammar-related criteria: grammar, vocabulary and knowledge of phrases (in the Simulated Conversations and Reactions in Situations tasks) on the tape-mediated test, and grammar, vocabulary, comprehensibility, and mother tongue interference in the case of the face-to-face test. In addition, rate of speech seemed to be an important criterion on the face-to-face test. Pronunciation criteria seemed to be the least weighted ones on both test modes.

Figure 11 Transformed grid for the face-to-face test

E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	
-	-	+	+	+	+	+	+	0	0	0	-	-	+	0	pronunciation / indiv. sounds
-	-	+	+	0	-	0	+	0	0	-	-	-	0	-	pronunciation / prosody
0	0	0	+	0	0	-	0	+	0	0	0	0	0	-	rate of speech
0	+	+	+	-	+	0	0	+	0	+	0	0	+	+	confidence
0	+	0	0	-	+	0	0	+	-	0	0	+	-	0	length of turn
0	0	0	+	-	0	+	0	0	0	0	0	0	0	-	transfer / interference
0	-	+	0	0	0	+	0	0	0	0	0	0	0	0	grammar
0	0	0	0	-	0	+	0	0	0	-	0	0	0	0	vocabulary
0	-	0	+	-	-	+	0	0	-	0	0	-	-	-	collocations and idioms
0	0	0	+	0	0	+	0	+	0	0	0	0	0	0	comprehensibility

6.4 Summary and discussion of results

The first research question in Study Three was whether there was a difference in the scores from the two tests. The answer to this question was a fairly clear no. The correlation between the scores was .85, a very high figure considering the small size of the data set. Although the scores from both tests appeared to be very reliable and the small difference potentially real, the numerical evidence was not strong enough for drawing the conclusion that the scores from the two tests would be different. This result concurred with all all but one of the previous studies where scores from tape-mediated and face-to-face tests had been investigated, and supported the first working hypothesis. The investigation of the second research question was thus motivated.

The first part of the second research question focused on the features of performance on which the scores from the two tests were based. The aim was simply to list the features that the assessors observed, as this would indicate what kinds of interpretations were warranted for the numerical scores obtained from the test. The results were that both language knowledge-oriented and language use-oriented criteria were used in the assessment of performances from both tests. It is notable that in addition to vocabulary and structures, the assessors also paid attention to control of collocations and phrasal knowledge among the linguistic criteria.

The assessors noticed features connected with pronunciation, but apart from rate of speech, appeared not to allow these aspects affect their assessment very much. Less language-based criteria such as creativity and willingness to talk were also attended to in the performances, and these aspects seemed to weigh more in the assessment of both tests than the candidates' pronunciation of individual sounds or the prosodic features of their speech. The inclusion of the less language-based criteria was in keeping with the view of language in the test, according to which the use of language automatically involves the speaker's person in social interaction. Assessing only knowledge of language in a test of speaking would be impossible according to this view.

Regarding the second part of the second research question, whether the features attended to were the same between the modes, two contrasting working hypotheses were offered. One stated that the criteria used for both tests would be the same and thus not very much affected by the kind of performance on which they were applied, the other that the criteria would be

slightly different depending on the mode that the performances came from. The results indicated that many of the features assessed in the two tests were the same, but that there were some differences as well.

Most of the differences in assessment foci were related to the differences between the test modes already encountered in Studies One and Two: interactiveness and task structuring. In the non-interactive tape-mediated test where the content of most replies was guided, creativity in expressing the expected content was valued, while problems expressing the content were diagnosed to be ones of vocabulary, grammar, or perhaps knowledge of phrases. On the face-to-face test, the assessors did not know what the candidates wanted to say, and in focusing on finding this out they sometimes noticed transfer from other languages. The features noticed were probably very similar to the ones that came up with the tape-mediated test, but since the assessors were focused on meaning, they categorised this as transfer.

Willingness to talk was the most clearly mode-related criterion named by the assessors, influencing judgements on the face-to-face test. The expectation with conversational interview as a test appears to be that the candidates should take initiative and be willing to perform, and if they do not, they may be assessed lower. Based on the results of Study Three, similar expectations do not seem to hold for the kind of tape-mediated test used here.

Neither of the working hypotheses on the second research question thus received full support in Study Three. The assessment constructs appeared to be similar but not exactly the same between the two modes, taking account of mode-specific features of performances when they were relevant. The differences seemed to be not so much between what features of language were attended to as between what it was reasonable to expect in each test, given the tasks and conditions of each.

It appears that there were small contextual differences between the assessment constructs connected to the two tests. Regardless of the differences, however, the test-specific features attended to seemed to co-occur in individual candidates, such that the final assessment was highly likely to be the same from both tests. Two immediate cautions are in order, though. Firstly, the explanations for the cases where the scores from the two modes were different were most often mode-related and had to do with test-taking strategies and affective reactions. Secondly, the generalisability of either score to performance in real life was not investigated in the present study at all.

6.5 Discussion of methods

In focusing on the interpretation of assessment results, both quantitative and qualitative methods were used in Study Three. The choice of the particular methods used, correlation and the Repertory Grid technique, was based on the size of the data set and the nature of the research questions. These in turn were determined by the stage of development of the project that provided the data. The following discussion focuses on the possible effects of the choice of methods on the results.

Regarding the correlation of the scores from the two modes, the distribution of the scores was skewed, while one of the preconditions for using the Pearson correlation is that the scores are normally distributed. However, the skewness was not considered a serious threat to the validity of the method, because there were no extreme scores in the data set. If anything, the concentration of the scores at the top end of the scale should lower rather than raise the resulting coefficient. The coefficient reported can thus be regarded as a lower bound for the correlation between the scores.

The aim with the qualitative assessment analysis was to distinguish aspects that the assessors regarded important in the holistic concept of proficiency. Repertory Grid was chosen as the technique because it suits the particular purpose of distinguishing dimensions within an overall concept. The initial list of assessment constructs was based on a small sample of performances, however, and it can be questioned whether the method had a substantial effect on the results. The possible constructs that are named during the elicitation are directly related to the material that has been used to elicit them, and should not be considered independent until proven so in subsequent analysis. The labels are also individual, and only connected to the two assessors' perceptions from whom they were elicited.

The quantitative analysis of the grids was doubtful, because both data matrices were small. The qualitative interpretations, though subjective, were considered useful in pointing out some of the main connections between the proposed criteria. According to the principles of Personal Construct Psychology, the best experts to interpret data collected through this method are the informants themselves. The present writer's role as one of the assessors can thus be considered an advantage, though the generalisability of the findings can be questioned. Generalisation without extensive individual case

studies would indeed be contrary to the central idea of subjectivity of reality in Personal Construct Psychology, where the method originated.

The assessors found the elicitation procedure very rewarding. Both the elicitation and the subsequent analysis helped them become conscious of at least many of the criteria they used in assessing oral proficiency. If not as sole method, this procedure is a promising candidate for at least one method through which operational assessment foci can be investigated. Even more appropriately, this method could be used as part of assessor training, because the consciousness-raising aspect of the procedure was strong. Furthermore, if the participants were trained assessors, feedback could be collected on how any analytic criteria that the system may provide are applied in practice by the assessors. This would benefit both the assessors and the test system.

The qualitative transformation of the scales into minuses, zeroes, and plusses, suggested to the present author by a psychologist who has used the technique extensively, was highly enlightening. The fact that this form of the grid corresponded with the kinds of diagnostic feedback that the assessors might have provided for the candidates lent support to the coherence of the initial assessments made by the assessors, and revealed a further potential use of the technique. Its usefulness for observing which criteria were the most important for the assessors was a further reason for suggesting the grid for assessor training.

Apart from difficulties with interpretation, a limitation of the repertory grid technique for investigating what assessors pay attention to is that it entirely breaks with the reality of normal assessment procedure. Analyses of the normal procedure, made for instance through talk-aloud protocols and retrospective interviews, could be conducted in addition to or instead of the grid procedure to investigate the practice of assessment of speaking.

7 CONCLUSION

The results of the present study indicated that there were both similarities and differences between the tape-mediated and the face-to-face test investigated. The clearest similarity between the tests was the scores produced. Many of the

criteria which the assessors paid attention to were also similar across the two modes. Moreover, the words, phrases, and grammatical structures used by the candidates in the two tests were the same. The clearest difference between the two modes was the presence versus absence of real-time human interaction. This was perceived as a highly salient difference by the candidates, and some indications of this difference were given in the linguistic analysis of the performances. For instance, the turns spoken on the face-to-face test were much shorter, because there was another person there to react to what the candidate was saying. There was also a clear difference in task structuring, such that the tasks on the tape-mediated test were much more highly structured. Lastly, there was a difference in what the assessors considered it reasonable to expect in a performance, given the tasks and the conditions under which they were performed. They tended to expect compliance with instructions on the tape-mediated test, and willingness to show ability on the face-to-face test.

A possible explanation for the similarity of the scores is that the connection between the tests through the words and structures of the language tested is strong enough to guarantee close comparability, as long as the performances in both modes are extensive enough. It could also be that the test developers' aim to develop matching tests, each of which suits its mode, had succeeded, and scores from one test could be used to predict scores from the other, because where different aspects of language ability were tested, these aspects develop simultaneously in individual language learners. If they did well in one test, they were therefore also likely to do well in the other. A further possibility was that the assessors buffer the possible variability in performances through compensating for the effects of the test tasks. They would thus not be assessing the performances as such, but what the performances indicate to them of the candidates' ability to use the language outside the test context. All of these alternatives are plausible hypotheses which could be investigated in an analysis of what assessors do when assessing test performances from the two modes.

The linguistic forms which the candidates used on the two modes were found to be highly similar. Though there were some statistically significant differences in the frequency of various parts of speech and individual words between the two tests, very few of the differences were meaningful. Evidence that pointed to the similarity of linguistic forms between the tests was also offered in the study on participant perceptions in the form of one

interviewee's comments. On reviewing her performances, she recognised the similarities in her language in both modes. Furthermore, some of the results of assessment analysis supported the similarity of linguistic forms between the two modes. The criteria that focused on linguistic forms — grammatical correctness, sophistication of vocabulary, and knowledge of phrases, collocations and idioms — were the same between the two modes.

The similarity between the modes regarding linguistic form is rather natural. Whatever needs to be accomplished in a test of speaking, it requires the use of the words, phrases and structures of the language, and if the performances are hundreds of words long, it is likely that similarities should appear.

The differences in language use between the two tests appeared to result from one key difference between the tests, which was the presence or absence of real-time human interaction. This was a highly salient difference to the candidates, who overwhelmingly preferred the face-to-face test because language use in it felt authentic while in the tape-mediated test did not. An important aspect of this was the possiblity for checking that they have understood and have been understood. The results of Task Analysis also pointed to differences in language use through identifying the ways in which the difference appeared in the settings, rubrics, instructions, input, and expected responses in the two tests. The key explanation for the differences was that one test instrument only took its final form during test administration, while the other existed well beforehand and did not alter in administration. Though most of the results of Performance Analysis suggested similarities rather than differences, two of the three mode-related differences that were found were differences in language use: discourse particles and hesitation markers. The distribution of both of these showed that at least some candidates used language differently in the two modes.

When the difference in language use is considered from the testing point of view, it can be construed as a difference in task structuring. This covers not only the performances elicited from the candidates but the nature of the elicitation instruments as well. Task structuring refers to the degree of control that the test developer exerts on the use of language in the test. The tasks on the tape-mediated test were highly structured, in that the test developers chose the simulated language use contexts as well as the roles of the interactants, and often specified the content of the expected responses. Even

the maximum lengths of individual turns of response were determined. The task of the candidates was to comply with the structure of the elicitation instrument and provide the performance. The specific task-by-task instructions were included to help the candidates follow the elicitation. The face-to-face test was much more loosely structured. The general topics and overall duration of the test were defined, as well as the two interaction types of peer discussion and individual interview, but the candidates had a fair amount of room to negotiate the specifics of the tasks within the overall framework. The interlocutors had some prepared prompts and probes, which they used as they deemed relevant on the basis of what the candidates had already said. The interlocutors thus mediated the structure of the face-to-face test through their conduct in the test situation, and the task of the candidates was to play along and help construct the discourse of the test.

When seen through the framework of task structuring, the candidate comments in Study One which were not directly related to real-time human communication can be identified as their negative reaction to the high structuring. As regards Task Analysis, task structuring can be used to explain why the instructions, input and expected response were different between the two modes. As Performance Analysis was conducted in Study Two, task structuring only figured in explaining why the functions elicited by the tape-mediated test were similar in all candidates' performances, but if the interplay between the elicitation instrument and the performance elicited were analysed, task structuring would undoubtedly figure more prominently in the explanation of the results. From the point of view of assessment, task structuring explains why the assessors were better able to predict the content of the candidates' responses in the tape-mediated test. Thus it points out an important difference in what is assessed in the performances on the two modes.

The difference in task structuring had clear implications for the assessment constructs of the two tests. The tape-mediated test measured whether and how well the candidates *could do* what the developers considered significant indicators of progress in language learning. The construct of the tape-mediated test was thus determined by the developers' choice of prompts and tasks. The benefit of the high structuring was that the items tested were the same for all candidates. On the face-to-face test, the way that the candidates displayed their ability depended to a large extent on what they *would do* in the

test situation. The construct measured on the face-to-face test was thus jointly determined by the interlocutor and the candidates, and there was little guarantee that the construct summarised by the score would be exactly the same in different tests.

It was possible that the difference in test focus between what the candidates *could* do and what they *would* do was also felt in the expectations of the assessors. A study focusing on the reasons why an assessor gives a particular score, or which features an assessor notes as indicators of ability while listening to a performance, might bring new insights into whether there is some difference in scoring between the two test modes.

The overall conclusion from the three studies was that while there was plenty of similarity between the tests, the constructs behind the two tests were to an extent different. Whether the tests were comparable is thus not easy to judge. Especially if comparability is interpreted to mean exchangeability, the yes/no decision depends on how important the differences are to the one asking the question.

If the scores are the most important concern for the decision maker, the overall conclusion of the present study regarding exchangeability of the two tests was affirmative. In this case, the choice of which mode should be used would depend on availability of resources and practicality. Other matters such as candidate preference may influence the decision, particularly if the existence of the test is dependent on the number of candidates choosing it and if alternatives exist. If the mode of the test is important from a language policy point of view, the administrators may choose the tape-mediated test to stress reliability and control, or the face-to-face test in the hope that it might encourage face-to-face discussion as a form of classroom exercise. The choice may not have the desired effects, however (cf. Alderson and Wall, 1993), and should always be investigated empirically rather than stated.

The presesnt study offered some practical conclusions for test development. Since the representativeness of the sample elicited on a tape-mediated test depends entirely on the content of the test instrument, it is important to cover a sufficient number of different language use situations which appropriately differentiate between the candidates. To ameliorate the negative affective reactions to the tape-mediated test, some system ought to be developed to allow candidates to familiarise themselves with the physical context of the language laboratory and the kinds of tasks that they are likely

to meet there. In the case of a face-to-face test of speaking, it is important that there is enough structuring in the test to ensure approximate comparability between different administrations of the test. It is also very important that the interlocutors share the test developers' view of what indicates proficiency in candidate performances so that they can use the relevant prompts consistently. By analysing the discourse from the two modes, hypotheses regarding the ways that proficiency appears in test performance could be more closely related to actual performance data rather than expectations based on theory. The discourse data could be used in developing assessment scales as well, particularly if features of performance are investigated together with assessor perceptions.

7.1 Limitations of the present study

From the point of view of generalisability of the results to the testing of speaking in other contexts, the present study has several limitations. The most evident of these concerns the numerical representativeness of the data. The number of candidates who participated in the tests was small, and even fewer performances were involved when the language elicited was analysed and when possible assessment constructs were investigated. The number of assessors involved was two, and several of the analyses were only conducted by the present writer. The small numbers increase the likelihood that the results and interpretations were heavily dependent on the specific characteristics of the data that were available, and might have been different if different candidates, assessors or analysts had been involved. The analyses that are deemed useful should thus be repeated with more representative data sets.

Some of the results of the present study may also have been affected by the fact that the data came from a pilot test. The tests contained some design flaws, and the atmosphere and the administration conditions of the pilot test were different from that of an operational test. That the participants were volunteers probably affected the distribution of scores, in that potential participants who thought the test might be too demanding for them were less likely to participate. The effects of the pilot condition are difficult to assess in the data, because comparable data from an operational test administration do not exist.

Another limitation regarding the generalisability of the results of the present study is that only two tests of speaking were investigated. The connections between the candidates' performance in the tests and their performance outside test situations were not investigated at all.

The methodological limitations of the present study were discussed in detail in Chapters 4.6, 5.5, and 6.5. Only the main shortcomings will be repeated here. Several of the methods of analysis used in the present study were exploratory, and particularly as regards analysis of test discourse, these were not necessarily the most fruitful ones. The elicitation instruments and the performances elicited were analysed separately, while a more fruitful way of looking at this might be to analyse the interplay between these. Such research would provide data on the variability in elicitation in face-to-face tests, such as studies by Ross and Berwick (1992), Lazaraton (1992, 1996), and Brown and Lumley (1996) have pointed out. It would also specify some ways in which candidates react to different prompts, studies of which have not appeared in the literature on language assessment. The present data set was too small and contained too much variation irrelevant to the central problem to warrant such analyses, but with a more representative set of data, this would be a very interesting aspect to pursue.

Like previous studies comparing tape-mediated and face-to-face tests of speaking, the present study could be classified as a validation study. Its questions were thus related to the particular test investigated rather than the testing of speaking in general. In combination with the limitations in the size of the data set, the answers as such were not generalisable to other contexts. However, the implications for further research might fit other contexts of testing speaking, and the experiences with methods of analysis might prove useful for other researchers.

7.2 Implications for further research

In the course of summarising the results and discussing the limitations of the present study, some directions for further study were already suggested. The common core in these suggestions is the connection between test discourse and assessment. It would be important to analyse the whole of the test discourse rather than investigating elicitation or performance in isolation. This was

evident from the results of the present study, and has also been suggested by Lazaraton (1996:167), who has investigated the language of interlocutors in live tests of speaking.

Test discourse as such is a worthy target for discourse analysis, as tests of speaking are clearly distinguishable events with distinctive characteristics such as the roles of the participants and the distribution of power, as well as the schematic structure of the event. The value of such research is that comparisons between test discourse and "real life" discourse can be made. However, the particular feature that makes the discourse analysis of test situations relevant to assessment is the combination of discourse information with assessment information. Only through making the connection will analyses of test discourse assist in the interpretation and correct use of test scores.

For making the connection between test discourse and assessment information, Lazaraton (1996:167) suggests statistical treatment of assessment data in combination with discourse analysis of test language. Such a plan is promising, but requires vast amounts of work. A large number of performances would have to be analysed for a large number of linguistic features, and several assessments of each of the performances would have to be obtained. After accomplishing all of this, it would be possible to model the importance of each of the discourse features analysed for the assessment, and to estimate how much of the variation can be explained through these features. The success of the analysis would depend to a large extent on the importance of the discourse features investigated for the assessment process.

Another way of tackling the problem would be to use the introspection of assessors as a guide in the search for features of performance which should be analysed. Trusting this guidance would mean that only features which the assessors can consciously reflect on would be analysed, but subsequent analysis might reveal different ways in which these features appear in the test discourse. An added advantage would be that the results of the assessment analysis could be used in reporting the results of the performance analysis, so that the concepts used in the description of the performances would be comprehensible to assessors. The assessment data could be collected through talk-aloud protocols and retrospective interviews. Douglas (1994) conducted a small-scale study where he analysed assessment quantitatively and performances linguistically. The results were that there was little connection

between the linguistic features analysed in the study and the analytic or overall scores given by the assessors. Douglas's (1994:135-136) conclusion is recommending qualitative analysis of assessors' judgement through think-aloud studies. This would increase our understanding of the bases of assessor judgements.

Both of the above approaches see the testing of speaking as a two-stage event, where one stage is the discourse during the administration of the test, while the other is a less easily observable interaction between the assessor and the test discourse, mediated by perceptions of proficiency. The two stages can be concurrent if assessments are made during the live situation, almost concurrent if the final stages of the assessment are completed after the discourse event is finished, or consecutive, if assessment only begins after the discourse event is over. The difference between the above approaches is that the first treats the assessment stage as an unknowable, and attempts to explain its workings through analysing the input — test discourse — and output — the score. The second considers the assessment stage analysable, and uses process data to get at the relevant concepts in the event. Both approaches hold promise, and could be taken in parallel projects. The results of both would be useful for writing assessment scales, training assessors, and training interlocutors.

Including both the discourse and the assessment stages in one design is difficult because there are so many dependent variables. One way of simplifying the design might be to use both of the above approaches to limit the object of study. Based on numerical assessment results only, proficiency level could be taken as the independent variable. Assessor perceptions could then be used to guide the analysis of test performances, and the results could be used to clarify what the proficiency level means in performance terms. A further stage, relating this line of research to language learning, would be to compare the features in the test discourse and the perceptions of the assessors across several proficiency levels.

REFERENCES

Alderson, J. Charles. 1988. New procedures for validating proficiency tests of ESP? *Theory and Practice. Language Testing* 5, 220-232.

Alderson, J. Charles and Dianne Wall 1993. Does washback exist? *Applied Linguistics* 14, 115-129.

Alderson, J. Charles, Caroline Clapham, and Dianne Wall 1995. *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Austin, J. L. 1962. *How to do things with words.* Oxford: Clarendon.

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

Bachman, Lyle F. 1991. What Does Language Testing Have to Offer? *TESOL Quarterly* 25, 671-704.

Bachman, Lyle F. and Adrian S. Palmer 1996. *Language Testing in Practice.* Oxford: Oxford University Press.

Bachman, Lyle F., Antony Kunnan, Swathi Vannirajan, and Brian Lynch 1988. Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries. *Language Testing* 5, 128-159.

Bachman, Lyle F., Fred Davidson, Katherine Ryan, and Inn-Chull Choi 1995. *An investigation into the comparability of two tests of English as a foreign language. The Cambridge-TOEFL Comparability Study.* UCLES Studies in Language Testing 1. Cambridge: Cambridge University Press.

Baker, Eva 1974. Beyond Objectives: Domain-referenced tests for evaluation and instructional improvement. *Educational Tehcnology* 14, 10-16.

Bannister, D. and J. M. M. Mair 1968. *The evaluation of personal constructs.* London: Academic Press.

Brown, Annie 1993. The role of test-taker feedback in the test development process: test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing* 10, 277-303.

Brown, Annie and Tom Lumley 1996. Interviewer variability in specific-purpose language performance tests. Paper presented at the 18th Language Testing Research Colloquium in Tampere, Finland, July 31 — August 3, 1996. Manuscript.

Canale, Michael 1983. On some dimensions of language proficiency. In Oller, J (ed.) 1983. *Issues in Language Testing Research.* Rowley, Mass.: Newbury House, 333-342.

Canale, Michael and Merrill Swain 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1, 1-47.

Carroll, Brendan J. and Richard West 1989. *ESU Framework. Performance Scales for English Language Examinations.* London: Longman.

Davidson, Fred and Brian Lynch 1993. Criterion-Referenced Language Test Development: A Prolegomenon. In Huhta, Sajavaara and Takala (eds.) 1993, 73-89.

van Ek, J. A. and L. G. Alexander 1975. *Threshold Level English.* Oxford: Pergamon Press.

Fortus, Ruth, Rikki Coriat, and Susan Fund 1995. Prediction of item difficulty in the English section of the Israeli psychometric entrance

test. Paper presented at the 17th annual Language Testing Resesarch Colloquium, Long Beach, CA 24-27 March 1995.

Fransella, Fay and Don Bannister 1977. *A Manual for Repertory Grid Technique.* London: Academic Press.

Fulcher, Glenn 1996. Testing tasks: issues in task design and the group oral. *Language Testing* 13, 23-51.

Garside, Roger, Geoffrey Leech, and Geoffrey Sampson, 1987. *The computational analysis of English: a corpus-based approach.* London: Longman.

Halliday, M. A. K. 1976. The form of a functional grammar. In Kress, G. (ed.), *Halliday: System and function in language.* Oxford: Oxford University Press.

Halliday, M. A. K. 1994. *An introduction to functional grammar.* Second edition. London: Edward Arnold.

Halvari, Anu 1994. Two communicative tests in comparison: What do the test desults show and what do test-takers say about the tests? Paper presented at the 1994 Language Testing Research Colloquium in Washington, DC, March 4-7, 1994.

Halvari, Anu 1996. Two communicative language tests in comparison. A study of the National Certificate and the ICC test. Unpublished Pro Gradu thesis, University of Jyväskylä.

Hatch, Evelyn and Anne Lazaraton 1991. *The Research Manual. Design and Statistics for Applied Linguistics.* Boston, Mass.: Heinle & Heinle.

Huhta, Ari, Kari Sajavaara, and Sauli Takala 1993. Recent developments in national examinations in Finland. In Huhta, Sajavaara and Takala (eds.) 1993, 138-159.

Huhta, Ari, Kari Sajavaara, and Sauli Takala (eds.) 1993. *Language Testing: New Openings.* University of Jyväskylä, Institute for Educational Research.

Hoekje, Barbara and Kimberly Linnell 1994. "Authenticity" in language testing: evaluating spoken language tests for international teaching assistants. *TESOL Quarterly* 28, 103-125.

de Jong, John 1991. Defining a variable of foreign language ability. An application of item response theory. Doctoral dissertation, University of Twente.

Kelly, George 1955. *The Psychology of Personal Constructs*, vols I and II. New York: Norton.

Kenyon, Dorry 1991. Using generalizability theory to study the parallel test reliability of two alternative testing methods: The case of the OPI and the SOPI. Poster session presented at the 13th Annual Language Testing Research Colloquium in Princeton, NJ, March 21, 1991.

Kenyon, Dorry and Charles W. Stansfield 1991. A method for improving tasks onperformance assessment through field testing. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL., April 1991.

Lazaraton, Anne 1992. The structural organization of a language interview: a conversation analytic perspective. *System* 20, 373-386. Pergamon Press.

Lazaraton, Anne 1996. Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing* 13, 151-172. Edward Arnold.

Lowe, Pardee and Charles W. Stansfield 1988. *Second language proficiency assessemnt.* Englewood Cliffs, NJ: Prentice Hall.

MacWhinney, Brian 1989. *The CHILDES project. Tools for analysing talk.* Hillsdale: Lawrence Erlbaum.

Manninen, Seija 1984. Communication Apprehension. A Study of the Factors Contributing to Anxiety Experienced by Finnish Speakers of English. Unpublished MA thesis, University of Jyväskylä.

Mathison, Sandra 1998. Why Triangulate? *Educational Researcher,* March 1988, 13-17.

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

Messick, Samuel 1988. The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity.* Hillsdale, NJ: Lawrence Erlbaum, 33-45.

Messick, Samuel 1989. Validity. In R.L.Linn (Ed.) *Educational Measurement.* Third edition. New York: American Council on Education / McMillan, 13-103.

Milanovic, Michael and Nick Saville 1995. Anything you can do, I can do better! The ALTE Can Do Statements Project. Paper delivered at TESOL 1995 in Long Beach, California, March 28 - April 1, 1995.

Millman, J. 1974. Criterion-referenced Measurement. In W. J. Popham (Ed.) *Evaluation in Education: Current Applications.* Berkeley: McCutchan.

Munby, John 1978. *Communicative Syllabus Design.* Cambridge: Cambridge University Press.

National Board of Education, 1995. *The framework of the Finnish National Certificate.* Helsinki: National Board of Education.

North, Brian 1993. *The Development of Descriptors on Scales of Language Proficiency.* Washington, D.C.: The National Foreign Language Center.

O'Loughlin, Kieran 1995. Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing* 12, 216-237.

O'Loughlin, Kieran 1996. The comparability of direct and semi-direct speaking tests: a case study. Unpublished PhD thesis, University of Melbourne.

Opetushallitus, 1997. *Yleiset kielitutkinnot ja kieltenopetus.* Helsinki: Opetushallitus.

Oppenheim, A. N. 1992. *Questionnaire Design, Interviewing and Attitude Measurement.* New Edition. London and New York: Pinter Publishers.

Osburn, H. G. 1968. Item sampling for achievement testing. *Educational and Psychological Measurement* 28, 95-104.

Pollitt, Alastair and Neil L. Murray 1993. What raters really pay attention to. Paper presented at the 15th Language Testing Research Colloquium, Cambridge and Arnhem, 2-7 August 1993.

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

Purves, A. C., A. Söter, S. Takala, and A. Vähäpassi 1984. Towards a domain-referenced system for classifying composition assignments. *Research in the Teaching of English,* 18, 385-416.

Ross, Steven and Richard Berwick 1992. The discourse of accommodation in oral proficiency examinations. *Studies in Second Language Acquisition* 14, 159-76.

Sajavaara, Kari 1992. Designing Tests to Match the Needs of the Workplace. In E. Shohamy, A. R. Walton and C. A. Morfitt (eds.) *Language Assessment for Feedback: Testing and Other Strategies.* National Foreign Language Center. Dubuque, IO: Kendall/Hunt Publishing Company, 122-144.

Scott, Mary Lee 1986. Student affective reactions to oral language tests. *Language Testing* 3, 99-118.

Searle, John R. 1969. *Speech acts: an essay in the philosophy of language.* Cambridge: Cambridge University Press.

Shohamy, Elana 1982. Predicting speaking proficiency from cloze tests: theoretical and practical considerations for test substitutions. *Applied Linguistics* 3, 161-171.

Shohamy, Elana 1994. The validity of direct versus semi-direct oral tests. *Language Testing* 11, 99-123.

Shohamy, E., Gordon, C., Kenyon, D., and Stansfield, C. 1989. The development and validation of a semi-direct test for assessing oral proficiency in Hebrew. *Bulletin of Hebrew Higher Education* 4, 4-9.

Shohamy, Elana, Smadar Donitsa-Schmidt, and Ronit Waizer 1993 / no date. The effect of the elicitation mode on the language samples obtained on oral tests. Revised version of a paper presented at the Language Testing Research Colloquium, Cambridge, England, August 1993.

Stansfield, Charles W. 1990. An evaluation of simulated oral proficiency interviews as measures of spoken language proficiency. In: *Georgetown Roundtable on Languages and Linguistics 1990.* Washington, DC: Georgetown University Press, 228-234.

Stansfield, Charles W. 1991. A comparative analysis of simulated and direct oral proficiency interviews. In Anivan, Sarinee (ed), *Current developments in language testing.* Singapore, Regional English Language Center, 199-209.

Stansfield, Charles W. and Dorry Kenyon 1988. Development of the Portuguese Speaking Test. Washington, DC: Center for Applied Linguistics. (ERIC Document Reproduction Service No. ED 296 589).

Stansfield, Charles W. and Dorry Kenyon 1989. Development of the Hausa, Hebrew, and Indonesian Speaking Tests. Washington, DC: Center for Applied Linguistics.

Stansfield, Charles W. and Dorry Kenyon 1992. The development and validation of a Simulated Oral Proficiency Interview. In *The Modern Language Journal,* 76, 129-141.

Stansfield, C., D. Kenyon, R. Paiva, F. Doyle, I. Ulsh, and M. A. Cowles 1990. The Development and Validation of the Portuguese Speaking Test. *Hispania* 73, 641-651.

Tagg, Steven 1994. E-mail communication 23 April 1994. Available from sluoma@cc.jyu.fi.

Thomas, Jenny and Andrew Wilson 1995. Methodologies for studying a corpus of doctor-patient interaction. In Thomas, J. and M. Short (Eds.) *Using Corpora for Language Research.* London: Longman, 92-109.

Weir, Cyril J. 1988. *Communicative Language Testing.* Volume 11, Exeter Linguistic Studies. University of Exeter.

Wigglesworth, Gillian and Kieran O'Loughlin 1993. An investigation into the comparability of direct and semi-direct versions of an oral interaction test in English. *Melbourne Papers in Language Testing* 2, 56-67. University of Melbourne.

Wilkins, D.A. 1976. *Notional Syllabuses.* Oxford: Oxford University Press.

Wilson, Andrew and Geoffrey Leech 1993. Automatic Content Analysis and the Stylistic Analysis of Prose Literature. In: *Revue Informatique et Statistique dans les Sciences humaines* 1, 219-234.

Yli-Renko, Kaarina 1989. *Suullinen kielitaito ja sen mittaaminen lukion päättövaiheessa.* Tutkimuksia 72. Helsingin yliopiston opettajankoulutuslaitos. Helsinki.

Zeidner, Moshe and Marsha Bensoussan 1988. College students' reactions towards written versus oral tests of English as a Foreign Language. *Language Testing* 5, 100-114.

Appendix 1 The overall skill level descriptions of the National Certificates

8 Communicates naturally, effectively and appropriately even in demanding oral and written tasks and situations. Fluent and in many ways native-like. Occasional problems with subtle stylistic distinctions and idioms.

7 Communicates rather effectively and appropriately even in many demanding oral and written tasks and situations. Usage is quite versatile and fluent. Slight inaccuracies and influence from other languages are not intrusive. Understands with ease both writing and speech on demanding subjects.

6 Communicates appropriately in familiar oral and written tasks and situations and manages adequately even in socially or lexically demanding situations. Occasional inaccuracies and inadequacies, which nevertheless seldom lead to misunderstandings. On demanding subjects may occasionally need repetition or consulting a dictionary.

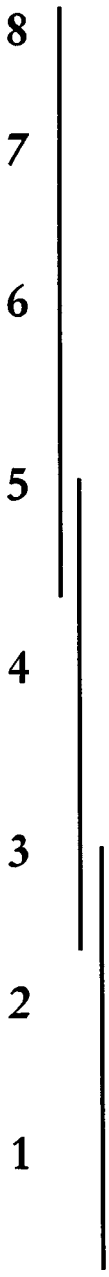
5 Manages familiar oral and written tasks and situations related to work and freetime rather well. Knows the basic structures and vocabulary and only occasionally needs to resort to requesting repetition or using a dictionary. Inaccuracies or interference from other languages occasionally hinder communication.

4 Manages familiar oral and written tasks and situations related to work and freetime. Interference from other languages can be intrusive. Vocabulary, grammar and fluency generally adequate but variable. Consulting a dictionary may sometimes be necessary for understanding main points of ordinary text, for instance, a newspaper article.

3 Manages to communicate in the most familiar oral and written tasks and situations but new situations caus communication problems. Understands slow and careful speech and can normally understand the gist of an easy text, such as a short newspaper article.

2 Manages to communicate in simple and routine tasks and situations. With the help of a dictionary can understand simple written messages and without one can get the gist. Limited language proficiency causes frequent breakdowns and misunderstandings in non-routine situations.

1 Knowledge of language is sufficient for coping with the simplest oral and written tasks and situations. Can understand the topic in newspaper articles and conversations that deal with familiar subjects. Knows some of the basic structures of the language.



The vertical lines on the left illustrate the scope of the three test levels: Basic (1-3), Intermediate (3-5) and Advanced (5-8).

Appendix 2 Taxonomy of topic categories in the National Certificates

Topics	Basic level	Intermediate level	Advanced level
A. Personal identification The candidate has to be able to tell about him/herself or others and to ask them about - personal information: name (also spelled), address, telephone number, sex, age, place and date of birth, marital status - nationality, languages (foreign language proficiency) - occupation, place of employment, unemployment - family and friends, members of family - hobbies	- at least by naming - briefly - using simple expressions and phrases	With more thoroughness and variation than on the previous level. Also: - telling about educational background: field, duration, level, degrees - telling about pastime activities - disposition, appearance	More extensively and profoundly than on the previous level.
B. Home and living The candidate has to be able to discuss living conditions, e.g. - blocks of flats, terraced houses., houses, rooms - furniture - place of residence, home town	- at least by naming - briefly - using simple expressions and phrases	With more thoroughness and variation than on the previous level. Also: - describing surroundings (town, countryside), landscape, pleasantness, recreational possibilities, connections	More extensively and profoundly than on the previous level. Also: - the social aspects of home and living, such as housing problems and policy
C. Work The candidate has to be able to tell about his work and ask about others' work (pensioners – former work or present daily routine), e.g. - occupation, place of employment, unemployment - working experience, description of duties - daily routine: working hours, beginning, end, phases, breaks, leisure time - annual working cycle: working periods, vacations, the impact of seasonal variation - target language requirements set by occupation	- at least by naming - briefly - using simple expressions and phrases	With more thoroughness and variation than on the previous level. Also: - pay, prestige, atmosphere, job satisfaction - job-related social benefits and looking after them (briefly) - occupation-related matters	General trends and topics in addition to the previously mentioned, e.g. - meetings and negotiations, making contacts - representing others , such as interpreting, negotiations
D. Society The candidate has to be able to convey and receive information about his or other societies, e.g. - economy - education - administration, political landscape - history - religion - language policy - equality - social security	Mainly understanding and telling briefly about basic facts using basic vocabulary	Understanding larger entities and describing the conditions of his native country in general terms.	Attains a more abstract level of analysis and comparison based on general knowledge.

Topics	Basic level	Intermediate level	Advanced level
E. Environment and geography The candidate has to be able to describe his country and its environmental situation. Must also know the main features of the target language countries and be able to find out more about them. e.g. - nature, environment and environmental conservation - population, economic life - weather , climate and seasons	Brief and simple description	Understanding larger entities and describing the conditions of his native country in general terms.	Attains a more abstract level of analysis and comparison based on general knowledge.
F. Personal relationships The candidate has to be able to communicate in social situations, e.g. - meeting people, introductions: personal information - extending and receiving invitations, acting as a host/hostess or a guest - polite conversation, such as starting or ending a conversation, thanking and saying goodbye	In familiar situations - using simple expressions and phrases - mainly oral	Semi-formal situations using linguistic and cultural conventions in addition to the previously mentioned Furthermore, - short written communications to both known and unknown recipients, e.g., greetings, invitations, inquiring about accommodation, and answering to them.	Also official situations as a representative of his employer or country .
G. Everyday life The candidate has to be able to describe his everyday life and enquire about others' way of living, news events and culture: - daily routines - public and private transport - shopping and using services - consumables and foodstuffs, preparing meals - eating in and out - measurements, such as currencies, prices, sizes, weights - (comprehension of) topical events - personal interests and hobbies, such as culture and sports - mass media	In broad outline - at least by naming - personal matters - using simple expressions and phrases	With more thoroughness and variation than on the previous level. Also on matters outside personal experience.	Attains a more abstract level of analysis and comparison.
H. Travel The candidate has to be able to handle travel-related situations, e.g., - arrival and departure: dealing with passport and customs officials (length and purpose of visit, goods to declare), exchanging money - announcements and signs in vehicles, in the customs, in public buildings and on the border. - mass transit: enquiring about arrivals and departures, purchasing a ticket - hiring and maintenance of vehicles - excursions, business visits and training sessions - getting information about destination and accommodation	Mainly understanding and telling briefly about basic facts using basic vocabulary	Also assisting others; in addition, - selecting the route and means of transport - traffic regulations, traffic signs and directions - introduction of destination and giving advice - common problem situations: lost luggage, lost-and-found offices, reporting thefts	Taking responsibility and representing fellow passengers in addition to the previously mentioned

Topics	Basic level	Intermediate level	Advanced level
I. Health and well-being The candidate has to be able to communicate in target language of e.g. - physical status: feeling sick, hunger, thirst, fatigue; states of mind; enquiring about others' well-being - accidents and diseases	- at least by naming - briefly - using simple expressions and phrases	More thoroughly and also e.g. - enquiring about medical services, making appointments, answering the doctor's questions, describing the malady - acting in emergencies: contacting the fire department and the police; calling an ambulance; giving relevant information on the phone; using an emergency phone	More extensively and profoundly than on the previous level. Also: - national health services - health education

Source: National Board of Education (1995)

Appendix 3 Taxonomy of language functions in the National Certificates

The language functions on all test levels are approximately similar; what separates the test levels is the range and diversity of linguistic expression expected in performing these functions. On the Basic level it is important that the participant can fulfill the function, however short and simple the expression may be. On the Intermediate level, more natural and context-sensitive performance is expected, and on the Advanced level the expression of functions should be effective, natural and fully context-adaptive. The list of functions presented here is not intended to be exhaustive but we have attempted to present a representative range of functions under each subheading.

	functions	Basic level	Intermediate level	Advanced level
A.	**Giving and asking for factual information,** e.g. - stating, naming, identifying - answering in the positive and in the negative - reporting, describing, narrating - explaining, clarifying; asking for these - asking	Roughly, using basic vocabulary, phrases and structures	More extensive and nuanced realization, socially more appropriate use	Extensive and nuanced realization in more abstract situations including e.g. official circumstances Additionally, - summarizing and presenting main points of discussion

	functions	Basic level	Intermediate level	Advanced level
B.	**Expressing one's point of view**, e.g. - expressing opinion and asking about others' opinions - arguing for and against something - agreeing, accepting something - disagreeing, denying something - expressing knowledge and lack of knowledge of something - expressing certainty/uncertainty of something - expressing positive and negative evaluations and reacting to these	Roughly, using basic vocabulary, phrases and structures	More extensive and nuanced realization, socially more appropriate use. Additionally, - expressing certainty and doubt - assuring - persuading - giving reasons for opinions and inquiring these - criticizing, expressing disapproval - settling disputes, reconciliating	Extensive and nuanced realization in more abstract situations including e.g. official circumstances
C.	**Expressing emotions and attitudes**, e.g. - expressing happiness, satisfaction and hope - expressing sadness, dissatisfaction and hopelessness - expressing irresolution, disappointment, andger and fear - expressing boredom and frustration - expressing interest, like and dislike	Roughly, using basic vocabulary, phrases and structures	More extensive and nuanced realization, socially more appropriate use. Additionally e.g. - expressing appreciation and disappreciation	Extensive and nuanced, more detailed and accurate realization
D	**Acting through words**, e.g. - purchasing and dealing with public officials - requesting, placing orders - making and cancelling appointments - requesting advice and information, offering and requesting help - suggesting, requesting, recommending, and reacting to these - asking for and granting permission - reminding, warning	Roughly, using basic vocabulary, phrases and structures	More detailed and accurate realization, socially more appropriate use. Additionally e.g. - applying for permission - giving permission - accusing - making complaints, demanding compensation	Also in official circumstances (e.g. meetings, negotiations)
E.	**Acting according to social norms and customs**, e.g. - greeting and leavetaking - addressing and reacting to being addressed - presenting oneself and someone else - managing telephone conversations - apologizing, thanking, complimenting and congratulating, and reacting to these - expressing condolence - sending and giving regards - presenting and accepting invitations - declining politely	In familiar situations; with rough appropriacy, using basic expressions and phrases	More accurate and socially more appropriate realization, in familiar and semi-formal situations. Additionally e.g. - opening conversations (also with strangers) - small semi-formal speeches in contexts like welcoming, proposing a toast, thanking	Extensive and nuanced realization in more abstract situations including e.g. official circumstances such as meetings and negotioations, representing others

	functions	Basic level	Intermediate level	Advanced level
F.	**Communication strategies**, e.g. - clarification questions, repeating, spelling and requesting these - signalling comprehension and difficulties in comprehending - correcting and clarifying oneself - hesitations, fillers - backchanneling	Roughly, without nuances, only when absolutely necessary	As a more natural part of communication. Additionally e.g. - turn-taking, keeping and giving up turn - changing and avoiding topics	Naturally and nuancefully. Additionally e.g. - interrupting - organising own turn

Source: National Board of Education 1995

Appendix 4 Translations of the candidate questionnaires

N.B. The original questionnaires were administered in Finnish. Translation by the author. Slight re-wordings with respect to the names of the tests have been made in the interest of consistency in reporting. Questions which are directly relevant to Study One are marked with an asterisk.

POST-TEST QUESTIONNAIRE, TAPE-MEDIATED TEST

This questionnaire concerns the speaking tasks in the test you have just completed. There is a set of questions for each task that you completed. To help you remember the tasks, each task is described briefly before the questions relevant to that task. Firstly, your responses to five questions are requested on a scale from 1 to 5, strongly agree to strongly disagree. **Please circle the answer that best describes your opinion.** Then there is a question asking whether you were familiar with the task type from before. Below that, space is provided for giving reasons for your opinions, or making further comments.

First, a few preliminary questions:

*Have you ever been in a language laboratory before?	yes / no
*Have you ever participated in a test in the language laboratory before?	yes / no
*Did you experience any technical or auditory problems during the test?	yes / no

And now on to the task by task questions.

Answer scale:
1 - strongly agree
2 - tend to agree
3 - can't say
4 - tend to disagree
5 - strongly disagree

In Task 1 you were asked to read aloud a newspaper article.

	agree				disagree
I understood what I had to do.	1	2	3	4	5
The task was easy.	1	2	3	4	5
There was enough time.	1	2	3	4	5
The task seemed to measure my language ability.	1	2	3	4	5
I liked the task.	1	2	3	4	5

The task type was ___ new to me.
___ familiar to me. From where? _____

Further comments:

_____

_____

In Task 2 you participated in simulated conversations where you heard your conversation partner's turns from the tape. In the first question you were in a store changing a jacket that your friend had bought, in the second you changed the date of your return flight, and in the third you talked with an American woman about Finland.

	agree				disagree
I understood what I had to do.	1	2	3	4	5
The task was easy.	1	2	3	4	5
There was enough time.	1	2	3	4	5
The task seemed to measure my language ability.	1	2	3	4	5
I liked the task.	1	2	3	4	5

The task type was ___ new to me.
___ familiar to me. From where? _____

Further comments:

_____

_____

In Task 3 there were ten descriptions of situations, and you were asked to say what you might say in this situation. You met an old teacher of yours, travelled in Britain and the United States, and met foreigners in Finland.

	agree				disagree
I understood what I had to do.	1	2	3	4	5
The task was easy.	1	2	3	4	5
There was enough time.	1	2	3	4	5
The task seemed to measure my language ability.	1	2	3	4	5
I liked the task.	1	2	3	4	5

The task type was ___ new to me.
___ familiar to me. From where? _____

Further comments:

_____

_____

In Task 4 you talked about your opinions on pop music, and leaders and leading.

	agree				disagree
I understood what I had to do.	1	2	3	4	5
The task was easy.	1	2	3	4	5
There was enough time.	1	2	3	4	5
The task seemed to measure my language ability.	1	2	3	4	5
I liked the task.	1	2	3	4	5

The task type was ___ new to me.
___ familiar to me. From where? _____

Further comments:

_____

_____

Questions concerning the whole tape-mediated test of speaking

*How well did the situations in the test correspond with situations in which you have been in the real world, i.e. how comprehensive did you think that the test was?

_____

_____

*What would you have liked to take away from the test or add to it? ___ Nothing

_____

_____

*Did your performance on the tape-mediated test give a fair impression of your ability to speak English? Why?

_____

_____

*With what kind of a test should your speaking ability be tested for the assessment to be as fair as possible? How does the tape-mediated test you just completed relate to that test?

_____

_____

THANK YOU!

QUESTIONS ON THE FACE-TO-FACE TEST

This questionnaire concerns the face-to-face test you have just completed. We hope that you will answer all the questions. If some of the questions are inappropriate for you, please tell us why, and describe your situation in your own words.

*Have you participated in a face-to-face test before? ___ Yes ___ No

*How tense did the test make you feel?

___ very tense
___ somewhat tense
___ middling tense
___ not really tense
___ not at all tense

*What did you think about the length of the discussion test?

___ far too short
___ a bit too short
___ just right
___ a bit too long
___ far too long

How did you feel about having a peer discussion in the test?

_____

_____

*What would you have liked to take away from the test or add to it? ___ Nothing

_____

_____

*Did your performance on the discussion test give a fair impression of your ability to speak English? Why?

_____

_____

THANK YOU!

Appendix 5 The interview protocol

The interviews were conducted individually, while viewing the interviewees' performances, in Finnish. The tape was stopped after each task, at points which the interviewer found interesting, and at points where the itnerviewee wanted to make comments. The structure of the interviews was loose and recursive.

At beginnings and ends of tasks, and at points of interest in the performances, e.g. planning times in tape-mediated test, hesitations, corrections:

- What did you do?
- What were you thinking?
- What did the task make you do?

Of specific interest

tape-mediated

- role of imagination
- role of listening comprehension
- listening vs. reading cues in the booklet

face-to-face

- relationships between peers in part 1
- attitudes towards peers in part 1
- turn-taking
- differences between peer discussion and interlocutor-interviewee discussion

both

- correspondence of performance with test-external language use
- fairness of the impression of your spoken English
- which factors other than speaking did you think had an effect on your performance

Appendix 6 Face-to-face test assessment sheet and questionnaire

Candidate A	**Candidate B**	**Candidate C**
Name: _____	Name: _____	Name: _____
Skill level ___	Skill level ___	Skill level ___
Discussion___	Discussion___	Discussion___
Interview ___	Interview ___	Interview ___
Why this skill level?	Why this skill level?	Why this skill level?

Describe the performances

How were the performances similar to each other?

How did the performances differ from each other?

How well did the elicitation succeed?

How fair was the test for the participants?

How useful was the sample for making a skill level assessment?

What would you have liked to add?

How confident are you of the correctness of the assessments you gave?

Appendix 7 Proposal for a new post-test questionnaire

QUESTIONNAIRE ON THE SPEAKING PART OF THE TAPE-MEDIATED TEST

background information:

Have you ever been in a language laboratory before?	yes / no
Have you ever taken a **test of speaking** in a language lab before?	yes / no
Were there any technical or auditory problems with the test?	yes / no
Have you already taken the face-to-face test of speaking today?	yes / no

This questionnaire concerns the **speaking part** of the tape-mediated test you have just taken. Most questions ask for your opinions on a scale 1-4, fully agree to fully disagree, or 0 = no opinion. **Please circle the number that best corresponds to your opinion on the test.** If you want to explain your reply, please write your comments below each question, or on the back of the page. There is some space for further comments at the end of the questionnaire.

1. I understood what I had to do (the instructions were clear).

1	2	3	4	0
agree	tend to agree	tend to disagree	disagree	no opinion

2. Test anxiety affected my performance to an appreciable degree.

1	2	3	4	0
agree	tend to agree	tend to disagree	disagree	no opinion

3. The test was difficult.

1	2	3	4	0
agree	tend to agree	tend to disagree	disagree	no opinion

4. I did well on the test.

1	2	3	4	0
agree	tend to agree	tend to disagree	disagree	no opinion

5. If I had taken the test on another day, I would have done better.

1	2	3	4	0
agree	tend to agree	tend to disagree	disagree	no opinion

Why?

_____

_____

6. I did not have enough time to answer the questions.

1	2	3	4	0
agree	tend to agree	tend to disagree	disagree	no opinion

7. Although I had to interact with a machine, the test sampled my strengths and weaknesses quite well.

1	2	3	4	0
agree	tend to agree	tend to disagree	disagree	no opinion

8. I liked the speaking tasks in the test.

1	2	3	4	0
agree	tend to agree	tend to disagree	disagree	no opinion

9. The test was too long.

1	2	3	4	0
agree	tend to agree	tend to disagree	disagree	no opinion

10. The topics and speaking situations in the test could happen in "the real world".

1	2	3	4	0
agree	tend to agree	tend to disagree	disagree	no opinion

11. The test was interesting.

1	2	3	4	0
agree	tend to agree	tend to disagree	disagree	no opinion

12. Taking a test of speaking in the language laboratory was a disagreeable experience.

1	2	3	4	0
agree	tend to agree	tend to disagree	disagree	no opinion

13. Had the test had different tasks or topics, I would have done better.

1	2	3	4	0
agree	tend to agree	tend to disagree	disagree	no opinion

What kinds of tasks / topics?

_____

_____

14. If you have now taken both tests of speaking (the tape-mediated and the face-to-face one), which test do you think gives a more fair impression of your ability to speak English?

1 The tape-mediated test 2 The face-to-face test 3 No difference

Why?

_____

_____

_____

Further comments (you can continue on the back if you wish):

_____

_____

_____

_____

_____

THANK YOU!

The two proposed post-test questionnaires are almost identical. However, two questions would be formulated slightly differently in the face-to-face test questionnaire, i.e.:

7. Although I only spoke with one person and on two or three topics, the test sampled my strengths and weaknesses quite well.

12. Taking the face-to-face test of speaking was a disagreeable experience.

Appendix 8 Reasons for including the questions, and working hypotheses

The questions of the proposed questionnaire are dealt with one by one. For each, the reasons for including it are listed first, followed by the working hypotheses attached to the question.

The non-numbered background questions:
Have you ever been in a language laboratory before?
Have you ever taken a speaking test in a language laboratory before?
Have you ever taken a face-to-face test of speaking before?

- Reason for inclusion: Basic background information that may explain some reactions.
- Working hypotheses:

1) Inexperience with the environment (language lab) or the test format will not be a significant factor across the whole population but there will be individuals who achieve a lower proficiency assessment because of it.
2) Inexperience explains negative affective reactions to the tests.

Were there any technical or auditory problems with the test?

- Reason for inclusion: if there were any problems, the reactions should be interpreted in the light of them. Positive cases must be investigated separately in case the technical problems colour other reactions to the test.
- Working hypothesis: very few problems.

Have you taken the face-to-face / tape-mediated test of speaking already?

- Reason for inclusion: it is impossible to know which test has been taken first
- Working hypothesis: order does not affect reactions or results.

1. I understood what I had to do (the instructions were clear).

- Reason for inclusion: may affect results or reactions.
- Working hypothesis: The majority will find the instructions clear. Instructions in the language laboratory may be more difficult to follow. Older people may find it harder to follow the instructions.

2. Test anxiety affected my performance to an appreciable degree.

- Reason for inclusion: the most frequently-researched variable. Previous results: small degrees of anxiety will not affect results but large degrees will. The results are clearest with low-achieving participants.
- Working hypotheses:

1) Tendency towards comparable effects as in previous studies.
2) The tape-mediated test causes more anxiety for many participants.

3. The test was difficult.
- Reason for inclusion: Highly salient perception for candidates. According to Fulcher (1993) this is a separate factor in candidate reactions. Enables investigation of relationship between difficulty and perceived difficulty.
- Working hypotheses: there is a tendency for low-achieving candidates to perceive the tape-mediated test as more difficult, while there is no difference in reactions to the face-to-face test.

4. I did well on the test.
- Reason for inclusion: Potentially independent type of reaction (Fulcher 1996)
- Working hypotheses: Correlates with test results and positive reactions to test. Many will think they did better on the face-to-face test than the tape-mediated one.

5. If I had taken the test on another day, I would have done better. Why?
- Reason for inclusion: A way of investigating some of the dimensions of not doing well. Verbal explanations may be interesting.
- Working hypotheses: Correlates with test anxiety, low negative correlation with question 4.

6. I did not have enough time to answer the questions.
- Reason for inclusion: Participant reactions to test quality. The results of the present study indicate that this is a highly relevant factor for participants.
- Working hypotheses: Correlates with test anxiety and negative reactions to test.

7. Although I had to interact with a machine, the test sampled my strengths and weaknesses quite well.
7. Although I only spoke with one person and on two or three topics, the test sampled my strengths and weaknesses quite well.
- Reason for inclusion: Perceived validity.
- Working hypotheses: Reactions will be positive overall. There will be no significant difference between the two tests. Correlates with other positive reactions, not with test results.

8. I liked the speaking tasks in the test.
- Reason for inclusion: affective reaction to the test.
- Working hypotheses: The face-to-face test will get more positive reactions. Correlates with other positive reactions. Correlates weakly with test results.

9. The test was too long.
- Reason for inclusion: Independent perception of test. Significant factor in the pilot test.
- Working hypotheses: Face-to-face test will tend to be perceived as too short (particularly by high achievers), tape-mediated test will be the right length.

10. The topics and speaking situations in the test could happen in "the real world".
- Reason for inclusion: Perceived authenticity.
- Working hypotheses: Reactions will be positive. Correlates with question 8 and other positive reactions.

11. The test was interesting.
- Reason for inclusion: perceived relevance.
- Working hypotheses: Reactions will be positive overall. Fulcher (1993) found an enjoyment factor = like/dislike, interest.

12. Taking a test of speaking in the language laboratory was a disagreeable experience.
- Reason for inclusion: Affective reaction, part of enjoyment factor
- Working hypotheses: Correlates with questions 11 and 8. Low negative correlation with test results.

13. Had the test had different tasks or topics, I would have done better. What kinds of tasks / topics?

- Reason for inclusion: Affective reaction, perceived irrelevance or inability to show proficiency. May yield interesting verbal responses. Connected with question 5.
- Working hypotheses: Correlates with test anxiety. Low negative correlation with question 4.

14. If you have now taken both tests of speaking (the tape-mediated and the face-to-face one), which test do you think gives a more fair impression of your ability to speak English? Why?
1 The tape-mediated test 2 The face-to-face test 3 No difference

- Reason for inclusion: Candidate preference.
- Working hypotheses: Face-to-face test will be favoured. Those respondents who favour tape-mediated test will feel that they did not do well on the face-to-face test. They may want to explain this.

Appendix 9 Sample of the rich transcript for the tape-mediated test

This extract is from the beginning of the rich transcript of the tape-mediated test.

Sp_in = spoken instructions
Wr_in = written instructions
Input_w = written input
Input_s = spoken input

Sp_in: In the speaking test you will give your answers by speaking. Your answers will be recorded on tape.

Sp_in: Task 1: Reading out loud

Sp_in: Please read the instructions for task 1 in your test booklet. You will have half a minute to do this.

Wr_in: PUHUMINEN

Wr_in: Tehtävä 1: Ääneenluku

Input_w: Olet matkalla englannissa. Olet lukemassa lehteä ja löydät mielenkiintoisen jutun, jonka päätät lukea ystävällesi. Teksti on alla.

Wr_in: Tehtäväsi on lukea tämä uutinen nauhalle. Sinulla on minuutti aikaa tutustua tekstiin, ja sen jälkeen minuutti aikaa tekstin lukemiseen nauhalle.

Wr_in: Tutustu tekstiin valmisteluajan kuluessa. Valmisteluajan loputtua kuulet nauhalta kehotuksen alkaa lukea tekstiä ääneen. Älä aloita ääneen lukemista ennen kuin kuulet kehotuksen!

Wr_in: Kymmenen sekuntia ennen vastausajan loppua kuulet nauhalta merkkiäänen. Äänen jälkeen Sinulla on siis kymmenen sekuntia aikaa lukea teksti loppuun.

Sp_in: In this task, you will have one minute to look at the text and one minute to read it out loud on tape. Ten seconds before the end of the reading pause you will hear a sound like this [ping]. After the sound you will have ten seconds to finish reading the text.

Sp_in: You will now have one minute to look at the text. Please do NOT start to speak on tape until you are asked to do so.

Input_w: Wally the 30-minute millionaire

Input_w: Mr. Wally Taylor, who roams the tropical northern beaches in Queensland, Australia in a mobile home, won a million Australian dollars on a lottery ticket and gave it all away in less than 30 minutes, at a rate of around $600 a second. He gave it to 15 relatives, a couple of friends, the Heart Foundation and cancer research. For himself he kept only a few hundred bucks, "for emergencies," in case the sunshine ends. "I am very happy to have traded all that money for a quiet life in the tropics. I'm not a nut. I reckon I spent the money wisely," he says.

Sp_in: Please start reading the text on tape now.

Sp_in: Thank you. This is the end of task one. Please turn to task two in your test booklet and read the instructions.

Wr_in: Tähän päättyy tehtävä 1. Tutustu tehtävän 2 ohjeisiin. Sinulla on minuutti aikaa lukea ohjeet ja tilannekuvaukset seuraavalta sivulta.

Appendix 10 Sample of the rich transcript for the face-to-face test

This extract is from the beginning of the rich transcript for one administration of the face-to-face test. NB. The written instructions are read by the candidates before the test itself begins.

Sp_in = spoken instructions
Wr_in = written instructions
Input_w = written input
Input_s = spoken input

Wr_in: DISCUSSION / INTERVIEW
Wr_in: This test consists of two parts: in Part A you discuss with another test-taker, in Part B you talk alone with the interviewer. The only language used in the room is English.
Wr_in: In the beginning, you introduce yourself to the interviewer and to the other test-taker, and tell some basic things about yourself.
Wr_in: Part A: Discussion
Wr_in: You will get a choice of three topics. First, you and your partner should decide which topic you want to discuss. **(Remember, only English!)** After you have decided what you want to talk about, you will read a short introduction into the topic. The interviewer will help you get started, and she may take part in the discussion, but you will mainly be talking with your partner. In the discussion, try to give your opinion of the topic, and give reasons for your opinions. Listen to your partner's opinion, and try to find out the reasons for his/her opinions. The interviewer will join in at the end to close the discussion.
Wr_in: Part B: Interview
Wr_in: In this part you will have a chance to talk about another topic, this time alone with the interviewer. Try to speak as much as you can: give your own point of view, give reasons, ask questions and/or give examples.

Input_s: please sit down here ## I don't even know your names so please could you give them to me # what's your name?

Input_s: Firstname # was it Lastname?

Input_s: Lastname # Lastname ok.

Input_s: and yours?

Input_s: are you something # are you related to each other?

Input_s: are you # that's nice.

Sp_in: all right <#> so you read <#> the piece of paper you were given.

Sp_in: aand <#> so we'll start with the discussion and it goes like this <#> I'm going to choose one card <##> let's take this one <#> and you'll have to decide between yourselves which one of the three topics you would like to <##> discuss <#> there you go <###>.

Input_w: DISCUSSION TOPICS 1. Stereotypes and prejudices 2. What can we do to save the environment? 3. Taxation (=verotus) in Finland.

Input_s: I'm sorry I &o <#> understand only only <laughs> english but let me explain <#> &erm <#> environment is something that you see outside it's the nature it's <#> where houses and trees are <#> what you have surrounding.

Sp_in: so it's number two <##> ok <#> this is some more explanation on it <#> you can read it and then we can begin <###>.

Input_w: What can we do to save the environment?

Input_w: Environment and pollution have been much discussed recently. Do you think pollution is a problem to us? If so, is there anything you could do yourself to help? What really are the chances of one individual making a difference in these matters?

Input_s: is it # a very important question for you ## saving # of the environment?

Appendix 11 Results of Task Analysis in tabular form

	Task Characteristics (Bachman & Palmer 1996)	**Tape-mediated test**	**Face-to-face test**
1	**Testing environment** nature and familiarity: - physical setting - participants - time of task	- language lab; 20-25 booths - headphones and microphone - participants: candidates in group but seated individually, invigilator - temperature & lighting good - familiarity of setting and participants varied - time: evening	- classroom - video camera and tape recorder - participants: candidate, 1-2 peers, interlocutor - temperature & lighting good - familiarity of setting and participants varied - time: evening
2	**Test rubric**		

	Task Characteristics (Bachman & Palmer 1996)	Tape-mediated test	Face-to-face test
	a. Instructions - language - channel - specification of procedures and tasks	- native and target language - written and spoken - task instructions fairly specificand different for each task; examples given - given task by task	- target language - written and spoken - instructions fairly specific and different for the two tasks - given in one block in writing, task by task in speaking
	b. Structure - number of parts/tasks - salience of parts/tasks - sequence of parts/tasks - relative importance of parts/tasks	- warm-up and three tasks - parts salient; clear transitions between tasks - order of increasing difficulty - no information given on relative importance of parts	- warm-up and two tasks - parts salient; clear transitions between tasks - order of difficulty vaguely increasing - no information given on relative importance of parts
	c. Time allotment	- each response timed individually - intended power test - both speed and power involved	- tasks timed, individual responses adaptive - power test
	d. Scoring method - criteria for correctness - procedures for scoring the response - explicitness of criteria and procedures	- main criteria comprehensibility and approximate appropriacy - rating scale and point scale - clear sequence, two types of procedure; same raters - criteria and procedures not explicit to candidates	- main criteria comprehensibility and approximate appropriacy - rating scale - holistic procedure, immediate + revision from video - criteria and procedures not explicit to candidates
3	**Input**		
	a. Format - channel, form - language - length - type (item, prompt) - degree of speededness - vehicle	- spoken and written language - target language (NS) - length one utterance/turn - structured prompts - variable degree of speededness - taped	- spoken and written, language and non-language - target language (NNS) - length one utterance/turn - prompts and backchanneling signals - low degree of speededness - live

	Task Characteristics (Bachman & Palmer 1996)	Tape-mediated test	Face-to-face test
	b. Language of input - grammatical and textual characteristics (vocab., morph., syntax, phon.; cohesion, rhetorical organisation) - functional and sociolinguistic characeristics - topical characteristics	- vocabulary intermediate (3 on a 4-level scale) - written syntax 2 on a 4-level scale, spoken syntax 1 - input in mother tongue may have influenced processing - cohesion within tasks, fragmented between tasks - regional Englishes, no personal interaction - 6-8 topics, little depth	- vocabulary intermediate (3 on a 4-level scale) - written syntax 2 on a 4-level scale, spoken syntax 1 - input in one language only - rather cohesive, only one abrupt task juncture - little sociolinguistic variation - two topics, some depth
4	**Expected response**		
	a. Format - type (limited/ extended) - channel, form - language - length - degree of speededness	- extended response, *guided* (short and extended turns, two short presentations) - spoken TL, recorded on cassette (i.e. linguistic/ vocal) - approx. 15 minutes - may be speeded to a degree	- extended response, *constructed* (mostly lengthy turns, but supported by backchanneling) - spoken TL, recorded on video (i.e. language and nonlanguage) - approx. 15 minutes - not speeded, but total length restricted
	b. Language - grammatical and textual characteristics (vocab., morph., syntax, phon.; cohesion, rhetorical organisation) - functional and sociolinguistic characeristics - topical characteristics	- vocabulary everyday, the more advanced the better - expected to master basic grammar, advanced structures may be shaky - phonologically comprehensible, may require effort from listener - variety of functions - approximate appropriacy to semi-formal and informal contexts - master common politeness formulas - variety of everyday topics and situations	- vocabulary everyday, the more advanced the better - expected to master basic grammar, advanced structures may be shaky - phonologically comprehensible, may require effort from listener - any functions that suit context - approximate appropriacy to context - basic skills for conversation management - two topics, abstract but adaptable
5	**Relationship between input and response** - reactivity - scope of relationship - directness of relationship	- nonreciprocal (uni-directional; master tape to candidate only) - amount of input to comprehend variable; towards narrow scope - response not supplied by the task; has to be provided by the candidate	- reciprocal (backchanneling; little feedback on correctness) - amount of input to comprehend variable; towards broad scope - response not supplied by thetask; has to be provided by the candidate

Appendix 12 Concordances of hesitations in two candidates' performances

NB The concordances are printed in a different font in order to maintain the alignment, which makes reading the concordances easier.

Voiced hesitations in Anja's performance

One to five words of context in the same speaker turn before and after the hesitation marker have been given below. The extent of the context depends on the position of the hesitation marker in the speaker turn as well as availability of space. # and ## stand for pauses.

Tape-mediated

yes	er	# my friend Sarah # bought yesterday
Sarah # bought yesterday this	er	# jacket and now she wants to ##
now she wants to ## change it	er	# in a # bigger one ehh
# in a # bigger one	er	# one number bigger than this is
o take it # Sarah gave it and	er	I left it and # and I left it on the
opened this # mm mm store and	er	# she 's er # about twenty ye
mm mm store and er # she 's	er	# about twenty years old # fair hair
he does # she comes here &alm	er	every week and # that 's very kind of
ppreciate this # emm do you #	er	how much do I owe ?
ery interesting to know # but	er	# sorry to inform that finland is an
y six years # and it 's not #	er	# s f does n't mean soviet finland .
good for the moment # because	er	# we have economical problems here in
here in finland but it is ##	er	# average european standard .
s if you wish to come here in	er	in the summer # I think it 's better
definitely see lapland and ##	er	# in # it 's very beautiful and
interesting town # but it 's	er	very # very different from the #
oin me # I would like to # mm	er	## would you like to join me #
	er	I 'm sorry # I 'm sorry to disturb
it possible that I could rest	er	for a while # is there any corner
think it was # it is stolen #	er	# I had it here when I game to this m
er # but now it 's gone # and	er	# I 'd like to # make an
t to be here and # and it was	er	# very # er mm
and it was er # very #	er	mm # I 've been I 've learnt a lot
## and I wish you also have a	er	er ponity # opportunity to come
and I wish you also have a er	er	ponity # opportunity to come # and vi
ourney was quite pleasant but	er	# look at my trousers ## this
restaurants # a group of	er	## for the same # working place
e going together to have some	er	## wine or glogi* and eat christmas m
ngs to christmas like ham and	er	# finnish specialities # it 's
sic every day # I think it is	er	# impossible mm er live without ehh
impossible mm er live without	er	# pop music er # because we a
live without er # pop music	er	# because we are so used to hear it a
# we can hear it always # and	er	# I think it 's # not very &am
nt that everything that we do	er	# it can be heared so # if you
ve also # classical music and	er	# my favourites in er pop musics are
sic and er # my favourites in	er	pop musics are # pop music is emm
many times # their music and	er	# evergreens like Beatles of course a
nd of festival # that we have	er	# we have in ## restaurants
hink it is er # impossible mm	er	live without er # pop music ehh
ould like to # change so my #	emm	flight to # back finland .
I # really appreciate this #	emm	do you # er how much do I owe ?
it was really a pleasure to #	emm	## thank you .
studying in # Jyvaeskylae ##	emm	# workers clup .
I 'd like to # make an #	emm	report of it .
at a pity # did you see the #	emm	the # plate in the car # the
there was a car that broke #	emm	# and the # it drove through &
and # now I 'm # so dirty but	emm	# maybe I can # you have if yo
pop musics are # pop music is	emm	# do you know Sting # and Dire
atles of course and # I think	emm	# life without music is not so fun

Face-to-face

ing systems we have huge em #	er	I do n't remember the word.
but # I 'm washing the # milk	er	?
tory we have # had rules that	er	make barriers to other people who wan
the only reason that there 's	er	&eno economical problems now ?
	er	# the finnish people have # been trav

ng only # so little time that	er	# the usual man in the street
mies" # he does n't know that	er	# the world is so wide and
# and	er	yes . how could we have #
and the swedes also	er	# wanted that the finnish come and wo
irty work and # and also from	er	turkey
	er	# we should n't have nuclear #
	er	# I do n't &kn know if it is much but
together # to make it # but	er	# who is going to give us the
it is possible but	er	# we need a leader .
english # I never need it I #	emm	that 's why I # I want to know if I
copying systems we have huge	emm	# er I do n't remember the word .
# first it # we should use it	emm	# more effectively .
o are planning # for instance	emm	traffic .
	emm	no # the marches are not enough they

Voiced hesitations in Eeva's performance

Tape-mediated

us english five years ago ##	er	I 'm missis Lastname if you do n't
nd she would like to change #	er	changing of this jacket # and she woul
no I 'm sorry	er	# er # she ##
parts of finland and # and er	er	# a little more in the southern part
o visit your company and er #	er	# it is very nice that we are able to
then it is not a relaxing mm	er	move me a relaxing moment ## and
gs should be done ## and # er	er	# maybe they then always want to say
always your opinion too # and	er	# they # should know what to
no I 'm sorry er #	er	# she ## did n't have it with
uld like to change my # emm #	er	trip # to mm a little later #
le to change it to tomorrow #	er	# I am # I am missis Lastname
he middle of the finland # mm	er	about a three hundred kilometres north
st parts of finland and # and	er	er # a little more in the southern
ong # I 'm going for dinner #	er	are you going to do the same ## is it
&co # keep me company ##	er	do you know what is # what there is to
per for my # travel insurance	er	er # the park was very crowded so that
for my # travel insurance er	er	# the park was very crowded so that I
us to visit your company and	er	# er # it is very nice that
ts ## and usually it is emm #	er	it comes out too loud and then it
oud and then it hurts my ears	er	# er I like sometimes background music
nd then it hurts my ears er #	er	I like sometimes background music but
I like it that it is ok but #	er	sometimes I can see that they have it
nce but sometimes I feel that	er	the man would know a little more how #
hings should be done ## and #	er	er # maybe they then always want to
are not asking from you but #	er	a good leader wants # to know always
# I would like to change my #	emm	# er trip # to mm a little lat
years independent so # oh #	emm	# finland has not been for a long time
igh # but now we have had a #	emm	# so many people without work
le without work ## so that oh	emm	actually there are five hundred
ause it is very natureral and	emm	# it is the part of finland you should
mp;ba to come back here # and	emm	# I am very happy and # all
ething to eat and drink and #	emm	# most # most often to dance
truments ## and usually it is	emm	# er it comes out too loud and then it
I go to the concerts too but	emm	# sometimes they play too loud too and
# emm # er trip # to	mm	a little later # one day later #
	mm	# just a minute # I do n't have it
	mm	# &wha # what about this
m the middle of the finland #	mm	er about a three hundred kilometres
Helsinki there are the most #	mm	oldest parts of finland and # and er
and then it is not a relaxing	mm	er move me a relaxing moment ## and
she who knows the best # and	mm	# sometimes I think that it would be

Face-to-face

so you think that #	er	women 's and men 's role in war is the
in in [laughs] usual life but	er	# er ## what was the thought I had I
[laughs] usual life but er #	er	## what was the thought I had I forgot
e because mm # though we have	er	quite good # equality in finland and
d # equality in finland and #	er	# they # now when there is a s
but	er	# how do you do that ?
there the big enough bomb and	er	# mm [laughs] +...

```
g in my # work that I could #  er   # bake # karelian pastries and
bake and I would do some # mm  er   something for them if they need me and
ld go there and help them and  er   +...
do with my hands knitting and  er   # crocheting and emm +...
                    because #  er   or maybe younger men learn already #
e are talking about it that #  emm  women should # stay home and take care
> their # their parents and #  emm  old people # and expect not to get pai
eed me and they pay me some #  emm  # some but like # if they have
tting and er # crocheting and  emm  +...
he roles are the same because  mm   # though we have er quite good #
 the big enough bomb and er #  mm   [laughs] +...
 finnish pulla " # and then I  mm   # would come to my old work and sell
 I bake and I would do some #  mm   er something for them if they need me
```

Appendix 13 Completed grid for the face-to-face test

The elements — the performance extractswhich the assessors rated— are in the columns. The potential assessment criteria that the assessors named were written one by one on the rows. The circles in the completed grid signal which three performances elicited each potential assessment criterion. The last three rows of circles represent the non-fruitful attempts to elicit more assessment constructs

E 1	E 2	E 3	E 4	E 5	E 6	E 7	E 8	E 9	E 10	E 11	E 12	E 13	E 14	E 15	
7	5	③	④	3	5	5	7	4	4	5	6	⑤	2	2	transfer / interference
7	5	3	5	③	⑤	3	7	5	4	5	6	5	2	②	rate of speech
⑦	6	4	4	2	6	4	⑦	6	④	6	6	5	3	4	confidence
7	④	5	3	3	5	6	7	④	4	5	⑥	5	2	3	grammar
7	5	3	3	②	5	6	7	4	4	④	6	5	②	3	vocabulary
5	④	4	4	5	6	5	⑦	4	4	⑤	5	4	3	3	pronunciation / indiv. sounds
⑥	4	④	4	3	4	4	7	4	4	4	5	4	②	2	pronunciation / prosody
7	7	3	③	1	6	4	7	5	3	5	6	⑦	1	③	length of turn
7	4	③	5	1	3	⑤	7	4	③	5	6	4	1	2	collocations and idioms
⑦	5	3	5	3	5	5	7	5	④	5	⑥	5	2	3	comprehensibility
							○	○	○						
			○										○	○	
○											○	○			

Appendix 14 Correlations between the criteria mentioned by the assessors

NB This appendix is printed in a different font and in landscape format to improve readability.

SPSS-Win output for Spearman correlation coefficients between criteria mentioned by the assessors on the **tape-mediated test**

CREATIV	,8604**										
SPEED	,9016**	,7892**									
ACCENT	,5347	,7554**	,5129								
VOCAB	,8388**	,8128**	,6974*	,7396**							
IDIOM	,7774**	,8735**	,6092*	,7814**	,9138**						
PHRASES	,9273**	,9640**	,8355**	,6852*	,8648**	,8891**					
PROSODY	,7155**	,7337**	,6791*	,6682*	,8751**	,7738**	,7488**				
CONFDNCE	,8243**	,8008**	,7794**	,6967*	,7206**	,7367**	,8215**	,5949*			
GRAMMAR	,7819**	,7771**	,7510**	,7176**	,8671**	,7245**	,7999**	,7784**	,5505		
	OVERALL	CREATIV	SPEED	ACCENT	VOCAB	IDIOM	PHRASES	PROSODY	CONFDNCE		

* - Signif. LE ,05 ** - Signif. LE ,01 (2-tailed)

SPSS-Win output for Spearman correlation coefficients between criteria mentioned by the assessors on the **face-to-face test**

TRANSFER	,8262**										
SPEED	,9400**	,7628**									
CONFDNCE	,8589**	,7653**	,8939**								
GRAMMAR	,8930**	,8876**	,7730**	,7591**							
VOCAB	,9102**	,9193**	,8326**	,8537**	,9401**						
PR_SOUND	,4967	,5014	,3754	,4018	,6570**	,4696					
PROSODY	,8648**	,8545**	,9055**	,8638**	,7926**	,8490**	,5227*				
TLENGTH	,9026**	,7417**	,9478**	,8887**	,7858**	,8768**	,3526	,8486**			
IDIOMS	,7412**	,9003**	,7800**	,8145**	,7821**	,8208**	,3743	,8545**	,7274**		
COMPREH	,8914**	,8793**	,8913**	,9078**	,8344**	,8698**	,4731	,8691**	,8330**	,9136**	
	OVERALL	TRANSFER	SPEED	CONFDNCE	GRAMMAR	VOCAB	PR_SOUND	PROSODY	TLENGTH	IDIOMS	

* - Signif. LE ,05 ** - Signif. LE ,01 (2-tailed)

or analysis of grid ratings for the tape-mediated test

rinted in a different font to improve readability.

analysis 1, Principal Components Analysis (PC)

ıality	*	Factor	Eigenvalue	Pct of Var	Cum Pct
	*				
00000	*	1	7.34145	81.6	81.6
00000	*	2	.56385	6.3	87.8
00000	*	3	.36726	4.1	91.9
00000	*	4	.28663	3.2	95.1
00000	*	5	.23367	2.6	97.7
00000	*	6	.13375	1.5	99.2
00000	*	7	.05142	.6	99.8
00000	*	8	.01448	.2	99.9
00000	*	9	.00750	.1	100.0

. factors.

ıality	*	Factor	Eigenvalue	Pct of Var	Cum Pct
	*				
92036	*	1	7.34145	81.6	81.6
78485	*				
70338	*				
90840	*				
86220	*				
91780	*				
69451	*				
71359	*				
83637	*				

: factors.

ıality	*	Factor	Eigenvalue	Pct of Var	Cum Pct
	*				
92583	*	1	7.34145	81.6	81.6
89454	*	2	.56385	6.3	87.8
85551	*				
90948	*				
87086	*				
95363	*				
80873	*				
83496	*				
85175	*				

ł in 3 iterations.

:ix:

›r 1	Factor 2
;713	.37766
ł947	.33668
;033	.52980
ł113	.61364
;186	.86335
;423	.82223
7129	.72483
)286	.71234
;237	.68611

Factor Transformation Matrix:

	Factor 1	Factor 2
Factor 1	.71895	.69506
Factor 2	-.69506	.71895

OBLIMIN converged in 33 iterations.

Pattern Matrix:

	Factor 1	Factor 2
CONFDNCE	.97226	
SPEED	.97116	
PHRASES	.81059	
CREATIV	.64769	.37412
PRON_ATT		.98321
PROSODY		.90542
GRAMMAR	.34139	.63736
IDIOM	.39320	.59976
VOCAB	.49255	.52516

Factor Correlation Matrix:

	Factor 1	Factor 2
Factor 1	1.00000	
Factor 2	.75596	1.00000

PC extracted 3 factors.

OBLIMIN converged in 22 iterations.

Pattern Matrix:

	Factor 1	Factor 2	Factor 3
SPEED	.96392		
GRAMMAR	.82292	.31599	
VOCAB	.62718	.36470	
PHRASES	.62672		.32007
CREATIV	.48865	.34820	
PRON_ATT		.95102	
PROSODY		.73009	
IDIOM		.64906	.33156
CONFDNCE			.74629

Factor Correlation Matrix:

	Factor 1	Factor 2	Factor 3
Factor 1	1.00000		
Factor 2	.71267	1.00000	
Factor 3	.55422	.41393	1.00000

Appendix 16 Factor analysis of grid ratings for the face-to-face test

NB This appendix is printed in a different font to improve readability.

Extraction 1 for analysis 1, Principal Components Analysis (PC)

Variable	Communality	*	Factor	Eigenvalue	Pct of Var	Cum Pct
		*				
TRANSFER	1.00000	*	1	8.14556	81.5	81.5
SPEED	1.00000	*	2	.81341	8.1	89.6
CONFDNCE	1.00000	*	3	.35434	3.5	93.1
GRAMMAR	1.00000	*	4	.25694	2.6	95.7
VOCAB	1.00000	*	5	.14791	1.5	97.2
PR_SOUND	1.00000	*	6	.12246	1.2	98.4
PROSODY	1.00000	*	7	.08728	.9	99.3
TLENGTH	1.00000	*	8	.04372	.4	99.7
IDIOMS	1.00000	*	9	.02606	.3	100.0
COMPREH	1.00000	*	10	.00233	.0	100.0

PC extracted 1 factors.

Final Statistics:

Variable	Communality	*	Factor	Eigenvalue	Pct of Var	Cum Pct
		*				
TRANSFER	.86508	*	1	8.14556	81.5	81.5
SPEED	.86173	*				
CONFDNCE	.83720	*				
GRAMMAR	.85659	*				
VOCAB	.90010	*				
PR_SOUND	.36399	*				
PROSODY	.88704	*				
TLENGTH	.81207	*				
IDIOMS	.86637	*				
COMPREH	.89538	*				

PC extracted 2 factors.

Final Statistics:

Variable	Communality	*	Factor	Eigenvalue	Pct of Var	Cum Pct
		*				
TRANSFER	.86604	*	1	8.14556	81.5	81.5
SPEED	.90052	*	2	.81341	8.1	89.6
CONFDNCE	.89424	*				
GRAMMAR	.89239	*				
VOCAB	.90203	*				
PR_SOUND	.95903	*				
PROSODY	.90659	*				
TLENGTH	.87129	*				
IDIOMS	.87113	*				
COMPREH	.89570	*				

VARIMAX converged in 3 iterations.

Rotated Factor Matrix:

	Factor 1	Factor 2
CONFDNCE	.92636	
SPEED	.91979	
TLENGTH	.91594	
VOCAB	.87052	.37977
IDIOMS	.86550	.34934
COMPREH	.85671	.40217
TRANSFER	.84810	.38309
PROSODY	.78315	.54154
GRAMMAR	.74669	.57866
PR_SOUND		.95858

Factor Transformation Matrix:

	Factor 1	Factor 2
Factor 1	.89710	.44182
Factor 2	-.44182	.89710

OBLIMIN converged in 5 iterations.

Pattern Matrix:

	Factor 1	Factor 2
CONFDNCE	1.00591	
TLENGTH	.99590	
SPEED	.99155	
VOCAB	.91280	
IDIOMS	.91210	
COMPREH	.89373	
TRANSFER	.88718	
PROSODY	.78893	
GRAMMAR	.74214	.33312
PR_SOUND		.94492

Factor Correlation Matrix:

	Factor 1	Factor 2
Factor 1	1.00000	
Factor 2	.46651	1.00000

PC extracted 3 factors.

OBLIMIN converged in 17 iterations.

Pattern Matrix:

	Factor 1	Factor 2	Factor 3
IDIOMS	1.03106		
TRANSFER	.94876		
GRAMMAR	.86815		
VOCAB	.83420		
COMPREH	.66309		
PROSODY	.40714	.30932	.39021
PR_SOUND		.98325	
TLENGTH			1.02455
SPEED			.92791
CONFDNCE	.33023		.68764

Factor Correlation Matrix:

	Factor 1	Factor 2	Factor 3
Factor 1	1.00000		
Factor 2	.52536	1.00000	
Factor 3	.84927	.39752	1.00000