Maija Saleva

Now They're Talking

Testing Oral Proficiency in a Language Laboratory

STUDIA PHILOLOGICA JYVÄSKYLÄENSIA

STUDIA PHILOLOGICA JYVÄSKYLÄENSIA 43

Maija Saleva

Now They're Talking

Testing Oral Proficiency in a Language Laboratory

Esitetään Jyväskylän yliopiston humanistisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston vanhassa juhlasalissa (S212)
marraskuun 8. päivänä 1997 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Humanities of the University of Jyväskylä,
in Auditorium S212, on November 8, 1997 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 1997

Now They're Talking

Testing Oral Proficiency in a Language Laboratory

STUDIA PHILOLOGICA JYVÄSKYLÄENSIA 43

Maija Saleva

Now They're Talking

Testing Oral Proficiency in a Language Laboratory

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 1997

Editors
Raija Markkanen
Department of English, University of Jyväskylä
Kaarina Nieminen
Publishing Unit, University Library of Jyväskylä

URN:ISBN:978-951-39-8307-9
ISBN 978-951-39-8307-9 (PDF)
ISSN 0585-5462

ISBN 951-39-0070-3
ISSN 0585-5462

Cover
Pirjo Viitanen

Jyväskylä University Printing House,
Jyväskylä and ER-Paino Ky, Lievestuore 1997

ABSTRACT

Saleva, Maija
Now They're Talking. Testing Oral Proficiency in a Language Laboratory.
Jyväskylä: University of Jyväskylä, 1997, 185 p.
(Studia Philologica Jyväskyläensia,
ISSN 0585-5462; 43)
ISBN 951-39-0070-3
Yhteenveto: Punnitaan puhetta. Kielistudiokoe lukion vieraan kielen suullisen taidon päättökokeena.
Diss.

Efforts to increase the teaching of oral FL skills in the Finnish senior secondary schools have often been less successful, mainly because oral testing has not been part of the influential national school-leaving examination. Therefore, an attempt was made to develop and try out an oral test of FL English, which could be used to test all secondary school-leavers - more than 30 000 at a time - simultaneously.

In order to develop a valid test the nature of oral proficiency was analyzed. Bachman's model of language ability was chosen to be the basic framework, and as criteria of proficiency the following features were used: pronunciation, fluency, coherence, amount of information provided, appropriateness of the language. For the instrument of assessment a SOPI type of test, called the LLOPT or Language Laboratory Oral Proficiency Test, was designed. The test has a contextual communicative frame and consists of six parts: warming up, reading aloud a letter, interpreting the Finnish part of a dialogue into English, conveying a Finnish newspaper story in English, reporting on the Finnish school system, and coping with everyday situations and expressing opinions. The whole test lasts 40 minutes, and the recorded sample of the student's speech about 20 minutes. The subjects, 60 school-leavers from two schools, were also tested with the ACTFL interview.

The main research task was to find out whether the LLOPT was a reliable, valid, and efficient means of testing the students. It was also explored whether the LLOPT could be validated with the ACTFL interview. Student attitudes towards speaking and testing English were investigated, as well as the effect of spending time abroad on oral proficiency.

The LLOPT proved to test the students reliably and validly. The correlation coefficient of the LLOPT with the ACTFL interview was .78, and 60% of the subjects received the same result in both tests. The LLOPT turned out to be more efficient than the interview, but the efficiency could be further increased by shortening the test. It was discovered that the ACTFL interview is not a perfect means to validate the LLOPT, because the two tests highlight partly different aspects of proficiency. The students' attitudes towards speaking and testing the foreign language were positive. It could not be shown that staying abroad would have had significant influence on the speaking skill.

The investigation indicates that it would be both feasible and beneficial to start testing FL oral proficiency in the school-leaving examination. At least in the first foreign language the most practicable means would be a language laboratory test.

Keywords: oral proficiency, ACTFL oral proficiency interview, language laboratory, testing spoken language

ACKNOWLEDGEMENTS

When one writes the last words in a dissertation, one cannot help thinking of the numerous people whose encouragement and kind assistance has made the work possible. To thank everyone by name would be impossible and even embarrassing, because writing such a long list would raise the question whether the writer herself had anything at all to contribute. In any case, I wish to extend my gratitude to both those mentioned below and those present only in my grateful thoughts.

The chain of people is long. It may be that the experiences of my late father made me realize for the first time how important it is to know foreign languages. A man with a mere primary school background, he sat frustrated at the meetings of the board of directors of a big company and listened to the Latin quotations circling around him.

I am especially grateful to Professor Kari Sajavaara. It was his expertise in applied linguistics that tempted me to take up postgraduate studies at the University of Jyväskylä. Now I have had the opportunity to enjoy his personal guidance, in which he has shown a deep understanding of both the intricacies of science and the equally complicated ups and downs of the writer's mind.

Docent Sauli Takala has also been a figure who has attracted both linguists and educationalists to the University of Jyväskylä. There has hardly been a doctoral thesis in applied linguistics which would not, at some stage, have passed through his hands. His wide knowledge and extensive library have been readily available to me, too. At the various stages of the work he has offered sensitive critique and pertinent suggestions, for which I am greatly indebted.

Special thanks are reserved for Assistant Professor Irma Huttunen, who kindly agreed to sacrifice part of her summer to read my thesis as an external reviewer. Her comments were subtle and discreet but revealed some essential shortcomings in my dissertation.

Having fun is not an expression commonly associated with doing research, but that is exactly what working with Hanna Jaakkola and Leena Vaurio meant. The idea of cooperative learning inspired us to form a team of "weird sisters" and see what a research coalition of three mature teacher trainers could accomplish. Enlivened by good cooking and animated conversation, the cooperation has so far produced a

licentiate thesis and a doctoral dissertation. In the hope that the synergy will continue I present my contribution in the shape of this book.

Among other good friends whose expertise I have been lucky to profit from I would first and foremost like to mention my dear colleague Heini-Marja Järvinen. Subject matter or form, English or Finnish, day or night, Heini-Marja's reliable advice and faithful support were always available. She also consented to act as one of the assessors, as did my other colleague and unswerving stand-in Irmeli Kaustio. With them I had many instructive conversations about the nature of good speaking. Eivind and Irene Kristiansen took great trouble in reviewing parts of the manuscript and offering many useful suggestions. Professor Heikki Hakkarainen kindly read and commented on the chapter on phonetics, and Ari Huhta kept me briefed about the world news of testing. Were it not for Matti Koponen's enthusiasm I would be far less knowledgeable about fluency. Many thanks to everybody!

An essential part of this research were the schools whose staff and students made the testing possible. The test was piloted in Kaarinan lukio with the benevolent assistance of Tuula Sutela and tried out in Halikon lukio and Joensuun normaalikoulun lukio. At both schools the principals - Matti Lassila and Matti Karjalainen respectively- and the English teachers - Irja Huolila, Jukka Lappalainen, Raija Kangaspunta and Kirsti Relander - showed great flexibility and goodwill in arranging suitable timetables and making all the resources available. Thanks are extended also to the students for showing such openness and excitement.

To an illiterate in statistics like myself the help of Annikki Poutiainen has been of great value. The figures were painstakingly drawn by Jarmo Huunonen and the cover designed by Pirjo Viitanen, whose art has always been a great source of beauty to me.

For financial support I thank the Academy of Finland, the National Board of Education, and the Emil Aaltonen Foundation. Teleste Educational Ltd provided me with the recording material and the Language Center at Turku University helped with the recordings. The text on the master-tape was read by Richard Duke. In addition, thanks are due to the University of Jyväskylä for accepting this work to appear in their series Studia Philologica Jyväskyläensia.

Last but not least, special thanks are due to my friend Juha Talvitie, who has faithfully kicked me onwards in moments of disbelief. He has kept stressing that research is not an end in itself. Had he been less occupied by his work and travels, this publication would not be complete yet.

Turku, August 29, 1997

M.S.

I dedicate this volume to my daughters Kati and Outi. I hope that their generation will be able to live in a world of growing international understanding, to which the knowledge of foreign languages is an essential resource.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The incentive for this study came from practical work. It has long been obvious to the language teacher educator that the emphasis in foreign language teaching should be shifted from written to spoken language, but this viewpoint has not met with unanimous understanding among other language teaching professionals. Many older practicing teachers have felt forced to cling to tradition. It has been argued that as long as the spoken language is not tested in the final school-leaving examination, the matriculation examination, teachers have to teach what best prepares the pupils for the examination. One way to forward the desired change was in this case to begin to work for reforming the examination itself.

The great advances in technology and communication have increased the need for foreign language teaching in general and increasingly brought people to face-to-face contacts. While school has traditionally concentrated on teaching the written form, at the end of the 1980s the Council of Europe for Cultural Cooperation (CDCC) proposed that the emphasis of foreign and second language teaching should be given to the development of the oral skills. It proposed that all important language examinations should contain a speaking test (*Suullisen kielitaidon kokeen työryhmän raportti*, 1989; Takala 1993). Some countries have had an oral part in the school final examinations even before, while others have now followed the CDCC suggestion. Finland is among the few countries where the speaking skill is not assessed in the final examination (Pohjala 1995, 13). Since 1995 there has, nevertheless, been an opportunity to take a voluntary test in many municipalities.

In Finland learners have long expressed a wish to have more emphasis on the teaching of oral skills. Surveys and opinion polls carried out among Finnish students at different levels and among employers have revealed that learners would like the FL[1] instruction and practice to involve more speaking (Koskinen 1994; Takala 1977; Yli-Renko 1985, 1988, 1991; for a compendium of needs inquiries see Suontausta 1993). Luckily the Finnish school administration has answered the public call. It has actually been well ahead of its time, for the National Board of General Education

---

1 The term *FL (learning/teaching)* is in this thesis used to refer to any language other than the speaker's native language. The term *L2* is used about the language which a Finnish child learns as the first foreign language at school (in Finnish: *A1 kieli*).

(later the National Board of Education) made its first proposal to introduce an oral part in the matriculation examination as early as 1958 (Saleva 1993, 8). In the 1980s, when the idea came up once more, the National Board of General Education was again the initiator.

In 1988 the National School Board set up a working party with the commission to study possibilities of arranging an oral skills test in connection with the matriculation examination and to investigate the necessary measures. The result of the undertaking was a report in which the working party suggested that the conditions for teaching the speaking skill be improved and that a project group be appointed to conduct and report on an experiment of teaching and evaluating the oral skills. It was also stressed that the project should be supported by research. In 1990-1994 the seven training schools of the national teacher education institutes and four municipal senior secondary schools were involved in the experiment. It was as a member of this experiment group that I in 1989 began to investigate the possibilities of developing an oral section for the school-leaving examination.

A test is not created for a vacuum and is not good or bad as such. Essential for the quality of a test is how well it fits the purpose for which it is intended. In the case of a national school-leaving examination the criteria are many and demanding. A test like the Finnish matriculation examination is traditionally very influential with plenty of power over the young test-takers' future lives. It should therefore be particularly valid and reliable. Because work in the schools is to a great extent guided by the examination (see e.g. Pasanen 1977), it should also have a beneficial washback effect (cf Messick 1988, consequential validity). It should be flexible and productive so that its basic elements could be used to create similar tests for other languages or new versions for the same language. As the number of testees often exceeds 30,000 per language, the test should be maximally efficient. And naturally, it should be based on the Finnish curriculum for the senior secondary schools.

The two central and inseparable questions of test design are the *what* and the *how*. As part of a school-leaving examination the oral test would have to fit the existing parts, the listening, reading, writing, and grammar tests, which already form quite an extensive measuring apparatus. If one is to believe that the existing sections of the examination work satisfactorily, what would then be the aspects of language proficiency which they do not measure and which would therefore need a new instrument? What is oral language proficiency? The question seemed even more pertinent after the publication of a Finnish dissertation which claimed that school-leavers' oral proficiency could equally well be tested by a writing test (Hellgren 1982; for similar thoughts see Kristiansen 1990; Norris 1991).

To decide what the *what* of a test is, the tester must form a mental image of the phenomenon being tested, X. To construct a valid and reliable testing instrument, the tester can use an existing model of X. A test along with its criteria is an operationalization of the model. The operationalization can be pictured as a lens focusing the light from the model and reflecting it on reality (Figure 1). Lenses A and B (= tests A and B) operationalize only a fraction of the model and they illuminate an even smaller fraction of reality. A strong model (= a valid construct) gives a more powerful light, and a big lens (= a versatile test) throws light on a larger area of reality. If there are two tests, the area they cover of both the construct and the reality is usually partly common, partly separate.

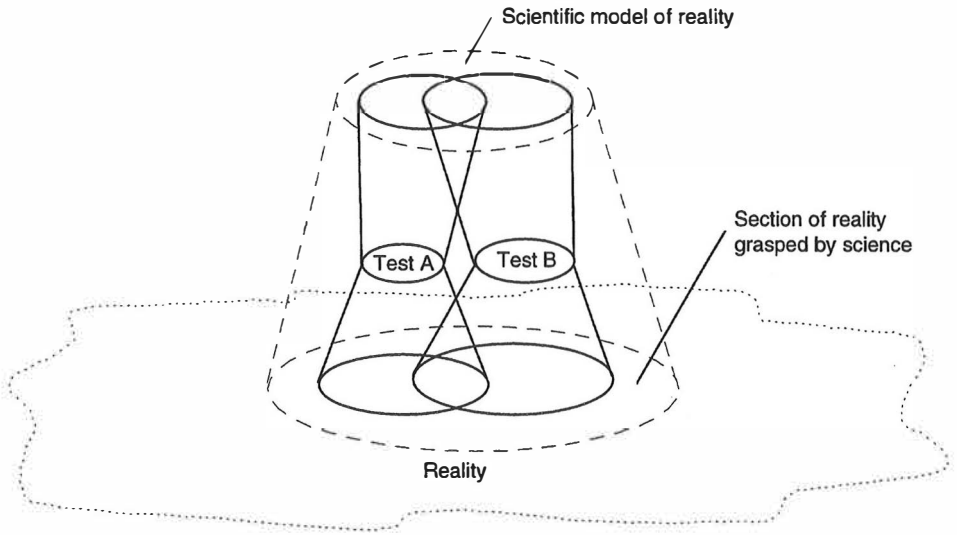

FIGURE 1 Language proficiency reflected by tests. Reality = (here) language proficiency, scientific model of reality = (here) e.g. communicative competence.

For language proficiency the generally accepted current model is the construct of communicative competence. In the present study, too, this model was used to describe oral proficiency, the object of measurement. However, the most commonly used versions of communicative competence do not make any distinction between oral and written proficiency, so part of the *what* question was still open. As there did not seem to be any model of just the oral sector, different descriptions of oral proficiency were studied and an attempt was made to compile a domain specification based on them.

The other principal question of test design is the *how*, i.e. what kind of instrument to use or construct for the measurement. As this particular instrument had to reflect the senior secondary school curriculum, it seemed unlikely that any of the existing tests could be used as such. The main internationally used instruments were an interview, a language laboratory test, and, to a smaller extent, a pair or group discussion. The use of each of them was problematic. As for the interview the main concern was the cost and the required tester expertise, and for the language laboratory test the very availability of laboratories. The pair or group discussion seemed to lack both sufficient validity and reliability. During a study period at the university of Reading the present writer took up the problem of a suitable test with two language testing experts, Arthur Hughes and Cyril Weir. Considering the number of the testees they concluded that the most efficient way in the long run would be to use the language laboratory. Their suggestion was then followed.

The language laboratory test was compiled to consist of the following five parts: reading aloud, interpreting Finnish dialogue, reproducing a Finnish newspaper text, transmitting information, and reacting in situations. Because the language laboratory test did not give opportunity to the commonest form of communication, the face-to-

face interaction, it was decided that an internationally established interview test, the ACTFL, would be used for concurrent validation of the test. Sixty testees, 25 from the Halikko Senior Secondary School and 35 from the Senior Secondary Practice School at the University of Joensuu, were then tested with both the language laboratory test and the interview.

The aim of the language laboratory test and the present study was to answer the following main questions:

1. Can the L2 English oral proficiency of senior secondary school students be assessed validly, reliably, and efficiently using a language laboratory test?
2. Can the language laboratory test be validated by means of the ACTFL oral proficiency test?

In connection with the main questions two further questions were asked:

3. What are the students' attitudes towards speaking English and testing speaking?
4. Is the students' oral proficiency improved by having had an opportunity to stay in an English speaking country?

The language laboratory test is here considered a proficiency test and not an achievement one. It is true that the test is supposed to be based on the senior secondary school curriculum, but the new 1994 version of the curriculum only gives the general framework, and every school is free to choose the exact contents. Besides, a great deal of English may be learnt outside school. This made the concept of oral proficiency even more central in the present study. The writer carried out a domain specification of the nature of the concept, and it seemed to confirm the everyday hypothesis that the factors which distinguish oral proficiency from writing proficiency are the following: rules of speaking, pronunciation and nonverbal features, and fluency. A large part of this study is therefore devoted to the description of these aspects.

The latter part of the study describes the use of the ACTFL oral proficiency interview and the compiling and carrying out of the LLOPT (language laboratory oral proficiency test) and the results of the two tests. There is little documentation of the success of ACTFL interviews as carried out by non-native testers. The account of the procedures in the two Finnish schools may thus be of interest as such. It is hoped that the study of the LLOPT results indicates whether language laboratories as testing instruments would be worth the investments. The results of the two schools are also presented separately to see whether the differences that the pupils showed in the written school-leaving examination have an equivalence in speaking. Because recent studies of lower level school achievements in English and Swedish (Huttunen & Kukkonen 1995; Karppinen & Sarkkinen 1995; Pasanen & Hietanen 1994) have shown notable differences between boys and girls, the genders are also studied separately.[2]

---

[2] Since the majority of the subjects in the present study were females, the pronoun *she* is used to refer to them. Otherwise both *he* and *she* are used without special significance.

2 WHAT IS LANGUAGE PROFICIENCY?

One of the crucial elements of a good test is test validity, above all construct validity. To be valid the definition of a construct has to be based on a sound theory of language and language proficiency. However, though numerous linguists have attempted to create a comprehensive theory, none of them have yet succeeded in formulating an altogether satisfactory construct, which could be used as the basis for test design. Nevertheless, there is some agreement: for the last twenty-five years, most research seems to have concentrated round one concept, that of communicative competence. There are several definitions of it, but I will in this chapter retrace the development of the concept as it has been formulated by Hymes, Canale, Swain, Bachman, and Palmer. Their thinking has been described and analyzed by many writers, from the testing point of view by, for example, Huhta (1993) and McNamara (1996).

In attempts to formulate a theory of language proficiency researchers have been concerned with both the content domains of knowledge and the way the various elements interact and are processed. Though the concept of declarative and procedural knowledge belongs to the nomenclature of cognitive psychology (see e.g. Anderson 1985), the conception of this duality has existed in linguistics since Saussure. He spoke of *langue*, the system of language, and *parole*, the utterances people actually produce when using the language. Chomsky's distinction of *competence* and *performance* (1965) is similar, but not the same. While Saussure was mainly interested in language as corpus, Chomsky was also concerned with the underlying competence. He describes the language learner as an active and creative being, who has an innate device for acquiring language. It is this innate competence that makes it possible for the learner to produce and understand verbal behavior, including expressions that he has never met before. Competence makes him ideally able to produce correct language even though his performance, his actual language use, may occasionally be imperfect.

Chomsky's theory became very influential, but it was not universally accepted. One of the controversial points was the relationship between knowledge and performance. To account for the actual production of language Chomsky had made a further distinction between grammatical competence and pragmatic competence. Pragmatic competence was seen as a kind of mediator between grammatical

knowledge and actual utterances. Nevertheless, there were many who claimed that Chomsky's theory was idealized and did not pay attention to the reality of communication, in which the ideal speaking situation may be disrupted by various extralinguistic as well as inter- and intrapersonal factors. Not only the grammatical use of the language, but also its usage, the speech and writing habits of the community, should be taken into consideration.

Chomsky's opponents were mainly sociolinguists and stressed such features as language usage, appropriateness, and constraints (Crystal 1991, 271). It was one of them, Hymes, who defined his concept of language proficiency as the notion of *communicative competence* (1972). He was referring to the native-speakers' ability to produce and understand sentences which are appropriate to the context in which they occur. Like Chomsky he made a distinction between the model - the knowledge and capacities for use - and the actual use in real-time situations (McNamara 1996, 54-7). But in Chomsky's juxtaposition competence refers to the speaker's internalized grammar of language and performance to his actual use of it. In Hymes's model ability for use is something that underlies the performance and is thus part of the model of communicative competence.

Though Hymes was speaking about native-speakers, his construct was eagerly welcomed in foreign/second language teaching. In the 1970s it was common to refer to up-to-date foreign/second language teaching as communicative, but the model of communicative competence was actually introduced into the theory of L2 teaching and testing by an article of Canale and Swain in 1980 and one by Canale in 1983.

According to Canale and Swain (1980), communicative competence consists of three factors: linguistic competence (morphology, syntax, semantics, and phonology), sociolinguistic competence (sociocultural rules and textual rules), and strategic competence (ability to make up for a lack of knowledge of grammar or vocabulary in a communication situation). In a later article Canale (1983) develops the concepts further and distinguishes two parts within sociolinguistic competence: textual competence (coherence and cohesion), and sociolinguistic competence (on the one hand the appropriateness of meaning: to what extent are language functions appropriate to the situation; on the other hand the appropriateness of form: to what extent has the meaning been put into a form that fits the sociolinguistic context, a division which existed also in the Canale and Swain model).

Unlike Chomsky and Hymes, Canale and Swain (1980) do not explain how the knowledge, which is represented in the three competences, is processed so that it appears as utterances in actual usage. Canale (1983), for his part, makes a distinction between actual communication and the knowledge and skills underlying it. Like Hymes, he considers both knowledge and ability for use to be part of communicative competence. As for processing and interaction he is rather vague, but sees strategic competence as operating in relation to the other three competences. He regards strategic competence as universal: it is acquired as part of learning the native language, whereas sociolinguistic competence has both universal and language-specific features. However, both have to be developed also in connection with the foreign language learning process.

Partly at the same time with Canale and partly later, another well-known version of the construct of communicative competence was being developed. The developers, Bachman and Palmer, even tried to validate the construct empirically

(1981, 1982). They continued their research either together (e.g. 1983, 1996) or Bachman did it alone (e.g. 1990, 1991a, 1991b, 1991c). It is particularly the 1990s versions that have brought new dimensions to the model. The test in the present study was designed after the 1991 articles and the text below is mainly based on them.

As a linguist Bachman is primarily a tester, and thus his theory of language ability forms part of his theory of language testing. He claims that to ascertain the authenticity and validity of language testing, we have to be able to define and assess the relationship between performance in the testing situation and language use in other situations (Bachman 1991c). For that purpose we have to develop a common framework for language use and language assessment. Within this framework, the performance in the testing situation is regarded as a special occasion of language use. The framework has to include both a model of language ability and features of the testing method. As other factors that contribute to the test result he mentions random factors and personality factors, in the latest models also affective factors (Bachman 1990, 1991a, Bachman & Palmer 1996). For the factors in language use and language test performance see Figure 2 (in the figure test method factors are not included). Bachman's concept of language ability will be discussed here, whereas test methods are dealt with in Chapters 6-8.

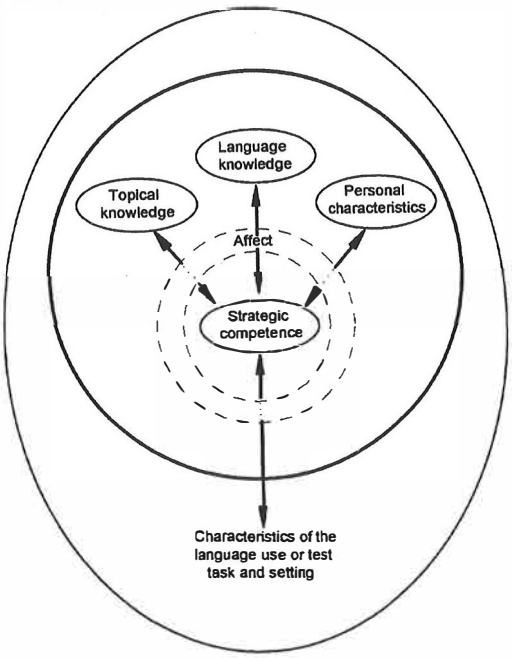

FIGURE 2 Some components of language use and language test performance (from Bachman & Palmer 1996, 63)

Bachman sees language ability as an essential component in language use (1991c, 682-3). (To avoid the too many connotations of the word 'competence' Bachman has replaced it by 'ability.') He defines *language ability* as a capacity to create and interpret meaning in language use contexts with the help of language knowledge. Language ability consists of two factors: *language knowledge* and

metacognitive strategies (strategic competence), which make it possible to put the knowledge into procedural use. Language knowledge is a cognitive component that is distinctive to the linguistic area only, whereas metacognitive strategies are also used in other mental processes.

Language knowledge is divided into two main areas: organizational knowledge and pragmatic knowledge, a division similar to one found in cognitive psychology (see for instance Sternberg 1985, 1988). *Organizational knowledge* determines how texts are organized, and there are two domains in it: grammatical knowledge and textual knowledge. The former covers the organizers of individual utterances, i.e. syntax, phonology, and graphology, while the latter comprises the devices by which utterances and sentences are joined to form texts, that is cohesion, rhetorical and conversational organization.

Pragmatic knowledge describes how utterances or sentences and texts are related to form meaning, and it consists of three domains of knowledge: propositional knowledge, functional knowledge, and sociolinguistic knowledge. Propositional knowledge directs how utterances and sentences are connected to the propositional content, and contains the vocabulary in the 1991 version of the model. In the 1996 version propositional knowledge is left out, and vocabulary is included in grammatical knowledge. Functional knowledge determines how utterances or sentences and texts are related to the communicative goals of language users, and sociolinguistic knowledge describes the relationship to such linguistic conventions as register and style as well as culture-bound references and idioms (in the 1996 version also dialects and varieties). It is sociolinguistic knowledge that guides the user to express himself in a way which is appropriate to the situation and context.

In its present form Bachman's model is extensive and diverse. It is further complicated by the fact that none of the language ability domains functions alone, but is always interrelated with the rest. Just as the language knowledge part consists of various categories and subdivisions, the strategic section has a complex structure of different substrategies which function simultaneously and are in an intrinsic way connected with one another and the different components of language ability. Though the model is both theoretical and, to a certain extent, empirical (Bachman 1991c), it is as such hardly possible to operationalize. The part that seems particularly difficult to concretize is the strategic component in its variety.

Though Bachman and Palmer have originally developed their model for testing purposes (Bachman 1991c, 680-1), from what Bachman writes (684-7) it seems unlikely that the whole model was even meant to be concretized to serve as a basis for test creation. What it does is to give a tentative explanation of the variance in language ability, which is manifested in test results. In addition to the fact that different language users possess different degrees of language knowledge, the extent to which they can make use of metacognitive strategies varies over time and task not only interpersonally but also within the same individual. The total variance in language test results is the product of language ability, on the one hand, and of the testing method, on the other.

Since neither Bachman and Palmer nor anybody else have succeeded in completing a final model of language proficiency, a test designer can either try to design one himself or base his test on one of the imperfect models. Because the present writer had neither the possibility nor the competence to design a model of her

own, she decided to use Bachman's construct as the general basis of test design. The model itself has probably not yet found its final shape, which meant that only a broad sketch could be made. What was left out was also to some extent decided on the principle of testability.

The hypothesis that strategic competence is a central agent functioning between the other elements was accepted as a main principle, but it was not considered to be either easy or necessary to assess metalinguistic skills in a proficiency test. Of the five knowledge components all except functional knowledge were tested. The tester did not know of any model to be used to operationalize it, and neither could she develop one that could have been used for the purpose. However, there was the feeling that functional knowledge is covertly present all the time and affects the result in subtests such as interpreting dialogue or reacting in situations.

In the final version also another component was left out. It had been planned that textual knowledge would have been assessed in Subtest 4, but the attempt did not succeed. Instead, fluency was tested. According to Faerch, Haastrup, and Phillipson (1984), fluency seems to be a distinctive factor in language test variance and is included in their model of language proficiency. Their model is also based on Canale and Swain. The rest of the elements it contains are phonology, orthography, grammar, vocabulary, pragmatics, and communication strategies. In the Bachman and Palmer model there is no mention of fluency, but in their thinking fluency might be seen as a result of the metacognitive strategies functioning efficiently. Table 1 shows which domain of knowledge was mainly tested in the various subtests.

TABLE 1 Domains of knowledge assessed in the LLOPT subtests

Domain of knowledge	Assessment
Grammatical knowledge: syntax	Subtest 2
pronunciation	Subtests 1 and 3
Textual knowledge	(Subtest 4)
Propositional knowledge	Subtests 3 and 4
Functional knowledge	-
Sociolinguistic knowledge	Subtest 5
* * *	
Fluency	Subtests 1, 3, and 4

The development of the construct of communicative competence shows language proficiency as a more and more complex phenomenon. In the same way as structuralism had stressed the division of language corpus into a multiplicity of individual items, cognitive theories describe language proficiency as a network of different kinds of knowledge and skills, yet united by a common factor like strategic competence. From the testing point of view this implies a demand for many and versatile tests. However, along this line of diversity there were other researchers who stressed the unity of language skills. Chomsky himself considered the apparent diversity to be only a phenomenon of the surface structure, which disguised the common basic unity in the underlying deep structure. A similar conclusion was reached by Carroll (1961), and later by Spolsky (1973, 173-5), whose argumentation was based on psychological considerations of the factors of language proficiency.

The idea of unity culminated in the works of Oller, whose conception of the nature of language proficiency was known as the *unitary hypothesis*. According to him (Oller 1979, 24-5) all verbal activity is based on an internalized expectancy grammar, a concept partly derived from cognitive psychology. The spoken and written language occur in such a sequence of elements that it is possible, in a context, to partly hypothesize which element is likely to appear as the next. By elements Oller refers to sounds, syllables, words, phrases, sentences, paragraphs, or more extensive units of text. When using the language, the speaker is all the time making and testing hypotheses about the message, which are based on the knowledge that he already has about the language on the one hand and about the real world on the other. The linguistic conclusions are made possible by the redundancy of language. The testing of hypotheses and the whole language use, listening, speaking, reading, writing, and thinking, is based on one indivisible competence.

Oller's ideas aroused great interest and inspired research. In the following years he found supporters whose theories about the unitary competence were mainly based on factor analysis. There were, however, also opponents. Vollmer (1983) has collected material from 17 articles based on factor analysis during 1965-82. His inquiry shows that the researchers before Oller had generally claimed that competence consists of several factors, and even round 1980 most scholars seemed to think so. The dominion of the unitary hypothesis remained altogether brief, for soon Oller himself (1983) withdrew his statement and admitted that his former theory had been based on a faulty interpretation of a statistical method. He then took a mediating stand and maintained that the multifactoral concept of language proficiency and the competing concept of the indivisible competence are not incompatible but complementary. The global factor of language proficiency is dependent on the contributory factors, while the contributory factors can be meaningfully distinguished from one another only in relation to the realization of a more comprehensive goal in which they are all integrated (Oller 1983).

Oller's unitary hypothesis also influenced language testing in Finland: Hellgren (1982) based his proposal for a school-leaving test on that concept (Chapter 6.3). Now the situation is different; the concept of communicative competence has been established and conclusions about communicative testing have been drawn (Morrow 1979; Weir 1988). For the present writer it was natural to construct her plan along the established communicative line (for the principles of designing the test see Chapter 7).

3 SPEAKING

Writing was long regarded simply as spoken words written down, and perhaps this was the case when writing first occurred. Later on, however, the two skills developed into two markedly separate modes of expression with distinct features of their own. According to the analysts of the speaking mode (Halliday 1985; Takala 1983 Tiittula 1992), they are not just alternative ways of doing the same thing, but ways of doing different things. Though the distinct border between the two has again begun to blur, from the testing point of view we must still regard them as so separate that they have to be assessed with separate instruments developed for each genre. In this chapter the nature of speaking will be discussed with special attention to the features that have a bearing on assessment.

3.1 Conditions of speaking

There are two factors that particularly affect the production of speech and make the conditions of speaking very different from writing. The first of these is related to the internal conditions of speech, the fact that speech takes place under the pressure of time. The second involves the dimension of interpersonal interaction. Bygate (1987, 7) calls the constraints caused by the pressure of time the *processing conditions*, whereas the circumstances brought about through the participation of two or more people are called *reciprocity conditions*.

3.1.1 Processing conditions

The ability to produce speech at a normal speed under pressure of time means that the speaker has to make quick decisions, implement them smoothly, and adjust the conversation when unexpected complications appear. Nevertheless, this is generally not a problem in the first language, where constant practice facilitates the automation of subskills (McLaughlin 1990). In a foreign language, however, particularly if the

learners have used the language mainly in writing, often with heavy emphasis on accuracy, the on-line production usually involves difficulties.

Even in the mother tongue the fact that there is less time in speaking than writing to plan, organize, and execute the message means that the speakers are often monitoring their phrasing and meaning as they speak. Traces of this monitoring can be noticed in the discourse produced. According to Bygate (1987, 14), this gives rise to four common features of spoken language:

1. Spoken discourse is generally described as more simple than written text. The simplicity is to be noticed in both vocabulary and syntax. In vocabulary some general nonspecific words and phrases are very common (*got, nice, a lot of, a bit, sort of, thing, and so on*), and short Anglo-Saxon words are on the whole more frequent than in written texts, which have plenty of complex Latinate words. In syntax the sentence structure is paratactic (unsubordinated) with sentences marked as related to each other not so much by syntactic devices as by the way speakers say them. If conjunctions are used, coordination is more common than subordination, *and* and *but* being the most common conjunctions. Time is often expressed by time adverbials rather than by tenses. In narrative discourse, it is not always necessary to indicate changes of time, and the historical present is often the natural tense. Heavily premodified noun phrases and accompanying post-modifications as well as heavy adverbial modification are avoided. Instead of relative clauses, deictics or stacking of nouns is used. All these *simplification* features make the information seem much less densely packed than in written discourse. (See also Brown & Yule 1983b, 4-7; Brown, Anderson, Shillcock & Yule 1984, 15, 88; Chafe 1982, 36-49; Hatch & Long 1980, 13; Leech, Deuchar & Hogenraad 1982, 135; O'Donnell 1974, 102-109; Owen 1990, 244.)[3]

2. A time-saving device in spoken language is *ellipsis*. When the spoken utterance is shorter than the written one, people seem, in the pressure of time, to follow the road of minimum effort. On the phonetic side there are contractions such as *it is > it's, you shall not > you shan't*. The context makes it possible to omit parts of a sentence and use syntactic abbreviations; *The big one, On Saturday, Why me?, Does what?* Sentences and clauses are often left "incomplete". The abbreviated expressions do not, however, lead to misunderstandings, because the speaker and listener possess a great deal of shared knowledge. (Bygate 1987, 16.)

3. It is easier for the speaker to produce his message if he uses fixed conversational phrases. Particularly in routine situations such as greetings and partings he can use *formulaic expressions*, which saves him from monitoring his choice of words. (Cf. section 5.2.2)

---

3 However, both spoken and written discourse are so multiform that it is difficult to generalize across the codes. Some writers argue for the complexity of spoken syntax (Takala 1983, 12). The context and the register seem to be more decisive than code as such (Beaman 1984, 78-9; Järvinen 1988, 15; Takala 1983, 61). There are also mixed modes such as formal lectures and written stories or letters (Tannen 1982) and email texts.

4. Simplifications, ellipses, and formulaic expressions do not generally suffice to give the speaker all the time needed for the planning and production of his message. To show that he wants to continue his speaking turn, he can resort to *fillers* and *hesitation devices* such as *well, you see, kind of, erm.* (Bygate 1987, 14.) Another means to get more planning time is *repetition*, which is a common feature in spoken language (Brown & Yule 1983b, 9; Bygate 1987, 20).

In spite of the facilitating strategies the message does not often seem to turn out the way the speaker intends. He has, however, an opportunity to compensate for the defects, to use repairing devices. He can *start again, correct himself* and *rephrase* what he has said. His repetitions may contain *expansions* and *reductions*. When he lacks the needed linguistic element, he can resort to *paraphrasing* and other compensation strategies (Bygate 1987, 19; Faerch & Kasper 1983).

Looked at from the traditional linguistic point of view, these processing conditions make spoken language seem less perfect than written products. There are, however, great qualitative differences in the clarity of spoken discourse that have very little to do with grammatical and lexical accuracy. Deictic and other referential indicators have an important role. We know from mother tongue experience that there are speakers from whose speech it is not easy to understand which way to go to the police station or who did what and when. For directions, for example, it is important to learn to use prepositions and adverbs of locality, such as *above, in the middle of, on the opposite side.* To tell a story coherently, the speaker has to indicate clearly which characters or objects are involved at a particular point, to describe the main activities or events, and to indicate when any significant changes in time or place occur. (Brown, Anderson, Shillcock & Yule 1984, 153-5.)

As long as the foreign language student learns specificity of reference, preferably originally in the native language, he need not be too worried about syntactic accuracy in the spoken foreign language. Bennett (1977, cited by Hatch & Long 1980, 13) has in fact suggested that native speaker talk data of unplanned discourse share many features with pidgins, creoles, and second-language learner talk. In evaluation, learner talk in natural conversations (if conversations in a testing situation are ever natural) should not be compared with standard English based on planned or written text but with natural native speaker conversations. Students who have spoken the language only in school situations, run the risk of speaking too formally rather than casually. According to Brown et al. (1984) this is particularly true of Scandinavian students speaking English. Speaking too formally is a feature that will probably also show unfavorably on fluency, and, in the writer's opinion, if penalized in language assessment, should be dealt with in the criteria of fluency. For evaluation purposes no absolute rules of formality can be set, but the aim and the conditions of the speech situation must decide.

3.1.2 Reciprocity conditions

Writing and speaking are both interactional. Even the writer of a diary usually has a recipient in mind, maybe only his future self. In speaking, however, the listener is more often than not present. It is vital for the speaker to be aware of, and to heed, the feelings, expectations, and background knowledge of the listener. The feedback that

the speaker gets from the listener affects his production of message all the time. Interaction is negotiation of meaning.

To the speaker consideration of the listener means, first of all, choosing an appropriate level of explicitness and detail. The listener wants neither too much information nor too little. It is for the speaker to estimate what his interlocutor knows, what he needs to know, or can understand. According to the feedback he gets, the speaker adjusts his message, repeats and clarifies things, and uses paraphrases and different expressions. To accomplish this the speaker has to be aware of what has gone before and what he expects to follow. In this sense a speaker must be much more flexible than a writer. Bygate (1987, 29-33) uses the term "a good communicator" and describes the routines needed for keeping up good communication as interaction routines. For the speaker they involve among other things:

- announcing or indicating one's purpose in advance
- indicating friendliness
- checking that the other person has understood
- asking the other person for information or language that he has forgotten
- asking the other person's opinion
- responding to requests for clarification from the listener(s), for instance rephrasing, repeating, giving examples or analogies
- checking common ground
- adapting to points made by the interlocutor
- clarifying meaning or intention by summarizing.

From the listener's point of view there is a similar set of responses that complement the preceding ones. Lack of the ability to continuously adjust one's speech to the feedback from the partner may give the impression that the non-native speaker is rather stiff, formal, indifferent, or slow.

For the tester this kind of natural interaction poses a problem. It is not easy to create a speaking situation in which all the above conditions would genuinely materialize. The best arrangement is a homogeneous pair conversation, but it is not enough to guarantee the naturalness of the exchange.

3.2 Principles of cooperation and politeness

When a person engages in verbal interaction, he has a goal, be it nothing more than being friendly. To reach the goal, however, he needs the acceptance and collaboration of his interactant. It is therefore natural that cooperation and politeness should be the two central principles of verbal interaction. They are not often specifically tested, but their absence in speech products leaves the assessor unsatisfied.

3.2.1 Cooperation

The principles of cooperation cited in almost every article or book concerned with spoken interaction (see e.g. Brown & Levinson 1987; Cook 1989; Coulmas 1981; House & Kasper 1981; Laver 1981; Leech 1983; Levelt 1989; Levinson 1983; Richards & Schmidt 1983) were laid down by Grice in his William James lectures at Harvard in 1967 (partly published, see Grice 1975). As the rough general principle which the participants in conversation are expected to observe Grice formulated what is called the cooperative principle: "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged." From this general principle Grice deduced the well-known maxims of quantity, quality, relation, and manner.

It is not surprising that, with their wide spread, Grice's maxims have also met with criticism. A question of interest from the language learning and testing point of view is to what extent these principles - or any sociopragmatic principles - are universal. If they were universal, there would be one cause less for cross-cultural pragmatic failure (Tannen 1984; Thomas 1983) and one object less for oral testing. Leech (1983, 10) claims that Grice's maxims are rational in such a general way that one would expect them to be universal. It has, however, been argued that there are linguistic communities to which not all of them apply (Keenan 1976). As long as empirical evidence is scarce, most writers seem to take a somewhat cautious stand, accepting the general core of the principles but admitting that they operate variably in different cultures or language communities, in different social situations, among different social classes, etc.

An example of a maxim to which different cultures seem to react differently is the maxim of quantity. It is generally claimed that some cultures are more talkative than others (see e.g. Tannen 1985). An American is more prone to comment on what he sees than for instance an American Indian or a Japanese. An Arab text is full of elaborate sayings, which seem to lack the equivalent elsewhere. It is sometimes hard for the foreigner to keep the right balance between verbosity and silence. It would seem likely that a non-native speaker would tend to speak too little rather than too much, but this is not always the case. Excessive politeness may have an effect which is the opposite of what the speaker/writer had in mind, and foreigners can also be accused of too elaborate speech and redundancy (Blum-Kulka & Olshtain 1986). The test situation itself will naturally affect the quantity or quality of speech. Fear, anxiety, and the formality of the situation are not apt to elicit natural speech (Hembree 1988; Madsen 1982; Madsen, Brown & Jones 1991; Wine 1980; Young 1986).

3.2.2 Politeness

Even many writers who have been critical of Grice's cooperative principle have taken it as a starting point and expanded it. The most common addition seems to have been some notion of politeness. In analogy to Grice's maxims, Robin Lakoff (1973, 298) stated the following "rules of politeness": 1. Don't impose, 2. Give options, 3. Make X feel good - be friendly. Leech (1983, 79), for his part, felt that what would "rescue the cooperative principle from serious trouble" would be the adding of a complementary principle of politeness. Within the politeness principle he distinguishes various

maxims such as tact, generosity, approbation, modesty, etc. According to Leech (1983, 149-50), these maxims, together with the principle of cooperation, are the general functional "imperatives of human communication", and so more or less universal, but their relative weights will vary from one cultural, social, and linguistic milieu to another.

According to the *Oxford English Dictionary* the adjective *polite* means having or showing that one has good manners and consideration for other people. However, in the study of language and communication it has a more extensive use: it refers to the ways in which people pursue social and interpersonal goals in interaction (Piirainen-Marsh 1995a). To Leech (1983) politeness is something very rational and straightforward. It is a device used to reduce friction in personal interaction, something needed to maintain the social equilibrium and the friendly relations which enable us to assume that our interlocutors are being cooperative. To put matters simply: unless you are polite to your neighbor, the channel of communication between the two of you will break down and "he will no longer lend you his mower". This is more or less what many other writers feel about politeness, too. Thus for instance Gumperz (1987, xiii) stresses that politeness is basic to the production of social order and a precondition of human cooperation. According to Levelt (1989, 65), most talk can be successful only if the speaker respects or takes into account the rights, capabilities, propensities, and feelings of the other parties.

Politeness markers are often seemingly insignificant, such as mere intonation or a word muttered more or less automatically in passing by. However, their importance becomes conspicuous the moment they are omitted or not acknowledged. The same applies to foreign language speakers. They are excused accent, mistakes of grammar, and errors of vocabulary but not trespasses against good demeanor. The explanation is that among native speakers the rules of behavior are generally below the level of conscious awareness. A breach of rules arouses considerable ambiguous anxiety. (Ferguson 1981, 24; Wolfson 1983b, 63.)

The awareness of the fatal results of lacking pragmatic knowledge explains the recent years' surge of interest in the notion of politeness. There seems to be a consensus about the fact that communicative competence in a language involves an understanding of the appropriate politeness strategies. To be highly proficient in pronunciation, grammar, and vocabulary may be even dangerous, because if such a person violates the native speakers' norms of politeness, it may be perceived not as evidence of a lack of proficiency, but rather as a "sign of disrespect, hostility, or other negative attitudes" (E. E. Davies 1987, 76). This is confirmed by other scholars, who conclude that deficiencies in sociolinguistic, contextualized language competence are seen to be more serious than mere structural errors because they are felt by the native-speaker community to reflect adversely upon the personality of the L2 speaker (e.g. Janicki 1982, 54; Seelye 1974, 53; Thomas 1983, 96-7).

As Grice, the philosopher, did fundamental work for linguistics in setting up the principle of cooperation, an equally important contribution was given by the anthropologists Brown and Levinson when they published their research on three very different languages and stated the universal principles of politeness in 1978 (1987, revised edition; see also Piirainen-Marsh 1995a). Grice (1975, 47) himself had mentioned other maxims, such as the social one 'Be polite', as a source of

unconventional implicatures, but it was Brown and Levinson who established the principal concepts now associated with politeness.

Brown and Levinson's central concept is *face*. All human beings who want to enter into social relationships with each other must acknowledge the face of other people. Face here is equal to the positive image or impression of himself that a person shows or intends to show to other people. According to Brown and Levinson (1987, 67), *negative face* is the wish of every "competent adult member" that his actions be unimpeded by others, whereas *positive face* is the wish of every member that his wants be desirable to at least some others. Negative face stresses the interactants' autonomy and freedom of action, while positive face is connected with the participants' need to be accepted by others and to share things with them (Piirainen-Marsh 1995a). Social contacts between people involve *face-work*, that is, efforts by the participants to communicate a positive face and to prevent loss of face (Richards, Platt & Weber 1985, 102). The more serious the situational threat to a person's face is, in situations of maximum risk, the more face-work, redress, is needed. Again, to do this, also a foreign language speaker must know these target culture routines well.

Face-work that is directed to the addressee's negative face is called *negative politeness*, while the opposite is called *positive politeness*. Negative politeness strategies indicate deference and social distance between speaker and hearer. The imposition of a face-threatening act on the hearer has to be minimized. The hearer must be given options. A very central device particularly in British negative politeness strategies is using indirect expressions. According to Leech (1977,19), "the more tactful a directive is, the more indirect and circumlocutory it is". Many nations, for example Germans, Russians, and Lithuanians seem less polite to the British, which is mostly due to the fact that their language use is more direct (Drazdauskiene 1981; House and Kasper 1981; Thomas 1983).

Strategy	Syntactic/semantic features	Examples
7	Interrogative-Past tense modal	*Could you give me a pack of Marlboros?*
6	Interrogative-Present tense modal	*Can you give me a pack of Marlboros?*
5	Interrogative-No modal	*Do you have a pack of Marlboros?*
4	Declarative-Past tense modal	*I'd like a pack of Marlboros.*
3	Declarative-Present tense modal	*I'll have a pack of Marlboros.*
2	Declarative-No modal (*need/want*)	*I want a pack of Marlboros.*
1	Imperative	*Give me a pack of Marlboros.*
0	Imperative-elliptical	*A pack of Marlboros.*

FIGURE 3 Theoretical hierarchy of request strategies (according to Carrell & Konneker, 1981, 20-1)

Using direct versus indirect expressions can be presented as a continuum, as the table of requests by Carrell and Konneker (Figure 3) shows. The more freedom the addressee has to refuse the request, the more polite it is. It is also hypothesized that politeness increases with the complexity of the surface syntactic markers.

Other strategies for negative politeness include for instance minimizing the imposition, polite pessimism, and apologies. Positive strategies stress closeness, intimacy, and rapport between the participants. The speaker shows solidarity with the hearer's positive self-image to satisfy the hearer's need for approval and belonging. A characteristic element is exaggeration and praise.

The notions we have about the two adjectives "cooperative" and "polite" would indicate that the two principles are reciprocally supportive. This is most often the case, but there are instances, too, when they come into conflict. The maxim of quantity often seems incompatible with politeness (see for instance example 1 above), and quality and positive politeness seem likewise mutually exclusive. However, even in the case when the latter conflict leads into the use of "white lies", the very basic motive is cooperation.

Which of the two principles seems to dominate is in the end dependent on the contextual features of the situation such as the social role and distance of the participants as well as the length of their acquaintance. Even the same participants will behave differently in a different situation. In an emergency situation, for instance, the efficiency of communication does not allow all the conventional politeness figures. Politeness in formal language situations is different from politeness in situations in which colloquial language is used.

Though the forms of politeness vary from situation to situation, the variation from culture to culture is much bigger. Researchers seem to agree that politeness is a universal phenomenon in human societies but that the forms it takes are very different. The specific nature of face varies from society to society. Culturally different assumptions about how respect and consideration for others should be realized leads to misunderstanding and disappointment. Hasty conclusions based on intuition and superficial acquaintance are the source of notorious national stereotypes such as labeling the Germans as abrasive, the Americans as insincere and vulgar, and the Finns as silent (Bentahila & Davies 1989, 104; Sajavaara & Lehtonen 1985, 1997; Thomas 1983, 97). The positive thing about such clichés is that they may have aroused the interest of researchers and given rise to important cross-cultural studies, as for instance Blum-Kulka, House & Kasper 1989.

There are many areas of verbal interaction where ethnically different assumptions of politeness are to be seen. One area that is constantly interwoven into verbal behavior is culturally different attitudes toward *speech* and *silence* respectively. In an article *Silence: anything but* Tannen (1985) describes a Thanksgiving dinner with six participants who represent two different conversational styles. The three New York Jews represent a talkative, even noisy style with a fast rate of speech and turn taking, and tolerance of, even preference for simultaneous speech. The two Californians and one Englishman represent something that Tannen calls "mainstream" American style, which is less talkative, has slower speech rate and turn taking, and is less patient with simultaneous speech. Both groups experienced the other party as impolite. The interesting thing here is, however, that both parties, in their very different behavior, were following their own rules of politeness. The New

York Jews saw the abundant speech as involvement, appreciation and enthusiasm, indicating positive politeness. To them silence meant indifference and lack of interest. The "mainstream" Americans and the Englishman, for their part, regarded the abundant speech as imposition, infringement of their independence. To them sufficient silence meant showing consideration and deference, which are such central parts of negative politeness.

One of the central problems to the linguist as well as the foreign language learner is the relation between *form* and *function*. If we know how to say *I am sorry* in a foreign language, we still do not know when and to whom we should say it according to the interactional norms of the respective culture. Our knowledge of the corresponding form may indeed lead us to ignore or not recognize the functional constraints on its use so that we transfer the pattern of usage of the equivalent term from our own culture. This kind of transfer of pragmatic rules from one linguistic system to another may lead to inferential mistakes just like any other kind of transfer. This is the trouble with phrase books. If cross-cultural pragmatics is not paid sufficient attention to, the study of foreign language does not lead to increased intercultural understanding but to increased prejudices and hasty judgments of impolite and inappropriate behavior. (See Coulmas 1981.)

With the great number of languages and the abundance of speech events and speech acts, cross-cultural pragmatics will have an inexhaustible field of research to cover. Its results are also likely to prove more significant than those of early contrastive linguistics, whose object was the investigation of grammar and vocabulary. The most fruitful insights seem to come from the numerous studies that have compared some specific functions in two or more cultures. Apologies (Cohen & Ohlstein 1981; Faerch & Kasper 1984; Harlow 1990; Wolfson 1983b), requests (Carrell & Konneker 1981; Harlow 1990; House & Kasper 1981), compliments (Holmes & Brown 1987; Manes & Wolfson 1981; Wolfson 1981a, 1981b, 1983a), thanks (Coulmas 1981; Eisenstein & Bodman 1986; Harlow 1990) and offers and invitations (Conein 1986) seem to be among the most popular subjects. The conclusion from reading the results of these numerous studies seems to be caution: the rules and conventions are really culture-specific and cannot be inferred on superficial observation. Paying excessive attention to culture-specific routines may also lead to the increase of stereotypes (Piirainen-Marsh 1995b).

As to the English language in general, linguists seem careful to go by the culture, almost invariably making a distinction as to whether a research result applies to American English or British English or perhaps to some less common variety such as New Zealand English. In an extensive comparative study, the Cross-Cultural Speech Act Realization Project (CCSARP), led by Blum-Kulka (Blum-Kulka et al. 1989), three of the seven languages included are varieties of English: American English, Australian English, and British English. Studies in which two varieties of English are compared interestingly show that there is a great deal of deviance and misunderstanding between these cultures, too, (see for instance Gumperz's comparison of British English and Indian English (1977, 1978, 1979) and Holmes & Brown's (1987) investigation of compliments in American English and New Zealand English).

To a Finnish tester this kind of information is useful in creating caution as to target culture generalizations. The crucial factor, however, is the question how much

cross-cultural Anglo-American-Finnish research is available. There is a large amount (e.g. projects led by Nyyssönen in Oulu, by Lehtonen, Markkanen, and Sajavaara in Jyväskylä, and the various master's theses at different universities), and the number is increasing all the time. Only little testing of sociolinguistic competence has so far been undertaken. Many scholars seem to agree that much more research is needed before cross-cultural language teaching can be based on reliable scientific knowledge. Others, like Thomas (1983, 97), express their doubts whether judgments of appropriateness can "ever be spelt out sufficiently to be incorporated in grammars or textbooks as other than fairly crude rules of thumb". However, the writer of the present text decided to devote one whole subtest to testing sociocultural appropriateness, but found out later that that section was the most difficult to judge.

3.3 Creativity and convention

One of Chomsky's main ideas was that every individual has an innate capacity for combining elements of language in a way possibly never used before. An important aim of language teaching is considered to be giving learners opportunities for this creative ability to develop also in the foreign language. However, only a fraction of all the potential novel sentences are actually ever used in communication. Communication largely consists of the use of language in conventional ways. There are strict constraints imposed on the creative-constructive capacities of speakers, and these set limits to how speakers encode propositional meanings. Though both *Please post this letter for me* and *I request you to post this letter* are grammatically correct sentences, only the former has a status as a potential utterance, since the latter would never be used by native speakers of English (Richards 1983, 114). It is not enough for a learner to produce a grammatically correct sentence, it also has to be conventionally acceptable. Coulmas (1981, 6) claims that creativity in language should be regarded as an interplay of grammatical rules, functional adequacy, situational appropriateness, stylistic preferences, and norms of use.

Richards, Platt and Weber (1985) give the routine use of language several terms: *formula, formulaic speech/expressions/language, conventionalized speech, prefabricated language/speech*. Drazdauskiene uses the words *stereotypes* and *clichés*, pointing out that in this linguistic sense neither of the words has any evaluative meaning (1981, 67). "Routine" is by Richards et al. (1985) defined as a segment of language made up of several morphemes or words which are learned together and used as if they were a single item.

Formulaic expressions are many. A language like English with its isolating structure is particularly rich in them. It is estimated that there are a few thousand formulaic expressions, which make up 20 per cent of daily conversational exchanges (Coulmas 1981, 9; Pawley & Syder 1983, 205), but the figure is, of course, dependent on the way in which the concept is defined. Besides, formulaic expressions are a continuum from very fixed combinations such as *How are you?* to more or less occasional associations.

Besides being a valuable facilitator in the process of native-speaker communication, formulaic expressions also smooth the beginning learner's path to natural conversation. At a stage when the learner cannot yet actually construct original messages, they make it possible for him to enter into conversation (Wong Fillmore, 1979). Likewise they are a useful tool for the poorest learners, for whom the combining and creative use of the language are singularly difficult (Kristiansen 1992, 14-19).

There are many categories of formulaic expressions such as idioms, memorized clauses, proverbs, compliments, politeness formulae, and ceremonial and (religious) ritual routines. The most frequently occurring, and thus of the greatest importance, are the *conversational routines* or *gambits*. Since it is the task of these expressions to facilitate communication, it is natural that they should occur most frequently in those sections of conversation where it is important to maintain the flow of speech and ensure that the channels remain open, i.e. in phatic communion. Drazdrauskiene (1981,64), who did a comparative study of discourse in English and in her mother tongue Lithuanian, points out how much more numerous these linguistic devices are in English. She also claims that English discourse is much richer in emotional expressions than Lithuanian, a language in which very emotive and superlative evaluations, especially positive ones like *It sounds lovely, That sounds marvelous really,* sound affective and alien. The same would certainly be true of Finnish. Conversely, the use of the Lithuanian or Finnish type of conversational patterns when speaking English would probably sound unenthusiastic and bored.

E.E. Davies (1987, 79) suggests that, in addition to the sociopragmatic level, the user of routine expressions must be aware of their semantic and illocutionary levels. There are cases in which L1 and L2 do not correspond at any of the three levels, so there is no equivalence whatsoever. However, these cases may prove to be less problematic than the ones where there is a partial equivalence of one or two levels and the learner then supposes that the expressions are identical at the one or two other levels as well. An example might be the American phrase *We really must get together sometime* (Thomas 1983, 108), whose semantic content may be clear to the learner, but he does not know that its illocutionary function is not an invitation but just a polite ending of a conversation.

The command of formulaic language may also function as a natural touchstone in testing oral proficiency. Authenticity and fluency are key criteria under which it can be measured. The right greeting in the proper sociolinguistic context is a good benchmark, but an inappropriate use is equally revealing: *To my mind I'll have another cup of coffee* (the example is from Richards 1983, 119).

As gambits are the most frequent in everyday conversations, the testing tasks should offer chances of such activity. The most natural test format with a chance to vary the social status and an opportunity for initiative would seem to be role play.

3.4 Forms of speaking

A valid speaking test has to have many subtests, since speaking has many forms. The many forms can be categorized in different ways, but not all of them are equally important for testing purposes. For instance the terms *acts* and *moves* are important in discourse analysis, but less so in the testing of an individual's speaking proficiency. From the testing point of view, different categories also have different significance cognitively. Also in the speaker's native language some categories, such as argumentation, offer a greater *cognitive challenge* than, for instance, dialogue. In addition, in the realization of some of the categories there is an extra challenge in the foreign language because the way they are produced in one language differs from the use in another language. When compiling the tasks the tester has to be aware of how much cognitive strain the different forms will put on the speaker.

The basic categorization of speech into speech acts, speech events, speech situations, and genres is of importance in testing. *Speech acts* (Searle 1969) are one of the key concepts in modern pragmatics, something that a whole theory is based on. A speech act can be described as what we actually do when we speak, for instance, ask, request, suggest. Speech act theory - basically derived from Austin (1962) - analysizes the role of utterances in relation to the behavior of speaker and hearer in communication. An utterance has two kinds of meaning: propositional meaning (also known as the *locutionary meaning*), conveyed by the particular words and structures which the utterance contains, and *illocutionary meaning*, which is the effect that the utterance has on the listener. (Crystal 1991; Richards et al. 1985.) To make an illocutionary act succeed, the speaker must judge his position relative to his interlocutor by assessing his positions (e.g. roles, status, etc.), properties (e.g. sex, age, etc.), relations (e.g. dominance, authority), and functions (e.g. 'father', 'waitress', 'judge', etc.) (van Dijk 1977; 221).

In different cultures there are subtle differences in realizing speech acts, and it is crucial for interlocutors to learn to interpret the intended speech acts appropriately. As an example of an unsuccessful attempt to interpret the speech act, Richards (1980, 418) offers the following conversation between a professor, B, and a foreign student, A:

A. *Hello, is Mr Simatapung there please?*
B. *Yes.*
A. *Oh...may I speak to him please?*
B. *Yes.*
A. *Oh...are you Mr Simatapung?*
B. *Yes, this is Mr Simatapung.*

Other categories of speaking which are partly universal, partly culture-specific are the speech event and the speech situation. Coulthard (1985) and Richards et al.(1985) define the *speech event* as a particular instance when people exchange speech, such as an exchange of greetings, an inquiry, or a conversation. The components of a speech event are its setting, the participants and their role relationships, the message, the key, and the channel. The term *speech situation* is sometimes used as a synonym of

speech event, but it usually refers to any situation that is associated with speech, such as for example a classroom lesson or a party. A speech situation may consist of just one speech event - for instance two people meeting in the street and exchanging a greeting - or several speech events, which may even be going on at the same time, such as conversations at a cocktail-party. (Coulthard 1985; Richards et al. 1985.)

The structure of speech events varies considerably according to the genre of speaking they belong to. A *genre* is actually defined as a particular class of speech events that the speech community considers as being of the same type. Such categories are for instance interviews, lectures, speeches, poems. Genres are universal to a certain extent, whereas the realization of a particular genre may be different in different speech communities. A genre often has a norm for its structural organization. (Bhatia 1993; Richards et al. 1985; Swales 1990.)

The number of genres that different linguists distinguish varies. Cook (1989, 95), for instance, mentions as many as 40 different types of discourse, though some of them in the written medium. Bygate (1987, 22-23) does not speak of genres but refers to more or less the same concept when he differentiates two kinds of speaking skills, the organizing skills and the negotiation skills. The former refer to the organizing of typical kinds of messages according to certain patterns. The patterns correspond to recurring cognitive problems and help to automate the processing. Bygate calls these patterns information routines and defines them as frequently recurring types of information structures. The routines are of two kinds: expository routines and evaluative routines. *Expository routines* are those that involve factual information depending on questions of sequencing or identity of the subject. The principal types of expository routines are narration, description, and instruction. When suggesting planning tasks for the practicing and assessment of expository routines Brown and Yule (1983b, 109) make the further categorization presented in Figure 4. The figure is presented as an example of task types and not as a complete categorization of expository speech routines. The most typical categories are, however, represented.

1	Static relationships
i	Describing an object or photograph
ii	Instructing someone to draw a diagram
iii	Instructing someone how to assemble a piece of equipment
iv	Describing/instructing how a number of objects are to be arranged
v	Giving route directions
2	Dynamic relationships
i	Story-telling
ii	Giving an eye-witness account
3	Abstract relationships
i	Opinion-expressing
ii	Justifying a course of action

FIGURE 4 Categories of expository speech (in accordance with Brown & Yule 1983b, 109)

Evaluative routines, again, are often based on expository routines. They involve the drawing of conclusion, usually requiring the expression of reasoning. Evaluative routines typically involve explanations, predictions, justifications, preferences, and decisions, and they are used in for instance argumentative texts (Bygate 1987, 23-4).

A metacognitive awareness of the patterns that are characteristic of different genres helps the student to produce the required language. Having copious models and ample practice, the student will learn to organize what he has to say in accordance with the relevant genre. A major dichotomy which cuts through the spoken language and which the student should know is the division into interactional and transactional language. The aim of *interactional language* is to establish and maintain social relationships, to make the interactants feel comfortable and friendly. It is listener-oriented language and will be further described in section 3.5. In *transactional language* the chief goal is the transmission of information, it is information- or message-oriented speech. For the information to be transferred, the listener must understand the message. It must therefore be explicit and well-organized, often with specific vocabulary, while the listener-oriented language can have a much looser structure and more general vocabulary.

The primary function of speech has been social, and the most common form of speaking is listener-oriented, while most written language is message-oriented. So the natural form for a child to learn the language is listener-oriented speech, and information-oriented speech comes only much later and usually not without explicit practice. Brown et al. (1984,12), in their three-year research work on developing the spoken language of Scottish adolescents, found that almost all school-leavers were able to chat cheerfully and cooperatively with a visiting interviewer whom they had never met before, whereas many of them had notable difficulties in producing coherent and easy-to-follow information-oriented speech. Those with difficulties could just cope with a task of reporting a motor accident that involved two cars, but if there were three cars, the task was too complicated for many.

The ever-increasing complexity and technicality of working life and the present world in general have made it necessary for people to cope with more and more complicated communication tasks also orally. In Britain there have been repeated complaints from employers and administrators of the inability of school-leavers to express themselves articulately, i.e. to use information-oriented language properly. The concern has been manifested in an increasing number of conferences and seminars on aspects of "oracy" and in the demand that spoken language should be taught and assessed within the school curriculum. (Brown et al. 1984, 5.) Similar claims have manifested themselves in Finland, and there is a decision to start testing native language oral skills in both junior and senior secondary schools. Producing coherent and effective transactional speech even in one's native language is such a cognitively complex task that it requires specific training; using such speech in a foreign language will be comparatively easier if the training is given first in the native language (Brown & Yule 1983b, 19).

A major difference between interactional and transactional discourse is that the former generally consists of *short turns*, and the latter of *long turns*. A *turn* (see 3.5.1) is the period of time that each speaker has the floor. Telling a story or a joke, giving route instructions, reporting an accident to the police, or taking a stand in a debate are all forms of speech that typically require a long turn. A short turn, on the other hand,

consists of utterances of one word to one sentence. FL learning is usually begun with short turns, but in the complicated real-life study and work situations also presentations are necessary.

From the testing point of view it is important to note that the structure of long turns is different from that of short turns. Short turns need little planning and show an ample display of features typical of spoken discourse, such as co-ordination, time adverbials instead of tenses, deictics, and stacking of nouns instead of relative clauses. If, on the other hand, the speaker has to hold the listener's interest for more than a sentence or two, he has to organize the information so that the structure of the discourse helps the listener follow and understand the flow of thought. The longer the turn, the more planning is needed. (Brown et al. 1984, 13-16.) The training of short turns, which the fashionable communicative approach has stressed, does not automatically lead to the control of long turns. To be able to produce long transactional turns, learners need "adequate models, adequate practice and feedback" (Brown & Yule 1983b, 19-24). For an FL tester it is useful to remember that speakers have difficulties with long turns also in the native language. Even young learners can be tested on short terms, while commanding long turns needs both practice and certain cognitive maturity. To advanced learners, such as L2 learners at the end of secondary school, extensive transactional test tasks are, however, the type of challenge that makes them present their best ability.

3.5 Conversation

The commonest genre of speaking is conversation. For the assessment of speech the tester needs to know the basic structure and elements of conversation. In this section the following aspects of conversation will be dealt with: the definition, the functions and general characteristics with special regard to the features important in assessment, elements that cause special difficulties for Finns, and the possibilities to arrange and assess conversation in a speaking test.

Conversation seems difficult to define, and the term is used rather broadly and vaguely. Nolasco and Arthur (1987, 5-7), however, define conversation as a time when two or more people have the right to talk or listen without having to follow a fixed schedule, such as an agenda. As guiding principles in English native-speaker conversation Nolasco and Arthur enumerate the following five: usually only one person speaks at a time; the speakers change; the length of any contribution varies; there are techniques for allowing the other party or parties to speak; neither the content nor the amount of what is said is specified in advance.

Most other researchers are content just to mention different characteristics. It seems to be generally accepted that conversation is informal and takes place between two or more but not very many participants. A characterization like that is, of course, imprecise. It is not only common everyday chat that can be regarded as conversation but also the polished dinner table conversation between two heads of state. In this

respect the boundary between conversation and other discourse types is a fuzzy one and can, according to Cook (1989, 51), best be described as a cline[4]:

Formal spoken discourse----------Conversation

A foreign language student's needs in conversation are very similar to those of the native speaker. A language learner wants to learn to converse in a foreign language, because he wishes to give and receive information, collaborate in doing something, and share personal experiences and opinions with a view to building social relationships. (Nolasco & Arthur 1987, 5; Richards 1980, 1983).

Conversation skills are not equal to speaking skills (Jakobovits & Gordon 1979). It is in conversation that the learner has most opportunities to realize his pragmatic knowledge. Being an accurate and even fluent speaker does not always guarantee the appropriacy of the utterance in a given set of circumstances. There are skills specific to conversation that make it easier for people to talk to one another informally. These skills do not overlap to a hundred per cent with the skills needed in fluent speaking. It is one thing to speak relatively correct and fluent English and another to be able to engage in on-going, interactive, mentally satisfying conversation. Speaking skills are necessary for conversation, but they do not form a sufficient condition. (Maley 1987.)

A central difference between speaking skills and conversation skills is that the former can often be realized on sentence level whereas a central issue in conversation is always the ability to link one sentence to the next. A sentence or a clause alone is never a conversation. Even the case of the hearer not providing the (expected) answer or second part has its significance. For the discourse to be coherent both the speaker and the listener must be constantly aware of what has happened before and what is expected to follow.

Not only formally distinct speech events but all kinds of casual talk are rule-governed. Conversation takes place in a certain order. Kallmeyer and Schütze speak of the *organization of conversation (Gesprächsorganisation)* (1976, 6) and Kallmeyer uses the term *order of interaction (Geordnetheit der Interaktion)* (1988, 1097). The conversation must unroll appropriately in time with a beginning, a middle, and an end. It usually begins with greetings and then proceeds through various ordered moves: the speaker's and hearer's roles are ascertained, topics are introduced, rights to talk are assumed, new topics are raised, and, at an appropriate time, the conversation is ended in a suitable manner. Among the rules that the speaker must master are for example the knowledge of when it is proper to open a conversation and how, what topics are becoming to particular situations, which forms of address are to be used, and how such speech acts as greetings, compliments, apologies, invitations and complaints are to be given, interpreted and responded to. (Richards 1983, 118; Wolfson 1983b, 61.)

---

[4] It is to be noticed that Cook does not even give the name of conversation to the left end of the continuum. In this connection one has to remember that the English word *conversation* - unlike the German *Konversation*, the Swedish *konversation* and the Finnish *keskustelu* - refers to informal talk only, whereas the notion of a similar but more formal speech event may be covered by, for instance, the word *discussion*.)

There is no doubt that conversation is the most central and primordial genre of speaking (Gardner 1984, 102; Levelt 1989, 29; Levinson 1983, 43), and should, perhaps, therefore also be the focus of oral testing. Simple as it may seem, it poses, however, problems for the tester. It is true that an interview, a commonly used test format, can be regarded as a form of conversation, but the unequal distribution of power keeps it formal and the initiative one-sided. Another kind of disparity may haunt peer conversations in a test situation, and they, too, are often far from natural. In the following sections two conversational rules, those of turn taking and phatic communion, will be described and dealt with also from the testing viewpoint. Both of them have been claimed to cause special difficulties to Finnish learners of English (Tiittula 1992; Yli-Renko 1989a).

3.5.1 Turn taking

Turn taking is a conversational skill that is important to learn also in a foreign language context. A *turn* (cf. 3.4) in a conversation is the interval between two successive speaker switches, and the place when the speaker and listener change is referred to as *turn taking* or *turn switching* (Coulthard 1985, 59-69; Crown & Feldstein 1985, 32; Nolasco & Arthur 1987, 5-6). The tempo of turn switching varies in different cultures depending on the general attitude to speech and silence. In a culture of abundant speech and little silence turn switches are short, while in a more silent culture they are long.

It is customary for a learner to tend to transfer the turn taking patterns of his native language to the new surroundings. The different conventions may, however, easily become another source of cultural misundestanding. If a learner from a talkative society carries his swift turn taking practices into a slower culture, he may be regarded as arrogant and bossy. If, on the other hand, the learner comes from a more silent culture, it may be difficult for him to even get a turn or to keep it. Even if he succeeds in getting a turn, he may still be regarded as reserved, unsure, or even hostile. (Scollon & Scollon 1983; Tannen 1985.)

The speakers of Finnish are accustomed to plenty of silence. In the same book in which Tannen described American conversation and gave the article the name *Silence: anything but* Sajavaara and Lehtonen's article about Finnish speaking conventions was called *The Silent Finn*. As it is not only the Americans but many other nationalities who are faster speakers than the Finns, the latter have a hard time competing for the turns. The subskills that they should learn to master are, for example, recognizing the right moment for a shift, signaling one's desire to speak, and keeping one's turn by, for instance, filling the pauses with appropriate hesitation markers (Bygate 1987, 39-40; Levelt 1989, 34; Wardhaugh 1985, 148-55). These new techniques are hardly learned without deliberate practice.

But how can such skills be assessed in a test? Turn switching is a good example of a speaking skill that is important to master but almost impossible to bring out in a test. In addition to the difficulty of making observations of both the speech and the turn switching behavior there is the problem of how to create the right circumstances for a swift, more or less authentic conversation which would challenge the turn taking skills. This is one of the skills in which the tester must admit that not everything that can be taught can or need be tested.

3.5.2 Small talk

Another pragmatic area that is problematic for the tester is the use of *phatic communion* or *small talk*. The term is used to refer to communication between people that is not intended to seek or transmit information but has the social function of establishing or maintaining social contact (Richards et al. 1985, 214). The use of small talk has strict constraints and is to a great extent conveyed by formulaic expressions and gambits such as *How are you?* and *Nice day, isn't it?*

Phatic communion or small talk typically occurs at the beginning of conversation, sometimes with no other type of conversation to follow. Laver (1981, 301-2) distinguishes between three types of subjects. To begin a conversation with a stranger it is safest to resort to subjects that are common to both parties, such as the weather, the beauty of gardens, or the irritation of having to wait in a queue. The syntactic structures are typically abbreviated, thus making the phatic function recognizable. The second type of subjects are factors personal to the speaker, for instance:

This hill was not made for my legs.
I thought I was late for the bus, and now I've been waiting for twenty minutes.

Laver calls this self-oriented speech, and contrasts it with the third type, other-oriented speech, which has to do with factors specific to the listener, for example:

How's life/business/the family?
Do you come here often?

The first type, the neutral subject, is a safe choice with all kinds of partners. With good acquaintances, all three are possible, but with strangers other-oriented talk is possible if the status of the listener is inferior to that of the speaker, and similarly, self-oriented talk can be used if one is speaking to someone in a socially more dominant position.

A characteristic feature of small talk is the fact that plenty of words may be used with hardly anything being actually said. The stereotyped phrases that are an essential ingredient of small talk are perceived to be hackneyed expressions that have lost their meaning. According to information theory, frequency of occurrence and meaningfulness are inversely related. With little of significance being said, no outcomes are expected. The listener is typically listening to just the core of the message, and also typically, there is a great deal of agreement on whatever is said. (Coulmas 1981, 4; Gardner 1984, 107; Wardhaugh 1985, 47.) The Hungarian humorist George Mikes's proposal for an ever-usable conversation about the weather may serve as an example:

Nasty day, isn't it?
Isn't it dreadful?
The rain...I hate rain...
I don't like it at all...Do you?
Fancy such a day in July. Rain in the morning, then a bit of sunshine, and then rain, rain, rain, all day long.

I remember exactly the same July day in 1936.
Yes, I remember too.
Or was it in 1928?
Yes, it was.
Or in 1939?
Yes, that's right. (Mikes 1977, 20-22.)

The ability to chat is, according to Brown et al. (1984, 6-9), the very basis of social life. It is smooth, effortless talk with unimportant content and many topic shifts that keeps the wheels of social life oiled. However, not all native speakers are equally good at it. There are people who find it easy to unearth common topics of interest and to exchange amicable conversational turns with almost anybody anywhere, whereas there are others who, as the dictionary puts it, 'have no small talk' (Cowie, 1989). In a foreign language small talk is on the one hand easy with all the common formulaic expressions, on the other hand twice as difficult if the speaker comes from a culture of negative face where it is polite to leave other people in peace.

In the testing of small talk there are actually two levels. At one level the tester's concern is with the culturally correct use of the pragmatic rules: is small talk used appropriately according to the constraints in each situation? Even this level is difficult enough for the non-native tester, for it presupposes a good knowledge of pragmatic rules. However, from the behavioral point of view the second level is more crucial: does the speaker produce any small talk at all when there is an occasion for it? A sentence unsaid at a moment when it is expected is a much more serious fault than the sentence expressed though containing one or two mistakes of syntax. How could the tester create a social context which would presuppose an initiative to begin small talk? Perhaps the only - and not necessarily very satisfactory - solution is role-play.

4 PRONUNCIATION AND NONVERBAL COMMUNICATION

The necessity of testing FL speaking proficiency with an oral - not a written - test is in no domain of proficiency as obvious as in pronunciation and nonverbal communication. If a Finn had been taught written English but no pronunciation, the way he would decode English written text in speaking would be completely unintelligible to anyone except another Finn. In the same way, wrong prosody can be a source of complete misunderstanding or a pragmalinguistically fatal misconception. In the present visual era, also the nonverbal aspect is an important factor in communication. Would a candidate in a speaking test get the same mark if her speech sample were recorded by video as he gets when recorded on an audio-tape?

4.1 Nonverbal communication

One of the aspects that distinguishes oral discourse from written is the conspicuous presence of nonverbal communication. The importance of nonverbal communication is expressed by the common phrase "It is not what you say, it is the way you say it". When for instance the verbal code and the visual code come into conflict, people are bound to trust the visual one (Burgoon & Ruffner 1978, 140). What makes errors in the nonverbal and prosodic domain problematic is the fact that they are often reacted to subconsciously. When a native speaker hears a non-native mistake of grammar or vocabulary, he can easily attribute it to the foreigner's lack of competence. When, however, there is a wrong tone of intonation or an inappropriate gesture, the native speaker seldom thinks of a problem of communication but may make an unfavorable judgement of the interactant's personality.

The layperson belief goes that nonverbal communication should account for as much as 70 to 90 per cent of the message conveyed. According to Birdwhistell (1970, 158) only 30 to 35 per cent of the social meaning of conversation is carried by words. Yet here as anywhere else where a system as complex as human communication is

discussed it is impossible to assign an exact percentage to any of the components involved. In transmitting attitudes and feelings nonverbal means naturally play a much greater part than they do in conveying, for example, the intricacies of nuclear physics. Yet in all face-to-face interaction, even in erudite discussions of nuclear physics, the conveying of attitudes and feelings plays a crucial part in establishing well-functioning human relationships, which are a necessary prerequisite even for scientific cooperation.

There are many descriptions of the system of human communication. The writer studied those by for example Beattie (1981), Hurley (1992), Kohonen (1987), and Oksaar (1988) and found that each of them had described and categorized the elements somewhat differently. Because it was not possible to include all the elements in the present oral test, it did not seem relevant to present the various systems here but to choose one for a general view. For breadth and clarity the model of A. Ellis and Beattie (1986, figure 5) was selected.

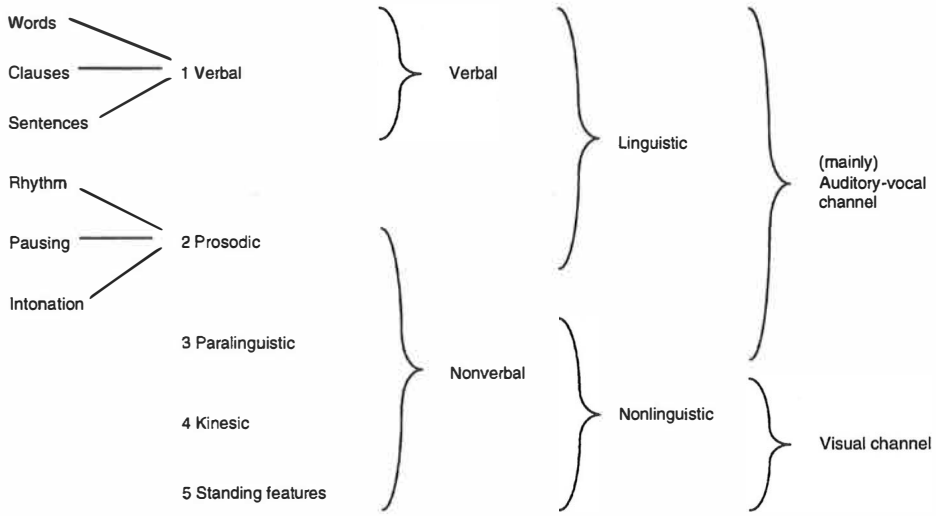

FIGURE 5 Systems of human communication (from Ellis and Beattie, 1986, 18)

In their description of the systems of human communication Ellis and Beattie include prosodic features (rhythm, pausing, and intonation) in the linguistic system, but not in the verbal one, which only comprises words, clauses, and sentences. It is to be noted that Ellis and Beattie present their description as psychologists and not as linguists. This explains the lack of categories such as phonemes and morphemes, which would be part of the verbal system. Pauses are dealt with under two headings. In prosody Ellis and Beattie include the ones whose position and function are linguistically determined, whereas those pauses that seem to have no clear linguistic cause are part of the paralinguistic system. Vocal phenomena, variations of tone, which are less systematic than prosodic features, also belong to the paralinguistic

category (see also Crystal 1991, 220). As examples of the paralinguistic system Ellis and Beattie mention the *ums* and the *ahs*, laughing and crying, whining and yawning.

If nonverbal communication is such a powerful message conveyer, it is only natural to wonder to what extent it should be assessed in evaluating oral communication. The traditionally most common decision has been to assess pronunciation but to leave out the other nonverbal parameters. There are several reasons why the same practice was continued in this study. To begin with, large-scale oral testing is new in Finland, which is why there was a need to keep it simple. Secondly, including non-linguistic features in the evaluation would have involved the use of video, which would also have meant a considerable cost effect. The third factor was the question of testing personality versus language proficiency (cf. Huhta 1994), a matter that still lacks a deeper analysis and discussion.

Apart from the justification of assessing personality, there are also the questions of norms and criteria. Each of the systems of human communication is dependent on various contextual and situational factors, so that all-purpose, clear-cut definitions and criteria for assessment purposes are difficult to give. A further complication is the fact that the different systems are in constant interaction with one another and exert constant influence on one another. As a result of this it is artificial to study and analyze any of them in isolation. The last complication is the fact that communication systems are often culture-specific and feasible to interpret only in the light of the culture-specific norms, which, for their part, may vary even at close distance.

The following discussion about the nonverbal factors of communication is thus confined to pronunciation and prosody. When deciding what aspects to handle, two criteria were chosen as particularly relevant to testing. To start with, one should know which deviant variables native speakers experience as specially irritating or detrimental to communication. For the second, one ought to be aware of which of the harmful features are particularly characteristic of the speech of Finns. We may know, for instance, that unfamiliar voice quality can be a source of irritation in cross-cultural communication, but if it appears that Finnish voice quality should not differ from English, why should we test voice quality? To decide matters like this it would be a great asset for testers to have an extensive contrastive literature available. As to differences between English and Finnish, there are studies on suprasegmentals, but in the other areas literature is still scarce.

4.2 Pronunciation

According to the Ellis and Beattie figure of communication, pronunciation has both verbal and nonverbal elements. It is a term commonly used in language and linguistic textbooks and dictionaries, but not often defined. Some articles make an exception, such as Anderson-Hsieh, Johnson & Koehler (1992) and Pennington and Richards (1986), though even they do not actually define the concept but enumerate the components included. The latter describe pronunciation not only as a part of the system expressing meaning but also as a central part of the interactional dynamics of the communication process. As the major areas of pronunciation they discuss

segmental features, voice-setting features, and prosodic features. By voice-setting (Anderson-Hsieh et al.: voice quality) features they mean the habitual positions of articulation in connected speech, which results in a characteristic voice quality. Because Finnish and English do not differ in basic voice quality, only segmental and prosodic features will be discussed here.

Pronunciation is geographically and socially much less uniform than many other variables of language, such as syntax or lexis. With hundreds of millions of native and non-native speakers of English from all parts and walks of life, the question of norm has often been dealt with (e.g. Abercombie 1956; Baxter 1980; A. Brown 1988; Hiltunen 1992; Kachru 1976, 1992; Newbrook 1986; Strevens 1974, 1981). The world-wide standard has traditionally been Received Pronunciation (RP), the educated speech of South-East England, but as the number of the speakers of this variant has constantly decreased while the position of English has risen all over the world, not only General American but also other widespread national or regional varieties now compete for the status of norm in many parts of the world. In Finland the general secondary school curriculum does not take a stand at this point, but a commonly established practice seems long to have been to accept Received Pronunciation with close variants as well as General American.

Another variable of pronunciation that is of crucial significance for testing is, of course, the standard required, the objective. Even as early as 1956 Abercombie relinquished the then traditional goal of native-like proficiency and stated that perfection was only necessary for intending teachers and intending secret agents. Others could be satisfied with a "comfortably" intelligible pronunciation, and "comfortably" intelligible was to be interpreted as pronunciation that could be understood with little or no conscious effort by the listener (Abercombie 1956, 93).

Van Els and De Bot (1987, 147) present intelligibility and acceptability as desirable aims for pronunciation and make the noteworthy point - actually similar to Abercombie's - that if the two come into conflict, the choice should be based on considerations of the aims of learning a language. For a Mediterranean migrant worker in the Netherlands intelligibility of the Dutch they speak should have priority over acceptability, whereas the reverse would be the case for people in representative or diplomatic professions. In general education, like secondary schools in Finland, we have to educate both intending teachers and intending secret agents, which means that, at least in the major foreign language, intelligibility as well as a certain amount of acceptability have to be the goal. The recent nation-level curriculum for senior secondary schools states as an objective that the learner "can actively participate in a dialogue using natural and fluent pronunciation, stress, rhythm, and intonation (*Framework Curriculum for the Senior Secondary School 1994*, 71). This, in fact, optimistically sets the standard quite high.

What are the factors that bring about the desired qualities such as intelligibility? (For a discussion see e.g. Albrechtsen, Henriksen & Faerch 1980; Eisenstein 1983; Johansson 1978). Lehtonen (1977, 41) points out the problematic nature of the error of pronunciation: How should for instance segmental errors be weighed in comparison to errors in timing of speech or errors in pausing, stressing, and dynamic patterning of speech? We do not know how native speakers react to various kinds of deviations from phonetic habits of their own or which errors have an irritating effect and which are only perceived as a feature of the foreign accent. The degree of "foreignness" may

also be a result of the interaction of phonic and syntactic factors, with the predominance of syntactic factors (Beatens Beardsmore 1979; Van Els & De Bot 1987). Not even native speakers can tell, for the native speaker judgement of foreignness is influenced by elicitation techniques, personality factors, and the judge's own linguistic background, particularly the number of languages he speaks (Thompson 1991; Beatens Beardsmore 1979).

Though the part of the different parameters of pronunciation is not clear, there is no doubt about its significance in testing. Phonetic factors vitally contribute to the intelligibility, irritability, and acceptability of speech (Anderson-Hsieh et al. 1992; Ellis & Beattie 1986; Pennington & Richards 1986; Thompson 1991), and they are always there. Mistakes of vocabulary and grammar can to a certain extent be avoided by skillful communication strategies, but pronunciation is present even in reading aloud, and it is rather straightforward to test.

4.2.1 Segmental features

In his proposal for the matriculation examination Hirvonen (1973b) tested the correctness of stress, intonation, and sounds, but since then the emphasis in language learning has changed towards a more holistic view. Besides, with the ever-expanding power of electronic media and entertainment industry, there has obviously been an essential improvement in the details of pronunciation. In today's situation, individual segments could be tested in achievement tests and at the end of shorter language courses, but after ten years of major language studies a more integrated test might do the students more justice.

However, though individual segments are not analyzed in the pronunciation criteria of the present test, the tester must at least subconsciously pay attention to them as well. When choosing what is worth special notice the tester should be aware of two guidelines: the differences between English and Finnish, and the criterion of the functional load.

Differences between English and Finnish. According to Wiik (1965, 15), difficulties in pronunciation occur if the sound systems of two languages are different. Unlike many other fields of linguistics, Finnish and English phonetics have been subjected to comprehensive comparative research (e.g. Lehtonen & Koponen 1977; Lehtonen, Sajavaara & May 1977; Moisio & Valento 1976; Suomi 1976; Wiik 1965). As the result of these studies we possess rich information on the special problems of Finnish learners of English. Both in vowels and in consonants as well as in linking words together Finns have difficulties which are to a great extent explainable by the transfer of mother tongue phonological characteristics.

Problematic sounds are those which are pronounced close to one another but in which two languages make the distinction in a different way. In Finnish a common means for distinguishing two *vowels* is the quantity whereas an equivalent means for English is making use of a tense-lax opposition, which is unfamiliar to a Finn. Accordingly, when trying to pronounce the English sounds /iː, ɪ/, a Finn has a tendency to replace the English qualitative difference by a Finnish difference of vowel length /iː, i/ and also to pronounce the English lax vowel as too tense. A similar

tense-lax variation occurs in the English back vowel pair /u:,ʊ/, which is equally difficult for a Finn.

Other new sounds to Finns are the central vowels /ɜ:/ and /ə/. The Finnish pronunciation of the former is seldom misunderstood, but can be a source of irritating foreign accent (Lehtonen et al. 1977, 116). The latter may be comparatively easy to produce but is not always used where it should be. In Finnish a syllable always receives the same stress, but in English the stress and the vowel involved vary according to the context. A vowel that is marked in a stressed position becomes the neutral /ə/ in an unstressed one.

A syllable in English is often unstressed because in an isolating language a single syllable carries less information than in an agglutinative language. The same is true about individual words. In Finnish, where a word carries a great deal of information, the word boundary is clearly marked. There is a glottalization in front of a word that begins with a vowel, and a Finn tends to use the same device in front of an English word, in which initial glottalization normally occurs only if there is a need to specially emphasize that syllable. The Finnish use of glottalization makes the speech sound disfluent and hesitant to English ears (Koponen 1990; Lehtonen & Koponen 1977; Lehtonen et al. 1977, 154).

The same tendency of a smooth transition between words is also seen in the English use of *consonants*. An example of such smooth linking is the introduction of an /r/ sound at the end of a word ending with an r in spelling if the following word begins with a vowel. A similar trend to smooth transition is observable in consonant clusters that occur at word boundaries or compound words. Pronunciation would be more simple for foreigners if they did not try to be too conscientious but allowed themselves all the assimilations and elisions that take place at word boundaries in native speaker speech. Instead they pronounce every consonant as carefully as they would if the words were isolated, which easily gives them away as foreigners.

It is not only at word boundaries that consonant clusters present difficulties for Finns. Another cause for the problem is the fact that the clusters include sounds that are difficult for Finns even in isolation. Among such consonant groups Lehtonen et al. (1977, 155) mention the following: /pθ, tθ, mθ, nθ, ŋθ, lθ, fθ, θs, ðd, ʃt, tʃt, dʒd, ntʃ, ndʒ, ltʃ, ldʒ/. As there are 24 consonants in English and only 14 in Finnish, the Finnish learner is bound to meet sounds like /θ, ð, ʒ/, which are altogether new to her. However, in consonants as well as vowels the greatest strain is often produced by sounds that seem very similar to the Finnish ones.

In the same way as the tense-lax opposition is problematic for vowel production, the fortis-lenis distinction causes problems for consonants (Lehtonen et al. 1977, 128). In a position where Finns produce only one sound such as /p/, the English have two consonants /p/ and /b/, which are distinguished by not only (1) the fortis-lenis opposition but also by (2) the presence or absence of voice and (3) the consequent lengthening of the preceding vowel and (4) in many cases also by initial aspiration. There are eight such pairs /p,b; k,g; t,d; f,v; θ,ð; s,z; ʃ,ʒ; tʃ,dʒ/. A common mistake for a Finn is to use only one of the distinctive features like the voice-voiceless opposition, in which case the native listener may easily misinterpret the intended /p/ as a /b/ so that a *pin* becomes a *bin*. A similar case in fricatives is the use of the Finnish semivowel /v/ to replace the English consonants /v/ or /w/.

Functional load. If the goal of the teaching of pronunciation is intelligibility, segmental errors matter only so far as they might be a source of misinterpretation. Should a learner have difficulties in the /u:/ - /ʊ/ distinction, it would not be a major concern because he would only have to make this distinction in four minimal pairs (*pool/pull, fool/full, who'd/hood, suit/soot,* (if the former is pronounced /su:t/) (A. Brown 1988, 218). A tool for weighing to what extent individual segments have

Vowels		Consonants	
10	/e, æ/	10	/p, b/
	/æ, ʌ/		/p, f/
	/æ, ɒ/		/m, n/
	/ʌ, ɒ/		/n, l/
	/ɔ:, əʊ/		/l, r/
9	/e, ɪ/	9	/f, h/
	/e, eɪ/		/t, d/
	/ɑ:, aɪ/		/k, g/
	/ɜ:, əʊ/		
		8	/w, v/
8	/i:, ɪ/		/s, z/
7	-	7	/b, v/
			/f, v/
6	/ɔ:, ɜ:/		/ð, z/
	/ɒ, əʊ/		/s, ʃ/
5	/ɑ:, ʌ/	6	/v, ð/
	/ɔ:, ɒ/		/s, ʒ/
	/ɜ:, ʌ/		
		5	/θ, ð/
4	/e, eə/		/θ, s/
	/æ, ɑ:/		/ð, d/
	/ɑ:, ɒ/		/z, dʒ/
	/ɔ:, ʊ/		/n, ŋ/
	/ɜ:, e/		
		4	/θ, t/
3	/i:, ɪə/		
	/ɑ:, aʊ/	3	/tʃ, dʒ/
	/u:, ʊ/		
		2	/tʃ, ʃ/
2	/ɪə, eə/		/ʃ, ʒ/
			/j, ʒ/
1	/ɔ:, ɔɪ/		
	/u:, ʊə/	1	/f, θ/
			/dʒ, j/

FIGURE 6 Rank ordering of RP phoneme pairs commonly conflated by learners (from A. Brown 1988, 222)

meaning distinguishing significance is the *functional load.* King (1967, 831) defines it in the following way: ".it is a measure of the work which two phonemes (or a distinctive feature) do in keeping utterances apart - in other words, a gauge of the frequency with which two phonemes contrast in all possible environments".

In making quantitative judgments on the functional load of various segments, for example the following criteria are used: cumulative frequency, probability of occurrence, occurrence and stigmatization in native accents, acoustic similarity, structural distribution of phonemes, lexical sets, number of minimal pairs, phonetic similarity. On the basis of the different criteria A. Brown (1988) has compiled a rank ordering of the proportional significance of several Received Pronunciation minimal pairs of vowels and consonants. The order is presented on a 10-point scale, where 10 represents maximal importance and 1 minimal importance. The scale is shown in Figure 6.

The conclusion we can draw from the information in this section is that the most important minimal consonant pair to teach and test in Finnish schools would be /p,b/ because it is number 1 on the ranking scale and also particularly laborious to Finnish learners. Similarly, the most useful minimal vowel pair would be the /e/, /ɪ/.

4.2.2 Prosodic features

The term 'functional load' is generally used about segmentals and the proportional relevance of individual segmentals, but if the term were used about the effects of the different subsystems of phonology, prosodic features would carry much weight. It is true, however, that studies comparing segmental and suprasegmental features have so far been scarce. Most studies in error gravity, i.e. native speaker reactions to non-native speech, have concentrated on syntactic and morphological factors (for reviews of the literature see Eisenstein 1983; Ludwig 1982; Ryan 1983), and some on pronunciation as a whole as weighed against other aspects (e.g. Politzer 1978; Varonis & Gass 1982). However, James (1976) and Johansson (1978) tried to compare the relative effects of articulation and prosody on native speakers' judgments of pronunciation, the former, however, in L2 learner French. Both studies indicated that good prosody combined with poor articulation was more acceptable than poor prosody together with good articulation. However, as no statistical studies of significance were reported, the results of the two studies cannot be regarded as conclusive. The most important study so far seems to be that by Anderson-Hsieh et al. (1992), which shows that of the three components of pronunciation measured, namely the segmental error rate, the prosodic score, and the syllable structure error rate, each correlated with the total pronunciation rating, but the correlation of prosody was always the strongest.

Anderson-Hsieh et al. (1992, 549) point out the difficulty of dividing the prosody into smaller segments, such as syllable duration, pitch range and direction, and measuring their relative significance, but they admit that it would be technically quite feasible and suggested that in future such measures should be taken. If the experiments showed that there would be a substantial correlation between all these elements, it might be reasonable to test only one of them. So far, as we have no such information, we must be satisfied with assessing the total and paying special attention

to those prosodic factors where Finnish learners seem to show most deviance: intonation and rhythm (Lehtonen et al. 1977).

4.2.2.1 Intonation

In the previous section it was suggested that to decide what aspects of language are worth testing the criteria should be cross-cultural differences, on the one hand, and significance for the transmitting of meaning, on the other. Intonation meets these two demands excellently. In Finnish the use and range of intonation are not as varied as in English (Karlsson 1983, 175), which makes English intonation difficult for Finns. According to Lehtonen (1978a, 61), it is likely that speakers of the Nordic languages also use the pitch level of their voice in a way different from that of the speakers of English. By using filtering techniques Van Els and De Bot (1987) showed that mother tongue intonational habits like this transfer into L2.

It is intonation that gives an utterance its decisive meaning (Callamand 1987). Intonation modifies and sometimes contradicts the literal content of an utterance. In cases where intonation and the other elements come into conflict, the interlocutors base their inference of meaning on intonation (Hurley 1992; Kreckel 1981; Lyons 1977, 63; Raith 1984). It is the prosodic factors that cause cross-cultural conflicts that are much more pervasive and fundamental than those associated with sentence level grammatical and lexical distinction (Gumperz 1982, 129). The Indian and Pakistani women working in the staff canteen at Heathrow Airport were hardly aware that offering more gravy with a falling intonation - *Gravy?* - would be considered a "cool, calm, phlegmatic, detached, reserved, dispassionate, dull, possibly grim, or surly attitude on the part of the speaker" (O'Connor & Arnold 1959, see Coulthard 1985, 98). Only after a new tone had been learnt were the customers satisfied.

The importance of intonation for meaning is easy to understand if we think of all the different tasks that it has. The commonest everyday function that it performs is to divide information into manageable units. When a speaker wishes to say something, he cannot say everything at once, but has to organize his message into chunks. Using a term attributed to Halliday, chunks into which propositions are organized are in discourse analysis called *information units*. With the help of intonation, information units are in speech realized as *tone units*, in which the key item of information is marked by pitch prominence. By the placement of pitch prominence the speaker can mark information as new versus given. The way he chooses to divide his message into tone units has great semantic significance. At school, however, the division into message units is problematic only at a more advanced level; utterances produced during the first few years are usually so simple that clause and message units coincide. For the advanced learner the problem of the right division is particularly manifest in reading especially when passages of complex text have to be read aloud (Gutknecht 1978, 261).

The importance of intonation is prominent in conversation, in which it has several functions to perform. It is the perhaps most significant cue to the judgement of turn beginning and turn end in interactional exchange (Brown & Yule 1983a; Callamand 1987; Local 1985; Schaffer 1983). In topic management and topic change intonation is also a significant signal, though, according to Schaffer, not as powerful as the total context (Brown & Yule 1983a; Schaffer 1984). It can also be systematically

used as a means to constitute and control participant cooperation in repair and problem handling sequences (Selting 1988).

When one considers the importance that intonation has for the creation of meaning, one could think that it is the most widely assessed aspect of language. And it is true that correct intonation is often mentioned in the criteria of pronunciation, and it is sometimes, but rarely, tested by itself. Hirvonen (1973b) did test intonation in his suggestion for a secondary school final test, and so did Lehtovaara (1978) in a test for primary school, but in today's communicative orientation such testing is rare. The reasons, although seldom mentioned in literature, appear rather obvious. The main reason is to be found in the complexity of the phenomenon itself. It is only natural that from this complexity results a secondary reason: the insufficient capacity of the tester.

Intonation is capable of expressing a most broad spectrum of nuances, but that is possible only because it is a very complex medium. The existing categorizations and descriptions of intonation are very complicated with long passages of elaborate text to depict a minor stretch of utterance (Currie & Yule 1982). What makes intonation so arduous to describe is its dynamism and lack of permanence. There is no constant relationship between particular acoustic phenomena and particular analytic categories, so that the interpretation does not depend on absolute values but on contrast with the previous values (Coulthard 1985, 97).

A further complication of the testing of intonation is the fact that, in spite of the various attempts at systematization, no standard has emerged. The American and the English school of phonetics go their separate ways. Also regional norms clash. A native speaker from England will as often as not misunderstand the intonation of a native speaker from India (Gumperz 1982, 118-29). Intonation can also indicate a person's socio-cultural membership and/or his personal style (Callamand 1987). In addition, former established theories change. One of the most commonly taught rules used to be the one about the wh-questions having a falling tone. Now this is partly disputed (Gutknecht 1978, 266).

To assess intonation the tester needs a trained ear, almost the knowledge of a phonetician, and certainly a good command of intonation himself. It is often difficult for even a phonetician to distinguish two tones from one another (Currie & Yule 1982, 272). Regarding the very complex nature of intonation it is not surprising that a nonnative teacher, even with many years at the university, does not have all the necessary qualifications to either teach or test the necessary subtleties.

However, even with perfect teachers and next to perfect learners, the intricacy of testing intonation remains. With a few exceptions, it is difficult to separately assess a phenomenon whose effect is always dependent on the context and the information that lexical, syntactic and the other prosodic elements bring to the utterance. The multifarious nuances of meaning brought about by intonation are not easy to describe in an unambiguous way that would guarantee an equal interpretation. And whose meaning is it anyway? How do we know that what the testee is saying is not what he means albeit it is impolite? Besides, how many cases are there in which another tone would be unthinkable?

Imitation, reading aloud a dialogue, and a role-play with quite explicit descriptions of the tones of individual lines might be used as a more mechanic means of testing intonation. For more integrated tests, the remaining solution so far has been

to include intonation in the pronunciation criteria and to test it rather holistically. This decision was taken also in this experiment.

4.2.2.2 Rhythm

Rhythm is defined as the perceived regularity of prominent units in speech (Crystal 1991, 302), which is brought about by the variation of stress (stressed v unstressed syllables) and/or length (long v short syllables) and/or pitch (high v low pitch). From an English listener's point of view rhythm is the organizer that divides the flow of speech into meaningful information-bearing units (Allen 1975, 84). In natural speech, rhythmic units are equal to breath groups and coincide with sense groups, whereas the uneven and jerky rhythm of foreign learners is often caused by faulty division into sense and breath groups (Taylor 1981, 237). At the same time breath groups are usually too short.

The listener engages in an act of communication with certain expectations concerning also the rhythm, and if these expectations are not met, i.e. if the non-native speaker's syllables are "of strange and unpredictable length", the native speaker will have much greater difficulty in understanding what is said (Faber 1986, 245). Listening to such talk is also tiresome. Correct rhythm is thus an essential ingredient of intelligibility (Taylor 1981, 242). The prosodic features rhythm, stress, intonation, and pitch function in constant interaction and interdependence with one another and also the sounds with the result that their respective role in communication is not always easy to distinguish.

Depending on the way in which rhythm is brought about languages are divided into two groups, syllable-timed and stress-timed (Taylor 1981, 235). In the syllable-timed group, to which also Finnish belongs, rhythm is created by an equal interval of time between each syllable, be it stressed or unstressed. In the stress-timed group, on the other hand, stressed syllables occur at equal intervals. English belongs to the latter group. In Finnish, stress is a word boundary signal, while in English it is used to indicate maximal information.

As Finnish and English belong to different groups of rhythm formation, it is only natural that Finnish learners of English should have difficulties with rhythm (Hackman 1978; Lehtonen et al. 1977, 63). However, also speakers of other stress-timed languages may have problems with the rhythm of English, which is perhaps, all in all, "the most widely encountered difficulty among foreign learners" (Taylor 1981, 235). A common cause for the dilemma is syllable duration. In English the stressed syllable is longer than the unstressed one. In an experiment conducted by Klatt (1975, 133) the average duration of stressed vowels was 132 msecs and that of unstressed ones 70 msecs. In another experiment (Adams & Munro 1978, 142), comparing native speaker English to non-native speaker English, it was discovered that in non-native speech stressed vowels were only 11 per cent longer than unstressed vowels, whereas in native speech they were 30 per cent longer. As there was only little intergroup variance in stressed syllables, the difference was due to the mispronunciation of unstressed syllables.

Rhythm is a more simple phenomenon than intonation, and, accordingly, also easier to test. As the difficulty of correct rhythm increases when language studies advance and sentence structure becomes more complex, the end of secondary

education would seem to be one of the appropriate stages to test it. According to Gutknecht (1978, 261), advanced students' problems of division into message units are very conspicuous in oral reading, so reading aloud could be a relevant form of testing. In the present study the available resources did not allow the separate testing of rhythm, but it was incorporated in the testing of pronunciation. In Subtest 1, Reading Aloud, some individual words were included for the purpose of testing word-stress. The different pronunciation criteria for Subtest 1 and the other subtests were also designed with testing stress and rhythm in mind.

5 FLUENCY

"Speaking is not a knowledge thing, it's a fluency thing", Paul Meara, in a lecture in the Summer School of Linguistics, University of Jyväskylä, June 10, 1993.

Language teaching and learning would be much more simple if speaking and writing were identical or explainable by some common factor g, as was supposed in the 1970s. Since this hypothesis was rejected, testers have been trying to identify the factor or factors which would account for oral proficiency as distinct from literary skills. In this endeavor, the concept of fluency - though used about the other skills, too (cf. e.g. Brumfit 1984, 54; N. F. Davies 1982; Lehtonen & Sajavaara 1985) - has had a central role. Meara's words above were probably meant to be taken as an ad hoc comment rather than a weighty definition of the nature of speaking. Of course speaking is also "a knowledge thing", but the statement as a whole seems to capture a common intuition about the quintessence of speech. And the intuition is verified by research. According to Feyereisen, Pillon, and de Partz (1991, 4), many features of speaking are loaded on fluency. Consequently, it is no wonder that in the teaching and learning of oral skills fluency has played an important role both in goal setting, materials selection and results evaluation (Lehtonen 1981, 322-3).

When evaluating oral proficiency there are at least two reasons why it would be important to find a universal and operational definition of fluency. Firstly, fluency is - together with vocabulary, grammar and pronunciation - one of the most commonly used criteria or bands in oral tests. For the second, the non-native speaker's fluency affects the native speakers' willingness to seek interaction with him. This was shown by Albrechtsen, Henriksen, and Faerch (1980), who studied the effects of Dutch students' English interlanguage on native speakers of British English. They found that native speakers reacted the most negatively towards speakers whose language showed extensive use of hesitation phenomena and - somewhat surprisingly- communication strategies. To have to try to infer the meaning suggested by some communication strategy or to have to restart the decoding of a message all over again was concluded to involve a cognitive strain. It would be only natural that - as e.g. Fillmore (1979) has suggested - the ease of understanding would encourage the native speakers

to renew contact with fluent learners, whereas the difficulty of comprehending would be disruptive to interpersonal communication and might discourage the interlocutor from seeking further contact (N. F. Davies 1982, 4; Olynyk, d'Anglejan & Sankoff 1990, 153).

One more reason for the need to look closely into the concept of fluency is the fact that fluency does not affect only the assessment of language proficiency but also the interlocutor's opinion of the partner's personality. Research results confirm (e.g. Koponen 1990, 181-2; Olynyk et al. 1983, 230) that listeners' evaluation of the speaker's personality, attitude and intellectual capacity is influenced by the speaker's oral fluency. Olynyk et al. made ten Canadian military cadets assess their peers, who spoke both French and English. The study revealed that the same people were judged as more intelligent and more acceptable, depending on whether they spoke the language in which they were more fluent or the other one.

In spite of the frequency and the importance of the term, fluency is a vague concept. Just as we have abundant literature about mental processes and the production of speech but very little actual knowledge, we still, in spite of copious research, lack sufficient understanding of the nature of fluency. The difference which makes one of two L2 speakers with equal communicative competence to be judged fluent and the other nonfluent has long occupied researchers' minds (cf. e.g. Olynyk et al. 1983, 213; Rehbein 1987, 100; Sajavaara 1987, 62; Sajavaara & Lehtonen 1980, 71). In compiling his band descriptions each test writer has, in a way, to give fluency his own definition. To increase the validity of testing, these descriptions should be based on common theory.

5.1 The concept of fluency

Though fluency has always played a role in EFL teaching, it became a major target of research only in the 1970s, at the time when developing communicative competence and the speaking skill became primary objects of language teaching. In the audiolingual writings of the 1950s and 1960s the concept is hardly mentioned, but in 1975 a whole book, Leeson's *Fluency and language teaching*, was published on the subject, but it did not bring about any major changes of the concept. More advanced ideas were put forward by Lehtonen and Sajavaara (Lehtonen 1978b; Lehtonen, Sajavaara & May 1977; Sajavaara 1977; Sajavaara & Lehtonen 1978). They showed fluency to be a diversified and debatable concept, for which very few norms or standards could be set.

The perhaps most-cited definition of the 1970s was presented by Fillmore in an article in 1979. His view of fluency was broad, covering both quantity and quality. A fluent speaker had to be able to speak at length, use coherent and reasoned sentences, have appropriate things to say in different contexts, and even be creative and imaginative in his language use. However, Fillmore distinguished these abilities as four different varieties of fluency which were not necessarily combined in one person.

In recent years interest in fluency research has accumulated. With growing understanding of human cognition and speech production, knowledge of the concept has deepened, but is still far from sufficient for an operational definition. Among the aspects that have interested researchers have been the following: the relationship of fluency and accuracy (e.g. Brumfit 1984; N.F. Davies 1982), fluency improvement (e.g. Lennon 1990; Varadi 1990), the influence of personality, genre and L1 on fluency, and the subjective and objective ways of assessing it (e.g. Koponen 1990; Korpijaakko-Huuhka & Moore 1992; Lehtonen 1978b, 1979; Lehtonen & Koponen 1977; Lennon 1990; Moore 1990, 1991a; Riggenbach 1991; Sajavaara 1987, 1988; Sajavaara & Lehtonen 1978). For the definition of fluency it has proved necessary to also define its antonym, disfluency or nonfluency. While the hesitation phenomena characteristic of disfluency were at first considered as mere speech errors, more recent research has also stressed the positive qualities inherent in them.

In their attempts to depict fluency researchers have come up with a wide list of epithets, from dictionary definitions to characterizations of their own. In the approximately 40 books and articles on fluency examined for this study, for instance the following synonyms or descriptions of the word *fluent* (usually attributed to some noun) were used:

acceptable, articulate, coherent, complex, continuous, diverse, easy, effortless, eloquent, facile, fast, flowing, garrulous, logical, many-sided, natural, normal, rapid, redundant, relaxed, rich, smooth, varied, voluble, witty.

It appeared that in the various definitions and descriptions some common features could be discovered. Behind the diversity of the different epithets it seemed possible to establish four components: temporal features, phonological features, dynamic/interactional features, and qualitative features. In the following I will discuss each of them.

5.1.1 The temporal aspect: fluency as smooth motion

The root of the word *fluent* is the Latin verb *fluere,* to flow. Accordingly, many descriptions of the word *fluent* have something to do with motion. Any motion takes place at some greater or lesser speed, and speed is connected with time. So it is natural to speak of the temporal element of fluency. For a long time the view was held that speech should always proceed at a more or less regular pace, a "normal" speech rate, an even tempo. The absence of movement, any longer pause or any other deviance was considered to be a sign of malfunction. This "more and faster is better" or "mind working as a machine" view was well in line with the general behaviouristic-structuralistic idea of regularity and order (Moore 1990; Scollon 1985).

It was soon realized, however, that speech rate was not such a straightforward matter. As a matter of fact, even in the structuralistic era, Goldman-Eisler (1968, 99) showed that pausing in prose read aloud was different from that in spontaneous speech. Sajavaara and Lehtonen (1978 and 1980; Lehtonen 1978b, 1979) were among the first to show experimentally that speed was not constant across different discourse genres but varied depending

on the cognitive task. They asked their subjects, Finnish university students of English, Finnish and Swedish-speaking Finnish business college students, students from a Swedish military academy, and some native speakers of English to read aloud simple and complex English texts from an articulation drill book and describe two sets of cartoons in English. In the oral reading task the speech rate was much faster than in the cognitively more demanding narration task. In the reading task there was no significant difference between the groups. Unlike many students, the native speakers made a clear distinction between the two texts by reading the complex text more slowly. In the narration task, on the other hand, there was a significant difference in the speech rate of the different groups. For instance in the Lehtonen 1979 experiment, the articulation rate (i.e. number of words or syllables per minute after subtracting the duration of pauses in the total time) of the Swedes was 76% of the native speaker rate, that of the Swedish-speaking Finns 61% and of the Finns 55%.

A perhaps not surprising finding in the Sajavaara-Lehtonen 1978 experiment was the discovery that some L2 students used a greater articulation rate for the reading task than did the L1 speakers. Too fast a speech rate with foreign accent has been shown to be detrimental to intelligibility even among native listeners (Anderson-Hsieh & Koehler 1988).

As a unit of speech rate Sajavaara and Lehtonen used both syllables and words per minute. The double choice is understandable, because for contrastive purposes neither words nor syllables are self-evident. The length of words varies in different languages, and in an agglutinating language syllables have more weight than they have in an isolating one and take accordingly more time to produce. Recently it has been suggested (Vanderplank 1993) that the syllables/words per minute unit be replaced by the more descriptive measures of 'pacing' and 'spacing', the former of which would indicate the tempo at which stressed words are spoken and the latter the proportion of stressed words to the total. So far the commonest technical measures of fluency have been those used by Lehtonen and Sajavaara: the syllables/words per minute unit, the total speech rate, the articulation rate, and the percentage of pauses.

As was suggested above, different speech rate medians have been found for different types of discourse. A study frequently referred to is that by Tauroza and Allison (1990), in which speech rates in British English were measured. These scholars chose the material from four common forms of authentic spoken discourse: (1) scripted radio monologues: radio news broadcasts and documentaries (2) conversations (3) interviews (4) lectures to audiences consisting mainly of nonnative speakers of English. As units of speech rate they used both words per minute (w.p.m.) and syllables per minute (s.p.m.) and calculated the rate for syllables per word (s.p.w.). They also compared their results with a previous study by Pimsleur, Hancock, and Furey (1977), who had assessed the speech rate of American radio news announcers. As a result of their study Tauroza and Allison recommend syllables per minute to be used as unit of measurement. Table 2, in which their results are presented, shows how the length of words and the rate of speech varies according to discourse type, the latter possibly also according to national discourse culture:

TABLE 2 Mean number of words per minute (w.p.m.), syllables per minute (s.p.m.), and syllables per word (s.p.w.) in the different categories of speech (from Tauroza and Allison 1990, 97)

Category	w.p.m.	s.p.m.	s.p.w.
Radio	160	250	1.6
Conversation	210	260	1.3
Interview	190	250	1.3
Lecture	140	190	1.4
Combined	170	240	1.4
Pimsleur et al.	180	300	1.7

Tauroza and Allison's table is a quantitative confirmation of what others had stated before, namely that there is no one right rate of speech applicable to all situations. In any case, the whole question of the unit of measurement is only relevant when there is an opportunity to use laboratory analyses. In classroom circumstances the teacher has to rely on his own judgement. For that purpose it is, nevertheless, gratifying to know that recent research has indicated human assessments to be consistent with results from laboratory analyses (e.g. Korpijaakko-Huuhka & Moore 1992, 20; Lennon 1990, 412).

5.1.2 The phonological aspect: fluency as pleasant sound

If the predicate of fluency is moving, its subject is sound. Most writers mention some phonological aspect as an important determinant of fluency. According to Starkweather (1987, 12), normally fluent speech has a characteristic rhythm, which disfluent speech lacks. Albrechtsen et al. (1980, 386) and Korpijaakko-Huuhka and Moore (1992, 14) mention intonation as a noteworthy factor. Dalton and Hardcastle (1977, 5) used the term transition smoothness to cover the main speech features involved in the assessment of fluency and found the following variables to be particularly important: pausing, rhythmical pattering, regulation of tempo, intonation and stress patterns (and other features including e.g. interjections and interruptions). Koponen (1990, 167) sees the learner's central task to involve the mastery of the idiomatic pronunciation habits of the target language and considers the Finns to be particularly lacking in the ability to link various phonetic elements, especially words beginning with a vowel, to the previous word.

5.1.3 The qualitative aspect: fluency as ease

When speaking of phonological accuracy or phonological confidence as a prerequisite for fluency, Lennon (1990, 409) shared a common linguistic opinion that temporal factors alone are not sufficient to account for fluency. The demand for holistic skill is old. Even in 1975, when Leeson (1975, 131) claimed that speed should be joined with quality to bring about fluency, he was referring back to the earlier work of Goldman-Eisler (1968). Not only phonological accuracy but also linguistic accuracy as a whole or linguistic acceptability is needed (e.g. Dalton & Hardcastle 1977; Hammerley 1991 51; Lehtonen 1978b, 12; Olynyk et al. 1983, 230; Sajavaara 1977, 23 and 1987, 62; Sajavaara & Lehtonen 1978, 34-5 and 1980, 71). The early writers demanded varied

vocabulary and precise expression (cf. Sajavaara & Lehtonen 1978) as well as stylistic variation (Dalton & Hardcastle 1977). Particularly in the material of writers interested in speech anomalies (e.g. Feyereisen, Pillon & de Partz 1991; Korpijaakko-Huuhka & Moore 1992, 14) the information content and complexity of sentence structure have been shown to be significant parameters.

5.1.4 The interactional aspect: fluency as communicative fit

In the Lehtonen and Sajavaara and Lehtonen studies of the late 1970s it was shown that fluency is a matter of genre. The two genres that these researchers experimented with were oral reading and narration, but a later investigation brought up another big difference, namely that between fluency in a monologue and fluency in a conversation. According to Riggenbach (1991, 439) fluency in a conversation involves the accomplishment of tasks quite different from a monologue. In a conversation the interlocutor has to be able to initiate topic changes and show comprehension not only through backchannelling but also through relevant comments and responses. In addition, the participants in a conversational exchange must see to the appropriate latching and overlapping of turns as well as to the proper amount of speech produced relative to the other interlocutors.

The dependence on the interlocutor and the context may be among the main reasons why some researchers (e.g. Sajavaara & Lehtonen 1978) find fluency to be difficult or even impossible to quantify by technical measures. A conversation is a product of an interactional enterprise where the quality of the product is assessed by the counterpart. The very difficulty but also the fascination of speech lies in its dynamic and unpredictable character. What is the right tempo of speech at this moment with this particular interlocutor may a couple of minutes later be wrong. Lennon (1990, 391) is probably right in claiming that fluency is purely a performance phenomenon with no permanent fluency store. According to modern linguistic theories communication is negotiation of meaning, in which the interlocutors, using all verbal and non-verbal resources, try to reach mutual understanding. To succeed, the participants have to use what Faerch and Kasper have called co-operational strategies (1983, 50-2).

According to this view, what out of context might appear as too slow a speech rate or too long a pause may be necessitated by the needs of the interlocutor. A fluent conversationalist can adapt the register of his language to the expectations of the listener and to the situational context. By at least unconsciously paying heed to the signals sent by the listener, the speaker knows whether the message has been understood and how it has been received. (Cf. e.g. Dalton & Hardcastle 1977,). The speaker has been fluent if the interaction leaves the participants with the feeling of communicative satisfaction, communicative fit (Sajavaara 1987, 62; Sajavaara & Lehtonen 1980, 73). As beauty is in the eye of the beholder, the right sound is in the ear of the listener.

5.2 Pauses/disfluency phenomena

In speech production sound and absence of sound, pauses, alternate. As a pause most researchers have counted a silence equal to or longer than 200 milliseconds. According to Richards and Schmidt (1983) pauses are "a commonly occurring feature of natural speech in which gaps or hesitations appear during the production of utterances". In this study the term pause will be used not only of the absence of sound, silence, but also of other hesitation phenomena, such as filled pauses (e.g. *er*), elongated syllables (e.g. *w-e-ll),* repetitions (e.g. *me-me-me*) and so on. To understand the proper and improper use of pauses in a foreign language it seemed natural to first discuss pauses in the native language.

5.2.1 Pauses in the native language

Pausing can be seen as positive or negative, a well-formedness or an ill-formedness phenomenon, depending on its origin and function (Arevart & Nation 1991; Moore 1990). A central originator of pauses in speech is planning. Producing speech is such a multifarious and complex cognitive phenomenon that it requires planning which takes place on various linguistic levels (morphemic, syntactic, semantic, discourse) concurrently. As speech proceeds in real time, both planning and production have to take place simultaneously. If the momentary load of the short-term memory exceeds its capacity, it leads to a slowdown or a block which might be heard as silence or other hesitation phenomena.

Among the factors that affect the amount and duration of pauses are the personality of the speaker, the interfering agents, such as interruptions and irrelevant sounds, and the complexity of the situation and/or the text to be produced. The three are often intertwined. Among L1 speakers there is as much individual difference in pausing as there is in any other form of human behavior. Every person has his own speech pattern (Fillmore 1979, 96), part of which is his individual way of pausing. There are for instance people whose ideational fluency is so poor that they find it difficult to think of anything to say even in their mother tongue. Similarly, some people are more easily distracted and brought off the main course of what they were trying to say than others. Individual variation in speech rate and pausing is bigger among Finns than among the speakers of many other languages (Lehtonen 1985; Moore 1991a). From a testing point of view it is important to know that a person's L1 pause profile is carried over to his L2 speech (Olynyk et al. 1990, 153). For an assessor of a person's L2 fluency it would, accordingly, be important to know the subject's L1 fluency and pausing profile.

Different discourse types are differently susceptible to outward distracters and inner cognitive constraints, which makes them differently susceptible to pauses as well (Riggenbach 1991, 439). In a speech or lecture -type monologue the speaker can usually concentrate on his message without being constantly interrupted, which is one explanation of less pausing in a monologue than a dialogue. That monologues can be planned in advance is another. But a dialogue or a monologue is not a homogeneous genre. Stylistically formal

speech is less liable to pausing than informal speech (Sajavaara & Lehtonen 1978, 31). Argumentation and presenting theoretical concepts are more difficult than describing something or telling a story based on pictures. The early pausologist, Frieda Goldman-Eisler (1968), showed that unfilled pauses were more frequent if the subjects had to interpret subtle cartoons than if they only had to describe them. The genre with the least cognitive constraints has proved to be oral reading, in which the speaker does not have to plan the contents.

Pauses caused by planning and the eventual cognitive constraints are called hesitation pauses. They are usually contrasted with another kind called *juncture pauses* (originally Loundsbury 1954, see Lennon 1990, 393). As the name suggests, the latter occur at the boundaries of natural linguistic units, such as t-units. Butterworth (1975) showed by experiment that these linguistic pauses were also idea boundaries in which the mind could be described to pull itself together, formulating the new idea. Lennon refers to the same phenomenon by the terms "sense group" and, citing Halliday 1967, "information unit" (1990, 393). In speech such units form prosodic phrases, which Dalton and Hardcastle (1977, 33) called "breath groups", Rivers and Temperley (1978) "meaningful mouthfuls" and Görding and Eriksson (1991, 45) defined as "a part of an utterance which is connected by special rhythmic and tonal pattern and demarcated by discontinuities in the range or general direction of the pitch contour". *Hesitation pauses*, on the other hand, often appear within a juncture and thus interfere with the flow of communication.

The proportion of pauses, hesitations, backtracking, restatements and the like out of total speaking time - between 30 and 50 % (Leeson 1975, 67) - is substantial enough to show that pausing is an integral aspect of native language discourse. Juncture pauses are not only natural but even beneficial in allowing also the listener the well-needed break for processing (Enkvist 1990, 21; Rivers and Temperley 1978, 83). For the speaker hesitation mechanisms are a kind of mental eraser to repent and improve what went wrong. Sometimes pauses may signify the affective state of the speaker, like situational anxiety or dispositional anxiety (Dalton & Hardcastle 1977, 36). It is not uncommon to make a conscious use of pauses for rhetorical purposes, such as allowing a point to sink in, or even to stimulate laughter or applause (Ellis & Beattie 1986, 119). In conversation the commonest function for a pause is to mark an opportunity for turn exchange.

5.2.2 Pauses in the foreign language

It goes without saying that pausing is more frequent in L2 speech than L1 speech and more frequent in elementary learner discourse than in the speech of an advanced student. As an example can be mentioned the pause/clause ratio in the Sajavaara & Lehtonen study (1978, 46), being 2.2 for Finns and 1.3 for native speakers of English (two informants only). Similar results were obtained by Olynyk, Sankoff, and d'Anglejan (1983, 1990), who showed not only that there appears more hesitation in L2 unplanned speech than L2 planned speech but also that there is more hesitation in L2 planned speech than L1 unplanned speech.

Processing in a foreign language takes longer time because the learner is not yet in possession of a sufficient association network, which would make the fast retrieving and combining of elements possible (Lehtonen 1990, 38). Another prerequisite for the development of fluency in formal L2 learning - like learning at school - is that declarative knowledge should become procedural knowledge (cf. e.g. Faerch & Kasper 1987). To what extent such development is possible at school - or anywhere - is a matter of discussion (for an account of recent thinking see Jaakkola 1997), but the rate of the process is dependent on the respective teaching approach. It is a well-known fact that people taught by the traditional grammar-translation method often have an extensive L2 declarative knowledge, but have difficulties in transforming that knowledge into smooth speech. Learners taught by traditional methods, which pay ample attention to errors, are more likely to monitor their speech than learners who are more geared toward the contents of the message. (See e.g. Lehtonen 1990, 43; Sajavaara 1987, 54.)

The main cause of the greater proportion of pauses in L2 than L1 is the lack of automation. With missing automation, micro-level planning, the retrieving and combining of individual elements, is slow. In the grammar-translation method the learners had too little practice in actual language use, simulating real life situations, which is a central means of developing automation. But whatever the method, achieving fluency at the elementary level of L2 learning, with all the new elements, is always problematic. A great asset in acquiring automation at this stage is the knowledge of *formulaic expressions* (cf. section 3.3). Even an elementary learner who is in possession of a good reservoir of fixed linguistic forms and set phrases need not plan every single item separately but can avail himself of chunks (Fillmore 1979; Lehtonen 1990). In L2 speech production formulaic expressions are resting places in which the mind can concentrate on formulating an expression out of the scattered building blocks. With practice automation improves. The experiments of Goldman-Eisler (1968) showed that making subjects talk for a second or third time on the same topic produced a greater amount of fluent, less hesitant speech. The same approach, letting students talk on the same subject for three times with decreasing time allowance from 4 to 3 to 2 minutes, has since proved a beneficial method in fluency training (Maurice 1983). As the fluency increases, the length of the chunks increases, too (Koponen 1990, 170).

There are also many similarities between L1 and L2 pausing. According to Lehtonen (1979, 35) too fast a rate of speech and particularly the lack of pauses or inappropriate pauses can be destructive to the understanding of the student's foreign language. Like in L1, in L2, too, *the cognitive complexity of the discourse produced* affects the frequency of pauses. In Lehtonen's and Sajavaara's tests (Lehtonen 1978b, 1979; Sajavaara & Lehtonen 1978) the subjects were asked to read aloud texts of various linguistic complexity and to describe sets of cartoons (cf. 5.1.1). It was found that the amount of pausing was much greater in informal speech (60 % of total speech time) than in oral reading (30 % of total speech time). It was also shown that the percentage of pauses was a differentiating factor in the informal speech test, but not in the oral reading test.

Another parameter in L2 disfluency is *L1 transfer*. If L1 and L2 are cognates, the transfer is realized in different forms than in cases where the two languages are wide apart, like Finnish and English. The Finnish language has some aspects whose transfer into learning English is particularly disruptive to fluency development. For the first, Finnish culture - together with e.g. Japanese - is claimed to belong to the so-called high-context cultures, which have a far greater tolerance for silence than the so called low-context cultures like those of America and Germany (Hall 1977, 91). In a low-context culture most communication takes place through words while in high-context cultures the bulk of information is conveyed through the physical context. Unlike his American colleague, a Finnish television ice-hockey commentator does not say the obvious just to fill air time (Moore 1990, 11). Because a lot of silence is allowed, also pauses are longer and more frequent. The habit of long silences is transferred into L2 interaction, so that the length of a Finn's pause often makes his foreign interlocutor wonder whether his Finnish friend is still with him (Yli-Renko 1989a).

The fact that the structure of the Finnish language differs from that of the Indo-European languages is another factor that has an effect on the Finnish way of pausing. Finnish is an agglutinative language, in which words are longer and syllables carry more information than in the isolating languages. Accordingly, the Finnish language has many means of demarcating word boundaries. The stress on the first syllable, word-internal vowel harmony and case endings serve to distinguish one word from the next. Syllables are pronounced unreduced. (Karlsson 1977; Lehtonen et al. 1977). The boundary of words beginning with a vowel is in Finnish marked by a glottal stop (Lehtonen & Koponen 1977; Koponen 1990). When Finns transfer this habit into the pronunciation of English, their speech is more halting and slower than that of e.g. Swedish-speaking Finns and Swedes; the Finns lack transition smoothness (Lehtonen 1978b).

5.2.3 Classification of pauses

Besides cases of total absence of sound, hesitation is signified by other disfluency markers, such as e.g. false starts, repetitions, self-corrections, and filled pauses. Among the many categorizations of disfluency phenomena Moore's taxonomy (Figure 7) is among the most comprehensive. It is based on Clark and Clark's (1977) categorization and used for the Finnish language, but I see no reason why it could not be used for other languages, e.g. English.

Many of Moore's terms are self-explanatory, but some may need comment. With speaker-based empty pauses Moore refers to pauses used by the speaker to plan or formulate the production of the message, whereas listener-based pauses are used to aid comprehension, marking for example change of subject. Juncture pauses are used for marking off larger segments of the message, focus pauses for drawing attention to something unusually relevant. Avoidance of the obvious is used about something that can be inferred from the situation by the aid of visual cues, as for instance some incident in a television sports broadcast. In filled pauses sounds or lexical items are uttered when the

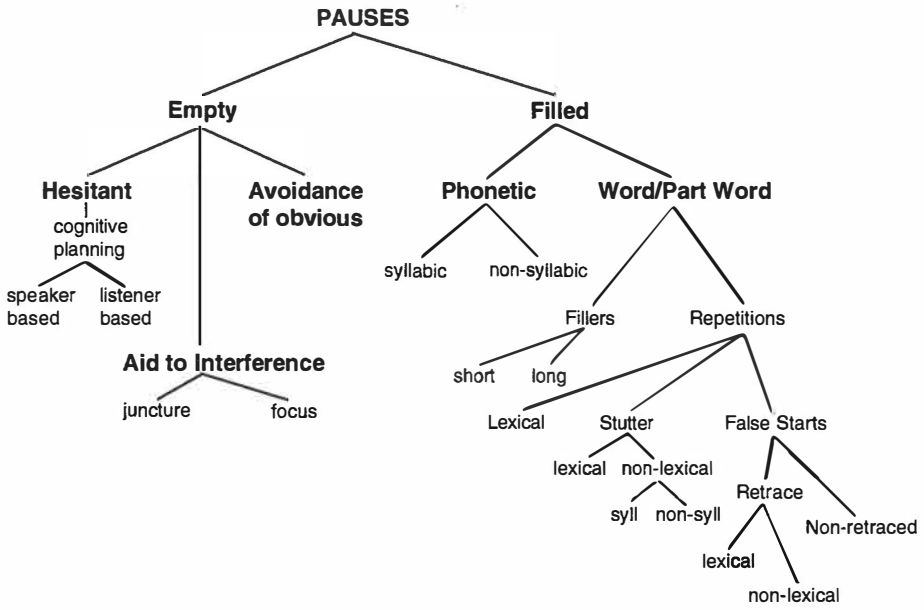

FIGURE 7 Classification of pauses (Moore 1991b, 146)

speaker needs more time but wants to maintain production and keep the channel open. It is an idiosyncrasy/shortcoming of Finnish L2 speech that it contains a more than usual amount of unfilled pauses (see e.g. Lehtonen 1978b, 1979). Syllabic phonetic pauses include, besides extended syllables, vowels and sounds like *m, mhm* (Finnish), *um, er, mm* (English), whereas non-syllabic phonetic pauses may be for instance glottal stops or creaky voice. Short fillers are equal or shorter than two syllables, long fillers are longer than two syllables.

This distinction of different disfluency phenomena is important, because it is not the mere existence of pauses and markers of hesitancy that help to determine fluency but what markers are used and how they are used. Linguists who have carried out experimental research seem to agree that among the most salient markers of disfluency are unfilled pauses and repetitions (Lennon 1990; Olynyk et al. 1983, 1990; Riggenbach 1991, 430; Varadi 1990, 4). As for corrections, the evidence seems conflicting. Olynyk et al. (1983; 1990, 148) who, among ten subjects with in other respects equal competence, distinguished between five very fluent and five disfluent ones, found that the disfluent ones had more repairs than the fluent ones. To Olynyk and his colleagues the position of the pause seemed a significant indicator of competence: the fluent speakers would resort to pausing at the planning stage, before the utterance, whereas the more disfluent ones would pause afterwards, trying to improve on the utterance. They call these breaks progressive and regressive speech markers respectively (1990, 151).

Olynyk and his colleagues' view of corrections as significant signs of disfluency is, however, not shared by all other researchers such as Lennon (1990, 413) and Riggenbach (1991, 433 and 438). On the contrary, Lennon found that after a 21-week stay in the target country, the number of self-corrections increased. He concluded that the fluency development of advanced learners may involve "an increased ability to reformulate, monitor, and self-correct production on-line". According to Riggenbach, subjects at the two extremes of fluency may use repair for different purposes.

In pauses, the total pause time is not as important as the type, frequency, length, and position, and the chunking of the different disfluency phenomena.

5.3 Fluency as a criterion in oral assessment

Out of the potential criteria for oral assessment, fluency is among the most central and the most complex. The paradox about fluency is that to be fluent in L2 one has to learn to be appropriately disfluent. A fluent speaker knows how to be silent, how to use hesitation signals and correction mechanisms, in a word, how to use communication strategies (Enkvist 1990, 25; Lehtonen 1978, 12; Sajavaara & Lehtonen 1978).

But how is fluency to be defined to work as a powerful criterion? Studying relevant literature has confirmed the view of the need for two definitions, one in a broad sense, according to which fluency is practically equal with oral proficiency or communicative competence, and one in a narrow sense, which would confine fluency to the temporal aspects and disfluency phenomena. Theoretically, the construct of fluency as a comprehensive phenomenon seems well-grounded. If, however, we try to find a distinct criterion which would aid the assessor to analyze the extremely intricate concept of speech into something tangible and measurable, saying that fluency is more or less the equivalent of communicative competence, another complex and highly debated concept, is of no help.

What is needed here is a definition of fluency serviceable in a broad assessment context with perhaps tens of thousands of subjects to be tested. Time and resources will be limited and fluency just one though significant criterion among others. A very exact definition with speech rate and pause time numerically specified would in such a case be equally futile. The difficulty lies in finding a definition which would be both optimally wide to include the essence of fluency and optimally narrow to give the assessors concerned the practical support required.

In this work, fluency will be used to denote a comparatively easy flow of speech with no unusual amount of hesitation phenomena such as unfilled pauses and repetitions. The three sets of fluency criteria (one for fluency in a reading aloud task, one for fluency in transactional speech, and one for the quality of the listening experience - fluency formulated differently -, Chapter 9) are operationalizations of this definition.

6 PREVIOUS ORAL TESTS

Since creating a many-sided oral test for a large number of testees is a demanding project, it was natural to first go through the existing tests and try to see if there was a suitable one among them. A search through the available tests proved once again that language tests are created for a specific use and not easily transferable to another context. The existing tests were, however, able to serve as partial models and to tell what is possible, or is not advisable, to do. Much of the available literature dealt with tests developed in the United States, and the two main American tests, the ACTFL OPI, the oral proficiency interview, and the SOPI, the simulated oral proficiency interview, will be presented here (but described in detail in Chapter 8). The ACTFL OPI will be discussed here, because it was used as an instrument in the present experiment, and the SOPI, because it is a language laboratory test with communicative elements. A short survey of oral tests in Finland will also be given, because it was natural to consider to what extent it would be possible to base the new test on those available in Finland.

6.1 The ACTFL oral proficiency interview

The development of oral testing procedures based on modern linguistic knowledge seems to have originated in the United States. The Second World War made Americans aware of how important it is for the military and the diplomatic corps to have personnel with good foreign language skills. With the collaboration of structural linguists, new language teaching programs were launched, and in 1956 a modern test of oral proficiency was introduced. This was an oral interview developed in the Foreign Service Institute to test its personnel to be sent on commissions abroad. In the next decade, the use of the interview spread to other government agencies, and in 1968 the skill-level descriptions used by various institutes were standardized under the title the ILR (Interagency Language Roundtable) Oral Proficiency Scale. (Liskin-Gasparro 1985; Lowe 1983; Lowe & Liskin-Gasparro 1986; Spolsky 1986.)

The oral interview was a face-to-face conversation conducted and rated by one or two highly trained testers. It was strictly structured yet individually tailored for each user. Depending on the test-taker's level of proficiency the interview lasted from 10 to 40 minutes. The resulting speech sample was usually recorded for later verification and rated on a six-level rating scale. The level descriptions or scales (from 0 for no proficiency to 5 for a proficiency comparable to that of an educated native speaker) were based on a linguistic analysis with an emphasis on functional language use. The ILR oral proficiency interview was a criterion-referenced test from the very beginning.

In the 1970s the use of ILR interview spread outside the federal agencies, and its further development was trusted in the hands of the Educational Testing Service and later of the American Council on the Teaching of Foreign Languages (ACTFL). At this stage the many years of use had led to increased understanding of the interviewing process and a consequential refinement of the instrument. However, with the ever-growing demand for accountability, an even more accurate and standardized measure was needed, and a three-nation project (the USA, Great Britain, and Germany) was set up to develop the interview into a "common yardstick". The resulting nine-level instrument was designed for wider use: levels had been added at the lower end and squeezed into one at the upper, which made it more functional at, for example, universities. New level descriptions were also created, and as the result of the enterprise the *ACTFL Provisional Proficiency Guidelines* were published in 1982. The present version, the *ACTFL Proficiency Guidelines*, was published in 1986. (For a description of the test see Buck 1989 or Liskin-Gasparro 1985.)

In the 1980s the ACTFL oral interview was put to many uses inside and outside the United States. Naturally it also became the focus of extensive research and criticism. In comparison with the structuralistic discrete-point tests it was easy to point out the merits of the ACTFL oral interview. It presents itself as a many-sided and integrative instrument encompassing vocabulary, syntax, pronunciation, coherence, functions, and situations. (Bachman & Savignon 1986, 381; Buck 1989.) Its great asset is its authenticity and undisputed face validity: it is real-life interaction between real people about at least partly unpredicted topics. Used by a skillful interviewer it is also a flexible device adjusting itself to the interests and level of each interviewee. According to most studies, it can pride itself on good psychometric qualities, such as high inter-rater reliability (see e.g. Dandonoli & Henning 1990; Magnan 1987; Meredith 1990).

In spite of its undeniable assets, the ACTFL oral interview has also met with severe criticism. It has been attacked for lack of both validity and reliability. In validity the greatest concern has been the theoretical weakness, the description of language proficiency and of the criteria, and the interview format itself. As for reliability, critics have particularly complained of insufficient attention to the method factors.

Validity. The most serious critique concerns the theoretical concept of 'oral proficiency'. In the *ACTFL Proficiency Guidelines* (Buck 1989) the proponents present an extensive analysis of the speaking skill and its components, but according to the

critics there is no empirical evidence for the alleged constituents of the proficiency construct or the claimed relative contribution of each constituent (Bachman 1988; Bachman & Savignon 1986; Barnwell 1987, 1989; Hymes 1987; Lantolf & Frawley 1985, 1988; Spolsky 1986). Neither is it clear how the constituent features should be assigned to particular linguistic levels of syntax, semantics, pragmatics, or sociolinguistics. On the contrary: communicative language proficiency (a term used by Bachman and Savignon) or communicative language ability (used by Bachman) consists of many parts, and these parts should not be lumped together for measurement but should each be assessed separately with a separate scale. It has even been claimed that the ACTFL criteria themselves constitute an absolute definition of competence, which may have nothing to do with real-world performance (Lantolf & Frawley 1985).

The ACTFL concept of oral proficiency may lack empirical verification, but so does any other concept of proficiency, with the possible exception of Bachman's model (Bachman & Palmer 1982). With that argument no standardized test could be used. However, the critics' claims about the excessive share of grammar and little attention to cultural and pragmatic aspects seem justified (Kramsch 1986; Raffaldini 1988; Savignon 1985). Likewise, the picture of a uniform educated native speaker as a norm seems to deserve criticism (Bachman & Savignon 1986; Barnwell 1987, 1989; Kramsch 1986; Valdman 1987).

Another essential aspect of ACTFL OPI validity is the very form of the interview. The main argument is that an interview is only one rather unusual genre of oral communication, and it is wrong to generalize the information received from it to other genres of speaking. Conversation is the by far commonest genre, and it is claimed that the validity of the interview depends on how far it is similar to or different from conversation. The critics have stressed the difference. The interview is said to differ in for instance pragmatics and in the affective side. The asymmetric control relationship is claimed to promote the world view of one participant at the expense of the other. Lantolf and Frawley (1988, 192) put it so strongly as to claim that "one speaker commits an act of symbolic violence against the other". Savignon (1985, 132), for her part, calls the traditional FSI/ILR interview format an interrogation rather than a conversation.

Reliability. When creating an oral interview, the test designer has to face two types of errors: the random error and the systematic error. The main causes of the random error are the test method facets: facets of the testing environment, facets of the test rubric, facets of the input, facets of the expected response, and the relationship between input and response (Bachman 1990, 119). If the test method facets vary from one test to another, they contribute to variance in test scores. If, however, the test designer tries to minimize the random measurement error by controlling the test method facets, for instance by standardizing the test, he meets with the other error, the systematic one. The test method facets will now affect all test takers' scores systematically and thus create potential sources of test bias. The test designer's dilemma is, accordingly, the fact that by increasing reliability he cannot help decreasing validity (Bachman 1988, 153).

It is the lack of attention to test method facets that Bachman and Savignon (1986) and Bachman (1988) see as one of the greatest defects in the ACTFL oral interview. They assert that in interpreting test results it is difficult to know whether a certain score is more indicative of the test taker's ability than of certain characteristics of the test. In two validation studies of the oral interview and the communicative proficiency respectively, Bachman and Palmer (1981, 1982) found sizable loadings on factors associated with the oral interview test method, and even factor loadings in which the method facets were consistently higher than factor loadings associated with the traits being measured. Bachman and Savignon and Bachman (1988, 154) therefore suggest that the ACTFL level descriptions be changed so that, instead of depicting the testee's proficiency, the characterization be narrowed to indicate more specific skills, for example the individual's ability of using grammatical structures accurately "in contexts and under the conditions that are included in the testing procedure" (Bachman 1988, 154).

As to the reported high interrater reliability figures of the ACTFL interview, they were achieved only after experiments with new sophisticated psychometric procedures (Dandonelli & Henning 1990; Liskin-Gasparro 1985; Magnan 1987). For instance in Meredith's study high reliability was attained only after three of the original ten testers were eliminated using the ANOVA analysis of variance. This meant the loss of 104 interviews out of the original 231. At some of the alleged reliability figures one may wonder: for instance Lazaraton (1992) claims 4-8 respective 5-10-minute interviews to be reliable and valid proofs of oral language proficiency.

Another cause for questioning the reliability of the ACTFL OPI is the discrepancy of the scores presented in different studies (see e.g. Carroll 1967; Dandonelli & Henning 1990; Henning 1992; Lafayette 1987; Magnan 1988; Meredith 1990). Carroll and Lafayette are examples of the strict line: they studied very advanced students - fourth year FL majors or FL teachers respectively -, but their results only represented the lower or middle levels. Of the Texan FL teachers tested by Lafayette half failed to score at the Advanced Level. Quite opposite results were shown by the rest of the above scholars, who studied less advanced or similar students and showed that many of them reached the Advanced and some even the Superior Level.

To summarize: the ACTFL oral proficiency interview has met with plenty of both positive and negative critique from the best-known experts. This is not the place to report it all (see e.g. Clark & Clifford 1988; Meredith 1990). It is only natural that an influential, standardized test should be exposed to extensive public criticism. What one may wonder is the fact how little impact this criticism has, in fact, had on the development of the test. The 1986 *Guidelines* preserved most of the criticized features of the earlier version untouched (lumping the different features together into holistic level descriptions, focus on grammar, native speaker as norm, etc.). What makes a critical study of the test onerous is lack of statistics. One might expect that those concerned with the development of an important test should have systematic figures to show about its use and results. However, from the American Council on the Teaching of Foreign Languages such statistics were, despite several attempts, not available.

All in all, the writer's closer investigation of the ACTFL OPI has helped make the picture of the test less idealistic than it was in the beginning. It is a test with many assets but also some defects. At the moment it is the best-known and perhaps also the best interview. The present writer's original wish that a new test could be solidly validated by an existing test was in any case unrealistic. A test can hardly ever be successfully transplanted into another context. In the present study the original plan will be followed: the scores of the new test will still be compared with those of the ACTFL OPI, but the value of the comparison will be dealt with due discrimination.

6.2 The simulated oral proficiency interview (the SOPI)

Even with the best of validity, the oral proficiency interview is too expensive to become a practical instrument for mass testing. The OPI demands time and high-quality expertise to elicit the needed sample and to assess it. It is, therefore, no wonder that test designers have keenly observed the development of other alternatives and, above all, the eventual rationalization brought about by technology.

The history of tape-recorded oral tests can be traced back almost as far as the history of the first interviews. The new instrument from the forties had many promising aspects: the tape-recorder standardized the eliciting phase, gave an opportunity to store the material for potential rerating, and, on the whole, helped to save costs by making it possible to test a number of people simultaneously. No wonder that semi-direct testing - elicitation by non-human means - soon became increasingly popular (Stansfield 1989, 1990a).

The tasks used in the first tape-mediated tests were in accordance with the structuralistic view of the period: reading aloud, mechanical repetition of words and sentences, giving pattern answers to pattern questions and substitution drills (Shohamy 1994; Shohamy, Reves & Bejarano 1986; Shohamy & Stansfield 1990; Stansfield 1989). Since the 1980s versions of tape-mediated tests have been widely used (Clark & Clifford 1988; Lowe & Clifford 1980), but only the introduction of the SOPI (simulated oral proficiency interview) really opened the international testing field for tape-recorded testing. The SOPI type of tests were developed by the Center for Applied Linguistics, Washington, DC. The first of the series of tests was developed for Chinese in 1986 (Stansfield & Kenyon 1992a), but they are now available for many of the less taught languages such as Portuguese, Hebrew, Indonesian, and Hausa (Stansfield and Kenyon 1992b).

The SOPI is distinguished from the rest of tape-mediated oral tests by three characteristics: its format is similar to the ACTFL OPI, so that it begins with a warm-up, uses different speaking tasks designed to produce samples ratable by the ACTFL OPI criteria, and the newest versions end with a winddown (Stansfield & Kenyon 1992a, 1992b). As means of elicitation it uses both visual and tape-mediated stimuli. Because the SOPI is planned to be

commensurate and interchangeable with the OPI, the sample is also rated using the ACTFL OPI rating scale.

The SOPI consists of six parts. Although the acronym is short for simulated oral proficiency interview, only the first section, the warm-up part, has questions about the interviewee's family, hobbies, education, etc., which are typical of an interview. The other parts would not normally be called an interview. Parts two to four present picture stimuli and ask the testee to give directions, describe in detail a drawing, and to tell a story in different tenses. Parts one to four thus present functions and topics whose command is essential at Novice and Intermediate Level, whereas tasks five to six are mainly aimed at Advanced and Superior Level candidates. Part Five (Topical Discourse) consists of five to six tasks, each offering a different subject. The testee is for instance asked to describe her favorite outdoor activity, discuss the advantages and disadvantages of public transport in the United States, or give advice on buying a used car. In Part Six (Situational Discourse) the candidate has to react appropriately in five everyday situations.

The SOPI tests for the different languages had many features in common. The total length of the test was 45 minutes and the amount of examinee speech 20-23 minutes (Stansfield 1989). For some tasks there was preparation time, and the answer time varied from 10 seconds to one minute 45 seconds. A clear distinction was made between the two skills of listening and speaking so that all the instructions were (except for Part One) given in the native language.

Research on the SOPI has concentrated on reliability, validity, and interchangeability with the ACTFL OPI. The interrater reliability was always quite high, .92 on an average, and there was remarkable agreement in the results of the parallel versions. The correlations in the studies in which the same examinees were assessed by both the OPI and the SOPI were also so high, averaging .93 (Stansfield 1990b), that the use of the SOPI as a parallel version to the OPI could be recommended without the least hesitation. However, the OPI with its shorter turns may suit the Intermediate or lower Advanced students better, whereas the SOPI with authentic longer terms may be more appropriate for the Advanced-Superior students (Stansfield & Kenyon 1992b).

To the present writer, who had from the beginning set out to create a language laboratory test, the publication of the SOPI was, of course, of utmost interest. It was interesting to note that there were similarities between the two tests, such as the resemblance of SOPI Part Six and LLOPT Part Six, as well as the length of the tests. The temptation to use a standardized test would have been great. However, there were also reservations. Stansfield himself had advised against using the SOPI for levels below Intermediate Low (Stansfield 1990a). Besides, the present writer learned about the SOPI only after she had designed and pretested her own test, and it seemed too late to forsake all the work done. Nevertheless, the most serious consideration was the fear that the large-scale assessing of the SOPI would need expertise on rating the ACTFL OPI that was not available in Finland. The use of the SOPI was thus given up, but there remained a desire to compare the two tests later on.

A new aspect to the juxtaposition of the ACTFL OPI and the SOPI was brought by Shohamy who carried out a linguistic analysis (1992, 1994). She

claimed that mere correlations were not sufficient evidence to argue that two tests measured the same trait and wanted to explore the samples further. She transcribed both OPI and SOPI tapes and compared the language from a qualitative point of view. The results did not give cause to generally place one test above the other, particularly not at the intermediate level such as it is represented in the Finnish matriculation examination. They showed, however, that the language elicited by the SOPI was more formal than the more intimate language produced by the OPI. The two test formats might be comparable in some areas and not in others, and a validation study should be carried out from multiple perspectives. If important decisions are based on the test results, the use of both types of tests may be recommended.

6.3 Previous oral tests in Finland

Before designing a test that was aimed to be used as a subtest in the matriculation examination it was considered natural to make an inventory of previous Finnish attempts to assess the senior secondary level speaking skill. For the younger age-groups Finland had been in the vanguard of oral testing: the 14- and 15-year-olds had in 1971 participated in an extensive international IEA research program, and there had been promising smaller-scale results also in individual schools (Takala & Saari 1979; Takala 1977). As for the senior secondary school, there were three earlier experiments of a test for either assessing the speaking skill in the matriculation examination or for finding a way to compensate such assessment: Hirvonen's (1971-74), Hellgren's (1982), and Yli-Renko's (1989b). Below there will be a short description of each one of them and also an account of the reasons why they could not be used for the present purpose.

Hirvonen's pioneering work. Hirvonen was ahead of his time. He criticized the matriculation examination strongly and claimed that the test format, the English-Finnish/Finnish-English translation, was a powerful impediment to the development of Finnish language teaching towards more practical lines. He wanted to show that it was feasible to create and set up an examination that would help learners build up a many-sided and usable proficiency.

Hirvonen carried out a multiform test in reading, writing, listening comprehension, grammar, and vocabulary with 1,023 subjects elected by means of systematic sampling, and even in the speaking test he had as many as 306 subjects (Hirvonen 1974, 12). The test formats that he used were mainly multiple-choice tasks, which could be assessed objectively. The speaking skill, however, was tested entirely with tasks that were assessed subjectively. The oral test was divided into two parts: pronunciation and oral expression. Hirvonen did not consider pronunciation very important for communication (1974, 60), but decided to test it, because it was one of the few aspects that was specifically mentioned in the contemporary curriculum. The pronunciation test was a reading aloud task in which he assessed phonemes, stress, rhythm, and

intonation in sentence context. In the test of oral expression there were two parts: answering general questions and reacting in contexts.

The very creation of a speaking test was revolutionary in the days when language teaching was guided by the translation-grammar method and the structuralistic approach to language. Seen from today's vantage-point, however, the test is dated. The speaking skill has improved so much since the early seventies that Hirvonen's test would now be far too narrow for L2. The structure of the test is atomistic, and the individual turns are so short that there would be little opportunity to measure for instance fluency. Transactional speech - an important discriminator among more advanced students - is not assessed at all. Hirvonen (1973b, 26-28) admitted himself that many other aspects could have been measured, but wanted to postpone it till a time when speaking would be properly taught at school.

Hellgren: Speaking can be tested by writing. Also Hellgren (1982) considered it important to test the oral skill, but he does not seem to have worried about the washback of the test format. He believed that an oral test would be so expensive and complicated that it would be unrealistic to think that it could ever be used. Therefore he set out to investigate whether the speaking skill could be assessed by means of a writing test. He chose a text, compiled 20 questions about its content, and had the text read aloud on an audiotape. The 406 subjects were matched by a cloze test and divided into two groups, one of which answered questions 1 to 10 first orally and then questions 11 to 20 by writing, while the other group performed the tasks in a reversed order. Regardless of whether the test-takers answered by speaking or by writing, they could use 40 seconds for the answer. As the students' oral answers correlated highly with the written ones, Hellgren felt justified to conclude that speaking could equally well be tested by writing.

Hellgren's test design reflects the thinking of the early eighties. Oller's Unitary Hypothesis, according to which there is a common factor in all the four skills, can be seen as the theoretical frame that makes Hellgren's conclusions understandable. However, the study had also many other sources of error. The very test format ignores one of the basic differences between the spoken and the written genre, namely the time constraint in speaking. If the test-takers are given the same amount of time for the spoken answers as for the written ones, it is no longer possible to speak of a valid oral test. Moreover, the twenty tasks were all questions about the content of the passage just heard, which meant that the candidate could get maximum scores by simply repeating what she had heard. Only short one-sentence answers were required, and the range of language, like range of syntactic features, was narrow. The criteria were also limited. Hellgren denied the significance of communicative competence in FL learning and testing, and he had, therefore, no use for such criteria as discourse competence, functional competence, or sociolinguistic competence. He used only two criteria: how much of the information had been mediated and how correct was the language usage. The oral responses were transcribed, and after that they were judged in exactly the same way as the written answers.

Yli-Renko's test of German. Though Oller disclaimed the Unitary Hypothesis a year after Hellgren's dissertation was presented (1982), Hellgren's ideas were not rebutted in Finland but supported. Both Kristiansen (1990) and Norris (1991) designed studies in which they claimed to have tested oral skills in writing. However, the next oral test that was aimed to be used at the end of the senior secondary school was more communicative. In 1989 Yli-Renko published a report on a test of German, which was based on the American Army Defense Language Proficiency Test of 1982. It was mainly an interview concerned with everyday affairs and matters of general interest to young people, but contained also a role-play and a section with picture description. The test lasted 10-15 minutes and was administered by a native speaker of German. Of the 228 subjects 38 had studied German as L2, 113 as L4 and 77 as L5.

It was important for Finnish foreign language teaching that at last a communicative test was published. Unfortunately, however, the psychometric qualities of the test were not up to the standard of the content. The first aspect that captures attention is the brevity of the interview. The Defense Language Proficiency Test recommendation is 10 to 40 minutes. Even if 10 minutes had sufficed for the L4 and L5 students, 15 minutes was certainly too little for the L2 students. It is also unusual how minute distinctions were based on so short a sample. In her criteria Yli-Renko used a 20-level scale (a scale from 0 to 5 as modified by +, -, and ½) and assessed with it not only the general impression but also the subskills of pronunciation, grammar, vocabulary, fluency, as well as understanding and reactions, each subskill separately. For accomplishing all this she only listened to a sample once. And moreover, despite the 20 levels and the short sample, the alleged correlation of the 2-3 assessors was as high as .84-.94. On the other hand, the correlations with the school final grades were unusually low (.35-.46), though German is commonly taught in so small groups that the teacher has a good knowledge of also the students' oral proficiency.

Two communicative tests in adult education. Though there have been three Finnish efforts to give an example of an oral school-leaving L2 test (in Hellgren's case actually a simulated oral test), they have not left much to be used in the present situation. Some of the test formats used by Yli-Renko might be used for an L4 or L5 test, but for L2 a more demanding test would have to be created. The present study has been carried out for that purpose.

There are now also two communicative tests in Finnish adult education, in which the oral part is either completely or to a great extent administered in the language laboratory. At the University of Jyväskylä research has been carried out which has led to the creation of two extensive examinations for adult language learners: *Työelämän kielidiplomi, TKD* (the Finnish Foreign Language Diploma for Professional Purposes, actually created in the late 80s but taken into wider use in the 90s) and *Yleiset kielitutkinnot, YKI* (the Finnish National Certificate of Language Proficiency, NC) (Huhta, Sajavaara & Takala 1993; Luoma & Takala 1993). These tests came too late to have any influence on the present study, but they could well be used as partial models for the oral section of the matriculation examination.

7 DESIGNING THE TEST

An advanced speaker's oral language proficiency is so versatile that to test it all, the tester should for a certain period of time follow the testee everywhere and record him continuously for a certain period of time. Because this is not possible, a choice must be made: a language test is only a broader or narrower sample of reality. In deciding which elements to test and which means to use for it, the most important consideration is the purpose of the test and the context in which it is used (Morrow 1979).

The present test was designed as if it were a secondary school final examination, i.e. attention was constantly paid to its purpose. That a test should serve as a national school leaving examination had at least four corollaries: a vast number of students would have to be tested; the examination would have to be based on the official curriculum; because the examination would be of great importance in the testees' lives, it should have first-class testing qualities; the design of the examination would have consequences on language teaching and learning in Finnish schools. These corollaries would, respectively, affect the efficiency, validity, reliability, discrimination, and washback of the examination.

In what follows the above testing qualities, except discrimination, will be discussed in turn. Discrimination is excluded because with a large heterogeneous population and cognitively complex tasks, discrimination will follow naturally. After a general discussion of each quality, the description will proceed to delineate the choices and decisions which were made in constructing the present test. The constructing and planning were, on the whole, considered a very important stage in the test development, for if the a priori measures of e.g. validation do not succeed, the a posteriori efforts will not be of great help, either (Weir 1989).

7.1 Efficiency

An efficient test produces maximum information with minimum resources. In a mass test like the matriculation examination it is difficult to avoid great expenses, but the

ultimate costs will depend on the required quality of the results. Does the present test have to have the same reliability as for example the matriculation examination, or is the oral test there only to make the teaching of speaking more credible? For the latter purpose a less expensive test will suffice, but if important decisions are made on the basis of the results, a more extensive, multiform examination is needed. To cite Hughes, "accurate information does not come cheaply" (1989, 37).

Any test involves both material and labor costs. Because labor is more expensive than material in present-day Finland, a means of saving would be to minimize the share of labor. Labor is needed at three stages: test design, administration, and assessment. At the moment it is not yet possible to design and assess a direct test without human input, and therefore the only way to cut the cost of labor is to try to save at the stage of eliciting the output. One of the commonest ways of acquiring the needed sample is to use an interview or a language laboratory or peer or group discussion or role-play. The most expensive means is the interview, but the use of language laboratory and peer group pose other problems. In the case of a machine the authenticity of genuine interaction is lost, but relying on another testee might spoil the whole interaction: even if the partners were carefully matched beforehand, the attempt would be risky. For instance, if one of the intended partners should have a bad day or be absent altogether, the whole discussion would be spoilt.

The interview was rejected for economic reasons and the peer group interaction because of the risk. It was now necessary to try to minimize the defects of the remaining alternative, the language laboratory test. A good language laboratory test would be as communicative as possible, with at least simulated interaction and natural time constraints. A good example seemed to be the new American-Israeli test called SOPI, the semi-direct oral proficiency interview. Though most of this test is mediated by tape, it has many communicative features (see section 6.2). In the name of the Finnish test the word interview was avoided, because the Finnish version contained many other subtests as well. The test was simply called the Language Laboratory Oral Proficiency Test, LLOPT. While the authenticity of the interview was lost, another kind of authenticity was gained: instead of a Finnish person acting as an English-speaking interviewer, the instructions and the American visitor's part were read by an American teacher. The frame of the test, the story, took the testee into an 'authentic' everyday situation (see Chapter 9).

The software cost of the LLOPT is comparable to that of the ACTFL OPI except for the fact that the LLOPT cannot be recorded on the video. The main material cost is, of course, the investment in the AAC-type language laboratories. If the test is carried out as part of a school final examination, it has to take place simultaneously all over the country. Even though students can be tested one group after another, which in many schools is the case at present in the matriculation listening comprehension test, there is a limit to how long the students can be kept enclosed in a room waiting for their turn. If the testing of one group takes 60 minutes, six successive tests seems the extreme maximum, which would mean that the number of language laboratory booths in a municipality should be the number of the candidates divided by six. To the municipalities still lacking the sufficient equipment this would mean a considerable expense, which would, however, be greatly compensated by the fact that by buying a language laboratory the municipalities would get an effective language learning instrument.

The labor cost and need of expertise involved in the LLOPT is, however, much smaller than the expense of the interview. A new version of the test is needed twice a year, but this involves only a few persons. The carrying out of the actual test can be done by any teacher of English, and the scoring is comparable to the present assessment of the matriculation essays. Compared with what the ACTFL OPI would require, the cost of the teacher training will be reasonable.

7.2 Reliability

The introduction of performance testing, which in the field of language teaching was more or less equal to communicative testing, brought about a changed concept of validity and also a less stringent attitude towards reliability. A complex test, which gives plenty of individual freedom in task performance, cannot be as neutrally assessed as a multiple-choice test. The same subjectivity which has been approved of in essay assessment must also be accepted in oral testing. When validity and reliability come into conflict, it is reliability which has to give way (see e.g. Weir 1989).

The various facets of testing are dependent on the context and purpose of the test. Because the matriculation examination is a high-stakes issue for the participants as well as for society as a whole, it is important not to undervalue reliability. While the a posteriori methods of controlling it are not available, the a priori validation, the test design, is even more important. In designing the present test, the following aspects concerning reliability were considered: elicitation, length, recording, affective factors, raters.

Elicitation A factor in which the OPI and the SOPI or the LLOPT differ and which makes the LLOPT more objective to score is elicitation. In an OPI the interviewer's moods, likes, and impulses may greatly affect the course of the conversation, whereas in the SOPI or the LLOPT the elicitation is the same for all. A similar principle applies to all tests of productive proficiency: the more open-ended the elicitation makes the task, the more diverse are the products and the more difficult they are to compare. Also pictures, particularly those with many details, leave a task more open than what interpretation tasks do. For this reason the present test was designed to be carried out without pictures.

Length The length of the test is a cost efficiency issue, as well as a reliability one. An administrator would applaud a short multiple-choice test, whereas a testing expert knows that there is no shortcut to accurate results (Hughes 1989, 37). It may be desirable to make the trial version of a test longer than the final examination would be, so that redundant and less successful items or subtests could be deleted and possibly replaced by new ones. It is also possible that the assessors will notice that they will not need so much material to make a decision on the grades. This was the procedure in the present test: just to be sure and to give the experiment material, some parts were made longer than would be needed in a national test. Yet, there is a limit to the length of the test regardless of cost. If the test is too long, fatigue lowers

reliability. In this case, practical concerns, like the normal working spell at schools, made 45 minutes seem a suitable total length.

Recording Any test is subject to a number of unintended variables which distort reliability. In this particular experiment, with the ACTFL interview and the SOPI test, the tester was especially concerned about two factors: technical considerations and the influence of affective factors.

A SOPI or LLOPT type of test is regularly recorded in a language laboratory, and a high-stakes OPI test, in which reliability is of great importance, has to be recorded, too. As the LLOPT lasted altogether 45 minutes, but maximally only 22 minutes of it was examinee speech, it was important to plan the test so that no pauses would be recorded. This was possible to accomplish by either human control or machine control. The latter alternative was chosen, and the recording was made in collaboration with the manufacturer of the language laboratory.

Affective factors are not usually considered in discussions of reliability, although their influence may be a major source of extraneous variance. To reach an optimum result, the testee has to be able to concentrate on the task fully; if part of the limited capacity of working memory is taken up by affective stimuli, such as anxiety and apprehension, the outcome does not reflect the testee's true proficiency (Wine 1980). Language may be affected by several types of anxiety such as evaluation apprehension, communication anxiety, and interpersonal anxiety. For instance in L2 pronunciation within-speaker variation may be affected by factors like the inhibition level, the identity and personality of the interlocutor, the emotional impact of what is being said, and the extent to which the speaker monitors himself (Munro & Derwing 1994, 254). Weaker candidates are more easily affected than good. In the demonstration of different skills, anxiety is at its worst in a speaking situation, and, of course, in a testing situation. In the present test the subjects were, in addition, exposed to a new test format and an unfamiliar interviewer (for the effects of test anxiety see e.g. Hembree 1988; Horwitz, Horwitz & Cope 1986; Madsen et al. 1991; Young 1986).

Although oral tests are more exposed to anxiety than reading comprehension and writing, there are research results which indicate that the face-to-face interview is generally experienced as positive (Madsen 1982; Shohamy 1982a). Language laboratory tests or other machine-mediated tests, on the other hand, are experienced as more negative than human-mediated tests (Shohamy 1993; Shohamy, Reves & Bejarano 1986). In Jyväskylä two comparisons of an interview and a language laboratory test were made (Halvari 1996; Luoma 1997). Both researchers found that the examinees preferred the interview. In both cases, however, the language laboratory test was a new test format to most participants. For the present test, the teachers of the classes tested were told in advance that there would be a language laboratory test, but because the tester did not want the subjects to train for any particular test format, no details were disclosed. However, it was considered important that every examinee should take the more pleasant test, the interview, first. The testees would thus have some practice of an oral testing situation with an unknown tester. Every effort was taken to make the interview a gratifying experience.

When the LLOPT was designed, special attention was paid to the affective factors. In addition to music and a pleasant, encouraging voice as the reader, three factors were particularly paid attention to: easy-to-difficult-to-easy sequencing, content, and sufficiency of time. The test was opened by very easy warm-up questions, and the first subtest proper consisted of reading aloud. Every subtest was more demanding than the previous one with the difficulty culminating in the last but one subtest, after which there was an emotionally appealing final section. All the five subtests were connected by a story which was meant to evoke pleasant associations.

The pauses for processing the material and producing the response were problematic. On the one hand, an attempt was made not to arouse anxiety by giving too little time to think and answer; on the other hand, there were the validity concerns relating to natural time constraints. Some people may claim that short pauses test also personality, and disfavor slow students. But so does real life. Time is precious, and few listeners have enough patience to wait for a message in which a word comes every 30 seconds. Anyhow, natural speaking speed is a quality worth striving at, and if the washback of the final test is apt to promote it, no harm is done.

Raters Like the length of the test, the number of the raters needed is a concern of cost efficiency as much as of reliability. For a trial test maximum information would have been needed, but the resources were very limited. Of even greater importance than the number of the raters is their expertise. Wesche's (1983) demands are great: communicative tests of global productive skills can be reliably assessed only by native raters with a long experience with the particular test format and scoring grid. To rate the school final test - and at the same time to comment on the new test - the raters should be familiar with language teaching in the Finnish senior secondary school. To rate the ACTFL interview, they should have the authorized training. For the language laboratory test, the training should be provided by the test author. For the present study, the assessment sessions were negotiations in which feedback from the two coraters influenced the final shaping of the criteria.

7.3 Validity

For measuring such a multidimensional phenomenon as speaking proficiency, no single measuring instrument is enough, and any instrument only gives us approximate information (see e.g. North 1993, 157; Spolsky 1993, 209). To know how exact or inexact the instrument is, it is validated. However, validity is a complex concept: according to Angoff 1988), 16 different validities can be distinguished. Many writers regard construct, content, and concurrent validities as the most fundamental. They agree that construct validity, the "mutual verification of the measuring instrument and the theory of the construct it is meant to measure" (Angoff 1988, 26), is the most essential of all and also the most difficult to establish (Anastasi 1986; Bachman 1990; Cronbach 1988; Cumming & Berwick 1995; A. Davies 1990 ; Messick 1988, 1989; Moss 1992; Weir 1988).

When the significance of construct validity has been stressed, many writers have renounced the tripartite division into construct, content, and concurrent validities. They speak of unified validity instead (e.g. Frederiksen & Collins 1992; Linn, Baker & Dunbar 1991; Moss 1992, 1994). At the same time a more holistic view of testing has gained ground. Psychometric discrete-point multiple choice tasks have been replaced by direct performance testing, in which more open, complex tasks give extended latitude to individual candidates. The test scores produced by such methods are not so easily comparable, and the whole purpose of testing is seen as a process in which individual scores do not matter as much as the wider consequences of testing on teaching and learning. The context and purpose of testing are central.

A school final test, however, must also differentiate the test takers, and it must be introduced in a form that seems valid to teachers and students. Accordingly, an ordinary tester like the designer of the present test had to disregard the newest theories and proceed in an unsophisticated way: take the existing scientific findings about language proficiency for granted and write the test on the basis of one of the models. For construct validation the test was derived from the concept of communicative competence and the nature of spoken language, and in the operationalization of the construct the principles of a communicative test were applied (see section 7.3.1). The rule of constructing a speaking test so that it should contain "a representative set of tasks that cover the spectrum of knowledge, skills, and strategies" needed for the activity being tested (Frederiksen & Collins 1989, 30) was followed. Because validity was regarded as very central, much effort was invested on the a priori validation of the three central aspects of the test: content, format, and criteria. What content to choose, which test formats to use, and which criteria to apply, were central decisions to be made. They will be described below.

7.3.1 The content

In a school final test it should be natural to evaluate its content validity by comparing it with the curriculum. In Finland, however, it is not possible. For foreign languages Finnish curricula have been rather general, and since 1994 no detailed curriculum for Finnish schools exists any longer. The common curriculum has been replaced by a *Framework curriculum for the senior secondary school 1994*, on the basis of which schools are to compile their own curricula. This is one of the reasons why the school final test (the matriculation test) can be called a proficiency test. Morover, for a proficiency test there is no content or content validity (A. Davies 1990, 83), and if the present test is considered a proficiency test, it has no content validity, either.

However, at the same time as Davies claims that there is no content validity for a proficiency test, he maintains that the proficiency test constructor must simulate a syllabus which can provide the content validity needed. This is the situation in the present case: the simulated curriculum is the 'hidden' school curriculum, i.e. what schools actually teach or what they are known to teach. In the Framework curriculum, six topics are mentioned, among them studies and school (see Subtest 4 below), and in addition, there is the material that is always a part of an extensive L2 syllabus, such as coping with everyday situations. The present tester's familiarity

with language teaching at school gave her an insider's view of this 'hidden' curriculum.

The Framework curriculum implies clearly that in Finnish schools foreign language learning and testing are communicative. This means that in a school final test the constructing principles have to be communicative. The best-known guidelines of a valid communicative test include those put forward by Morrow (1979) and Weir (1988). According to Morrow, the language should be authentic and unpredictable and processed in real time. To make it possible to judge the appropriacy, every utterance should have a purpose and a situational and linguistic context. Weir claims that the test should contain relevant information gaps and it should show whether the candidates are able to process appropriately sized input. It should be planned for the needs of the candidates, and it should possess face validity for them. The test designer should consider both the size of the total text and the representativeness of the individual tasks: lexical range, grammatical range and complexity, and functional range, i.e. the variety of illocutionary acts involved in the event.

For the present study an attempt was made to write the content of the test in accordance with the *General Framework* and the principles put forward by Morrow and Weir. The test consisted of six subtests, which were all of different sizes and formats. They were, however, all connected by a story, which acted as a kind of framework. This is the content of the subtests (for the complete test see Chapter 8; the content and the criteria are also presented in Table 5):

Subtest 0, Warm-up. The function of part 0 was to familiarize the testees with the procedure and to make them feel at ease. Some simple everyday questions about hobbies and preferences were asked.

In *Subtest 1, Reading a Letter Aloud.* The candidate received the written material and was asked to read aloud a letter for her classmates. In the letter an American youth orchestra inquired whether they could come and visit the student's school.

In *Subtest 2, Interpreting.* A member of the American orchestra has arrived in the testee's home. The testee's mother does not speak English, so the testee has to act as an interpreter.

In *Subtest 3, Telling a Story.* The testee tells her American guest an amusing story which she has read in *Helsingin Sanomat.* The story tells about Manya Joyce, who celebrated her 86th birthday by parachuting 2900 meters.

In *Subtest 4, Presenting Finnish Education.* The American guests are in the testee's school. It is the testee's task to explain to them the Finnish school system (delineated in Finnish in the written handout) and to particularly describe the senior secondary school.

Subtest 5, Reacting in Situations and Expressing Opinions takes the testee on a return visit to America and lets her spend an eventful day in Cincinnati. As the title says, she has to cope with different situations and present various opinions.

After the test had been designed, it was shown to several secondary school teachers for comments. Naturally also the teachers who participated in the National Board of Education experiment (see the Introduction) came into contact with it. According to the teachers, the test is representative of the language skills and contents taught in the senior secondary school.

7.3.2 The format

The examinee's score does not reflect his language proficiency alone, but also other things, like different elicitation tasks and test methods as well as the choice of raters (Bachman 1990; Bachman & Palmer 1981; Chalhoub-Deville 1995; Ellis 1985, 1987; Larsen-Friman & Long 1991; Shohamy 1983, 1984). According to Bachman (1990, 350) sources of variance in language test scores are communicative language ability, personal characteristics, random factors, and test method facets. Chalhoub-Deville (1995, 17) claims that the test method has at least the following four effects on the result: it affects the measured constructs, (2) influences students' scores differentially, (3) taps diverse aspects of students' L2/FL oral proficiency, and (4) produces varied attitudes on the part of test-takers.

Many of the principles Morrow and Weir mention as characteristic of a valid communicative test are applicable not only to the content of the test but also to the format. One of them is the requirement to embed every task in a life-like situation, for which the sociolinguistic factors are given in the instructions. Considering how many of such factors there are (e.g. the interlocutor's age, status, personality, manner, mode, tone, tolerance of linguistic and stylistic failure, purposive domain, task, delicacy, complexity, setting, medium, channel; B. J. Carroll 1980; Higgs & Clifford 1983; Munby 1978), it would seem obvious that by changing several of them the test writer can create plenty of variation even within one test format. However, more than one test format is needed, because an advanced examinee, like the senior secondary school last-year student, has to be able to function in more than one genre (Bachman 1990; Wesche 1983). Morover, different examinees may excel in different genres. Thirdly, it is important to see how well different formats serve the test development.

An attempt was made to implement the above principles for the LLOPT. In the test there were both short (Subtests 0, 2, and 5) and long (Subtests 1, 3, and 4) tasks, and special attention was paid to eliciting also transactional speech. The great problem that the tester met in realizing these principles in a language laboratory was the question of authentic interaction. The result in Subtests 0, 2, and 5 was simulated interaction which did not continue beyond one exchange, but that particular pair sounded genuine.

Below the subtests will be described in detail.

Subtest 1, Reading a Letter Aloud. The easiest oral test to arrange was *reading aloud.* The test designer knew that it is not a technique recommended in communicative testing theory (Hughes 1989, 110; Underhill 1987, 67), but it seemed worth probing how much information such a simple test would give. If it gave similar information about for example pronunciation and fluency as the other tasks, it could occasionally be

used as one of the alternatives. On the positive side, it is simple to compile for any level, always available, and also reliable in the sense that every testee has the same input. It can be designed so that some segmental or prosodic features are emphasized, or it can be constructed as a general test of pronunciation and fluency. The deficiencies of a read-aloud test are also easy to name: reading aloud is a skill rarely needed, and it does not tell anything about the examinee's ability to compose a message. Even native speakers differ in their skill in reading aloud. If it is excessively used as a test format, it may have a negative washback effect.

To reduce the feeling of artificiality in a read-aloud test, the text was presented in a communicative situation in which it would be natural to read it aloud. As all the subtests should also discriminate among students, some long and difficult words were implanted in the text. The words were chosen so that the best testees should, at least passively, be familiar with them all, whereas a weaker student's pronunciation might show that she did not know them. Another difficulty for the less able students was the numbers which had to be read aloud .

To give the students an opportunity to get used to the testing situation and to familiarize themselves with the story, the passage was made rather long. It was thought that such a long reading aloud at the beginning of the test would also give the raters a chance to get used to each examinee's voice and pronunciation. The information received from this experiment was expected to help decide how long a test would be suitable in the potential future school final test.

Subtest 2, Interpreting. In language testing literature little attention has been paid to *interpretation* types of testing. Nevertheless, interpretation into and from a foreign language is a common real-life activity, and its use in testing should give a valid picture of the testee's proficiency. A reason why interpretation is used rarely in testing may be the fact that many test designers still have less encouraging memories of the use of translation as the dominant testing method. However, interpretation does not require a word-by-word translation of the speaker's discourse, but only the conveying of the meaning. Such test format offers the testee ample opportunity to use different communication strategies. Its merits are the ease of compiling, flexibility for different language uses, good (face) validity, natural time constraints, and, in the language laboratory, similar eliciting to all. A disadvantage is the prospect that its use in testing might lead back to the frequent use of translation - not interpretation - as a form of exercise.

Because oral interpretation is not commonly exercised in Finnish schools, an easy version of it was planned for the second subtest. Another reason was the fact that this subtest was the first in which the testees were asked to actively produce language. The students' task was to interpret their mother's everyday conversation, mostly questions, to an American visitor. The emphasis was on simple vocabulary and basic grammar, particularly asking questions. It is the present tester's experience that even after ten years of English studies the weakest students have difficulties in formulating correct questions.

Subtest 3, Telling a Story. Considering the situations in which a young person would need her major foreign language, it seemed necessary to test also *transactional speech* and even to make the testee give a longer account of something. For the future

development of the test, it was felt useful to experiment with two transactional tasks with different types of elicitation. The story of the parachuting lady was meant to be rather easy to assess, because all subjects had exactly the same input, a short piece of news from *Helsingin Sanomat*. The testee did not have to strain her memory by trying to think of what to say, because everything was given in the article. However, the task involved a challenge to the knowledge of vocabulary: to gain maximum points the testee could not avoid rendering any of the colorful details.

Subtest 4, Presenting Finnish Education. The fourth part seemed a proper stage to give the students great challenge. In a test which has to differentiate between the candidates there should be a task which, beside the linguistic requirements, offers the testees cognitive complexity as well (Linn et al. 1991). In this *transactional task* the conceptual framework was given in Finnish in the handout, but the testee had to plan how to make her presentation consistent and informative. It was originally planned that one of the criteria for this task would have been the cohesion and coherence of the text, but the attempts to assess it, and particularly to design the criteria, proved unsuccessful.

From the very beginning the test designer's wish had been to include a part which would test the students' knowledge of culture. As the Finnish curriculum is quite general, it was difficult to think of a subject that every student should be familiar with. However, the third part in Subtest 4, in which the student was asked to compare the Finnish senior secondary school and the American senior high school, gave an opportunity to display such knowledge. The American school is dealt with even in junior high school textbooks, and school and education are central themes in the Course Three Senior High School Curriculum. It seemed well-founded that also students who had not been to the United States should get full points for this task.

The final part, *Subtest 5, Reacting in Situations and Expressing Opinions* gave the testees an opportunity to show another kind of cultural knowledge, their *pragmatic competence.* During her imaginary day in Cincinnati the student had to react politely in eight different situations. This was as near natural *interaction* as was possible in a language laboratory test. The situations followed one another with only a short pause for the student's answer, which created a time constraint similar to that in natural conversation. In addition to the eight situations, there were three tasks in which the testee had to express her opinion on a current issue. Also these tasks were included in the test with a beneficial washback in mind. If Finnish young peope are poor at expressing an opinion and at debate, as has been claimed (see e.g. Maude 1980), they should practise argumentation in both the mother tongue and the foreign language.

7.3.3 The criteria

If the test format is important, so are the criteria. The assessment of a communicative test should be criterion-referenced, and great care should be taken to design unambiguous criteria and to train reliable raters (Morrow 1979; Wesche 1983). But how to find valid criteria and how to weight different criteria and different subtests in the final assessment? Is for instance grammar more important than pronunciation?

And how many criteria and levels of criteria could the assessor's working memory reliably process at one time?

As the whole test was derived from the theory of communicative competence, it seemed natural to establish the criteria on the same model. Several criteria can be found in the models and descriptions of language proficiency and/or communicative competence. Huhta (1990) mentions the following seventeen for a speaking test: pronunciation (individual sounds), pronunciation (prosodic features), fluency, grammar, vocabulary, situational appropriacy, conversation strategies, organization of speech, accuracy, range, intelligibility, understanding, repair strategies, need of support in conversation, size of speech, content, and task achievement. Chalhoub-Deville (1995, 21) used different criteria for different test formats. Some, like 'confidence', were considered useful in all tests whereas 'linguistic maturity (simple versus complex)' was used for interview and narration, 'giving detail unassisted' for interview, 'proper temporal shift', and 'creativity' for narration, and 'student's ability to melodize the script to make reading meaningful' for read-aloud. In theory, the most accurate assessment would be achieved if many criteria were used, but in a practical testing situation the tester has to think of feasibility as well. On what grounds should the criteria be chosen?

If seventeen or more criteria were too many, a way to find the right set would be to see what had been done in the existing tests. The only printed communicative oral test for this level in Finland was Yli-Renko's L3 test from 1989 (1989b), and it really seemed to have a solution to the problem of criteria. As the theoretical basis of her test Yli-Renko had used Higgs and Clifford's (1983) well-known Hypothesized Relative Contribution Model of Speaking Proficiency, which shows how the five basic subskills of proficiency - vocabulary, grammar, pronunciation, fluency, and sociolinguistic skills, - are proportioned at different levels of learning a foreign language. To verify their model, they had introduced it to some fifty foreign-language teachers of the CIA Language School and asked them to rate the relative importance of the five contributory skills at each proficiency level in the language which they taught (17 languages altogether). The average of the fifty teachers' ratings verified the original Higgs - Clifford model (Figure 8).

It would have been tempting to accept Higgs and Clifford's model and to weight the criteria in accordance with it. However, in spite of the attention the article has aroused (see e.g. Bachman & Savignon 1986, 385-6; Lantolf & Frawley 1985, 341; Magnan 1988; Yli-Renko 1989b), it appeared to lack validity to the present writer. For the first, the teachers who were used to verify or default the hypothesis seemed too homogeneous a group: they all worked at the same school and evidently shared a more or less similar paradigm of language learning and teaching. For the second, an average of 17 languages does not necessarily tell a great deal about any particular language. The role of e.g. pronunciation is quite different in a language like English if compared to, say, Finnish. Thirdly, the choice of the five subskills relied on tradition and not on verification. Finally, in their argumentation the writers referred to "data reported elsewhere in the literature" without giving any references.

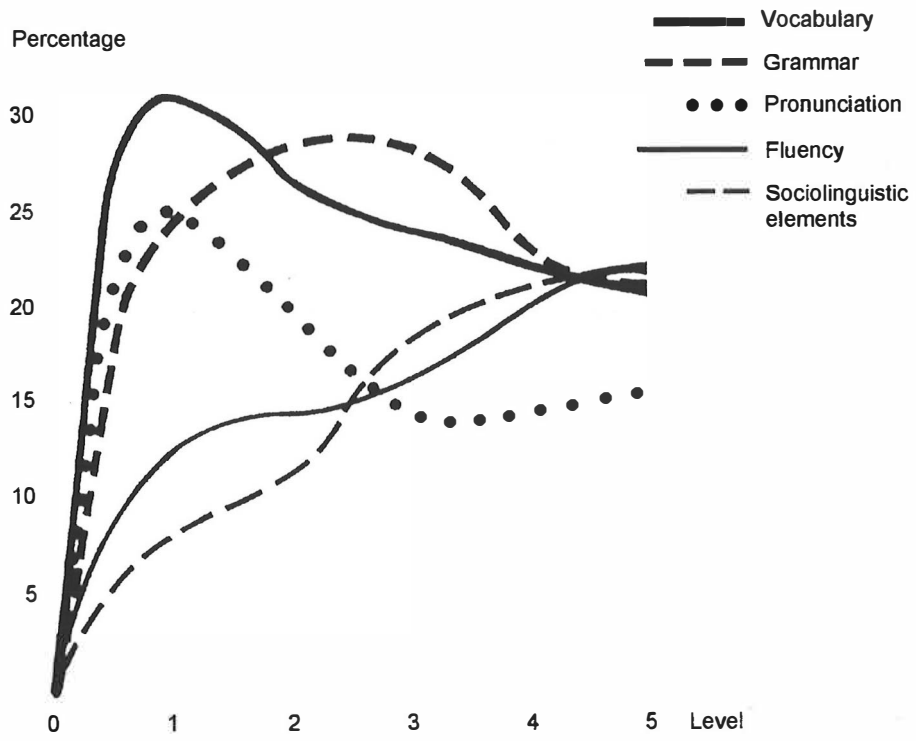

FIGURE 8 Hypothesized Relative Contribution Model, All Languages. The percentage refers to the hypothesized contribution in total language proficiency, the levels represent the Interagency Language Roundtable Oral Proficiency Interview levels (from Higgs & Clifford 1983, 69).

Since Higgs and Clifford's model could not be used, the only possibility was to construct the criteria for the test. They had to be based on the usual principles: simple to use, reliable, and having good washback. The following guidelines were also used:

- The criteria would be more reliable if they were based on a verbal description. A few examples could be given.
- It would be more economical to use criteria different from those used in the matriculation examination. If the resources are limited, why test the same aspect twice? In this test the criteria should preferably refer to certain qualities in the spoken language. The only exception was Subtest 2, in which the tester wanted to compare the command of grammar and vocabulary with those in the written examination (the matriculation examination).
- The criteria would have to depend on the format and length of the subtest. In a long subtest the criteria could be more complex than in a short test.

- Since the testees had all studied English equally long, the number of levels should be relatively small. If there were only a few categories, the raters with little experience of oral testing could judge more reliably.
- It should be possible, with some practice, to rate a sample by listening to it only once. Therefore not too many criteria should be applied to one subtest. Subtest 3 with four criteria would possibly make an exception.
- For research purposes there should be variety in the criteria. The items in Subtests 2 and 5 would be assessed each at a time, while the other subtests would be judged applying the criteria to the whole test. For comparison the same quality would be rated in different subtests, and if the correlations were good, fewer criteria could be used in the future. Pronunciation would be rated twice with the same criteria, whereas fluency would be rated three times with different criteria. It was hypothesized that disfluency in a read-aloud test where the material was given would be different from that in a presentation task in which the testee might not always know what to say and how to say it. For the third time fluency would be assessed with again different criteria and under a different heading, Quality of the listening experience.

The final choice of the criteria was a compromise achieved through intuition, knowledge, listening to the samples of the pilot test, and negotiations with teachers, theorists, and the other assessors. This is how today's theorists, too, recommend the criteria to be drafted (e.g. Chalhoub-Deville 1995). The following criteria were chosen:

TABLE 3 The subtest criteria (for the details of the criteria see section 9.2)

Test	Criteria
Subtest 1, Reading a Letter Aloud	Pronunciation Fluency
Subtest 2, Interpretation	Accuracy (of grammar and vocabulary)
Subtest 3, Telling a Story	Transmitting information Pronunciation Coherence Quality of the listening experience
Subtest 4, Presenting Finnish Education	Transmitting information Fluency
Subtest 5, Reacting in Situations and Expressing Opinions	Appropriacy

As no theoretical model was to be found for weighting the criteria, the tester decided to rely on experience. Two decisions had to be made: how to weight the different criteria within a subtest, and how to weight the five subtests in proportion to one another. The first decision was made by the tester in consultation with the two other raters, whereas the second was made in the Higgs - Clifford style as an average of the suggestions of six experts (the writer and the two other raters, two experienced FL teacher educators, and a language testing professional), who were all well familiar with the test. The result is seen in Table 4.

TABLE 4 Weightings of the LLOPT subtests. The figure indicates the percentage assigned to each subtest.

Rater Test	LLOPT Rater 1	LLOPT Rater 2	LLOPT Rater 3	Teacher educator 1	Teacher educator 2	Testing expert	Mean
Subtest 1	15	20	15	10	10	10	13.3
Subtest 2	15	10	10	20	25	10	15
Subtest 3	20	20	20	20	20	20	20
Subtest 4	30	30	30	30	20	30	28.3
Subtest 5	20	20	25	20	25	30	23.3

7.3.4 The test as an instrument of change: washback validity

The concept of *washback* (British: backwash, the effect of testing on teaching) is controversial. The most divergent view is presented by Alderson and Wall (1993), who claim that backwash does not actually exist at all or, at least, it has not been proven. The reverse view is represented by West (1952) and later e.g. Shohamy (1993b, 1993c), who argue that it is wrong to blame teachers for teaching for the test: on the contrary, if the students must pass a compulsory examination, it is the teachers' duty to help them. A. Davies (1990, 31) presents similar argumentation about testing: when designing a test, the tester should also aim at beneficial washback.

In the circumstances for which the present test was being constructed, considerations about washback were particularly important. The washback is in direct proportion to the power of the test (Alderson 1993, 122; Shohamy 1993c, 17), and if any test has any significance for the participants' lives, the matriculation examination does. Pasanen (1977) asked 424 Finnish upper secondary school teachers of English what the main guideline of their work was, and 53% of the 349 who answered named the matriculation examination as the primary factor. While 28 % placed the examination as second, the most influential parameter in the language teachers' work is, no doubt, the final examination. The situation has hardly changed since the time Pasanen conducted her study. If the matriculation examination does not exert its influence overtly, for instance so that teachers use past testing papers as teaching material, its covert hold can be equally powerful and can express itself as 'the hidden curriculum' (see e.g. Ahlroos & Muilu 1989).

A good test has a *positive washback effect*, i.e. it promotes the study of meaningful contents and the use of sound teaching practices, whereas a bad test with a *negative washback effect* does the opposite. A good example of intended positive washback is the Finnish matriculation examination reform of 1977, which changed the one-sided and dated teaching of translation and grammar so that more modern practices were adopted. The negative backwash is illustrated by Prodromou (1995) when he describes teachers who make the whole lesson resemble a test, an endless interrogation, which leaves no place for different learning styles or student-centered activity. Alderson and Wall (1993, 117) also mention learner anxiety and teachers' fear of their students failing as part of negative washback. The worst distortion in today's Finnish FL teaching which has been brought about by the dated matriculation examination is the scarce attention paid to developing the oral skills in the upper secondary school. Though mastery of the oral skill is mentioned as an objective in the *Framework curriculum for the senior secondary school 1994* (p. 71), the teaching of it is to a great extent neglected. It is natural that, in the limited time available, teachers should pay more attention to fostering such skills for whose command the students are directly accountable in a public examination.

With the exception of Alderson and Wall (1993) researchers seem to agree that washback is there whether test designers want it or not (A. Davies 1990; Hughes 1989; Pasanen 1977; Shohamy 1993c, 1994; Spolsky 1993; West 1952). But washback can also be used purposefully to bring about educational change. As the commonest ways of creating educational innovation A. Davies (1990) enumerates change of curriculum, change of teaching methods and/or materials, teacher education, and testing. He - and also Shohamy (1993c) - claims that the optimum result should be achieved when all the four factors can be affected simultaneously, and it should be even dangerous to use just one to influence the others. Davies, however, emphasizes that, if only one way can be chosen, the fastest impact can be achieved by changing testing.

From the 1980s and 1990s there are examples of language testing expertise being harnessed to promote innovation in teaching. The examples of oral skills assessment come from Sweden, the Netherlands, England, Turkey, and Israel. There is also an early instance from Finland. The experiments show that work is being done at all levels: the GCSE examinations in England and the work at CITO in the Netherlands represent junior high school level, whereas the long-term project for English, French, and German at the University of Gothenburg (Lindblad 1992) serves senior high schools. The bilateral work at the University of Bogazici in Istanbul is the clearest example of an enterprise which was aimed at a reform in teaching from the very beginning: a testing expert from England was invited to plan an examination which would change the old-fashioned and ineffective language teaching at the university (Hughes 1989).

Two experiments deserve to be mentioned specially, Hirvonen's in Finland and Shohamy and her colleagues' in Israel. Hirvonen's work at the University of Turku in the early 1970s is quite remarkable (see section 6.3), and so are the reforms accomplished in Israel. Hirvonen constructed a proposal for a new matriculation examination, and his motive was the same as the present writer's: he wanted to reform language teaching in Finnish senior secondary schools through a reform of the final examination. What was remarkable in those days was the fact that his proposal contained also an oral test, which he ran on 306 school-leavers. The test consisted of a

pronunciation part (with sounds, stress, rhythm, and intonation as criteria) and a communicative part. Hirvonen's test was behavioristic, but it was notably ahead of its time. Some parts of the test appear to have had influence on the development of testing in Finland, but it was too early for the oral section. As a whole the test did not receive the attention which it would have deserved.

The value of the Israeli experiments lies on several factors: they were well-planned with the best of expertise, and they led to long-term results. They produced plenty of interesting language testing information, and they showed that the language laboratory is a good means in oral mass testing (Shohamy 1992, 1993a, 1993b, 1993c, 1994; Shohamy, Reves & Bejarano 1986; Shohamy, Shmueli & Gordon 1991; Shohamy & Stansfield 1990). In the articles mentioned above Shohamy and her coworkers have written extensively about the planned and unplanned effects of washback. In the description of the improvement of the position of the Arabic language by introducing a new test they also showed how the short-term negative effects of an examination can finally lead to a positive conclusion (Shohamy 1993b, 1993c). In the SOPI type of examination of English Shohamy and her colleagues used many efficient formats, which could particularly well serve as models for the L3 or L4 examinations in the Finnish final tests. The comparison of language samples produced by an oral interview versus a semi-direct interview in the language laboratory is valuable material in making decisions about testing.

The conclusion to be drawn from material on washback is that it is a powerful means for promoting educational change, but for optimum results it needs the support of other factors such as curriculum, material, and teacher education (A. Davies 1990). For such use of a test Frederiksen and Collins (1989, 27) have launched the concept of *systemic validity*, which they claim a test to have if it "induces in the education system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure". Improvement in the skills concerned can only be seen after the test has been in use for some time. In the present situation in Finland the other three factors referred to by Davies seem to be there: the *1994 Framework Curriculum* emphasizes speaking, the teaching materials contain more and more oral exercises, and teacher education has the needed readiness. Why should testing any longer lag behind?

In the present case considerations of washback meant, above all, that the test should be made many-sided. If the test format is the same from year to year and only a few task types are used, negative washback is only too obvious. To have positive washback, the test should assess speaking in as many forms as possible. Both transactional and interactional speech should be presented, and different genres should be offered. Qualities typical of the speaking skill should be emphasized: pronunciation, fluency (and real-time constraints), conversational strategies, and pragmatic skills. For transactional speech the students should learn clarity and organization of presentation. There should be an opportunity to present knowledge of culture. And, naturally, all of this could be implemented only to the extent that the frame of a language laboratory would allow.

8 THE LLOPT TEST AND CRITERIA

The language laboratory oral proficiency test (LLOPT) was developed in accordance with the principles outlined in the previous chapters. Summed up the main guidelines were:

- variety of functions and discourse types: both interactional and transactional speech
- simulated authenticity (to a degree feasible in the language laboratory): real-life time constraints, both speaking and understanding of speech, "real-life circumstances": reacting in everyday situations, interpreting spoken and written discourse, reading a letter aloud for classmates, presenting an institution to foreign visitors
- considering affective factors: an agreeable voice as the "interlocutor", music between the subsections, a topic with pleasant associations, beginning and ending on a task which seems easy
- considering reliability: sufficient length of the whole test and the individual parts.

8.1 The test

An overall view of the test and the criteria is presented in Table 5.

TABLE 5 The LLOPT test and criteria

Title	Content	Preparation time	Criteria		Weighting
0 Warm-up	Answering personal questions	None			
1 Reading a Letter Aloud	An American youth orchestra would like to visit Finland	1 minute 45 seconds	1. Pronunciation 2. Fluency	1-4 1-4	13,3
2 Interpreting	An American musician has arrived. The subject interprets her/his mother's talk to him.	None	Discrete point: (grammar and vocabulary)	8 x 4	15
3 Telling a Story	The subject explains a story in a Finnish paper to the visitor	3 minutes	1. Pronunciation 2. "Fluency" 3. Propositions 4. Cohesion	1-4 1-4 12 x 1 1-4	20
4 Presenting Finnish Education	The subject gives a talk to the Americans	3 minutes	1. Fluency 2. Information school system secondary school comparision	1-4 0-4 0-4 0-2	28,3
5 Reacting in Situations and Expressing Opinions	On a return visit to the US the subject has to react politely and to express opinions	None	Discrete point:	11 x 3	23,3

BACKGROUND MUSIC

Part 0 Warm-up

You are very welcome to participate in a test of spoken English in the language laboratory. First of all, we need your name. So please, when you hear the signal, give your whole name. Now answer the following questions:
How long have you been studying English?
Do you think it is an easy language?
What is your favorite subject at school?
What do you do in your sparetime?
What kind of music do you like?

BACKGROUND MUSIC

This is where the test itself actually begins. The test consists of five parts. In the first part you are asked to read aloud a letter. We have just asked about your taste in music, because the letter is about an orchestra. You are, of course, allowed to study the text beforehand. In the second part you will have to put some simple Finnish sentences into English. In Part Three you are asked to tell an American guest a story. In the fourth part you are requested to tell the

American visitors something about school and education in Finland. You can make use of a printed diagram. In the last part you are given the opportunity to show that you can behave politely in simple everyday situations. You are also asked to express your opinion about some current issues. Now we move to part one. Good luck!

Part 1 Reading a Letter Aloud

Here is a letter which has arrived at your school from America. Your headmaster has given the letter to your English teacher. She now asks you to read it aloud to the rest of the class. You can now study the letter for a couple of minutes by yourself. So take the sheets of paper given to you and study Part One. When you hear the signal, start reading it. Read as naturally as possible, not too fast and not too slow.

Cincinnati Brass Band
Clarence Mansion
2213 Twelfth Street
Cincinnati, Ohio
10785
USA

Mr. Juhani Niittylä
Varismäen lukio
Raivaajankatu 4
61390 HÄRMÄLÄ
Finland

December 23, 1992

Dear Mr. Niittylä,

We got your name and the address of your school through Mrs. Hilkka Knightsborogh, a former neighbour of yours, whose son now plays in our orchestra. We have been invited to visit Finland by Harjulahti Youth Orchestra, but as we are coming such a long way, we would also like to visit some other towns. Our visit is scheduled to take place from May 7th to 17th, and we would prefer to come to Härmälä towards the end of our stay. There are 23 musicians (aged 15-19) and five adults in our group. We would like to stay in Härmälä only one night.

My request is: Would it be possible for you, together with the municipal authorities concerned, to arrange a concert for us in your town? And could the pupils of your school put us up for the night in their homes? We do not necessarily need any food, not even breakfast. In return, maybe one day a group of your pupils could come and visit us in Cincinnati.

We would be grateful for a prompt answer, because the arrangements for the trip have to be made in the very near future.

I am looking forward to hearing from you.

Yours sincerely,

Richard Stephenson

Richard Stephenson,
Bandmaster

Part 2 Interpreting

Your school decided to invite the orchestra to come and stay with you. Your family agreed to put up one member of the orchestra. His name is Michael Smith. It is now May the 14th. The orchestra has just arrived and you have been to the school to collect Michael from the bus. Now you are entering your home. Your mother has been waiting for you to come, but she does not speak any English. So you have to interpret what she says to Mike. If you do not remember a word in English, try to convey the main idea using some other words. The main thing is that you try to say everything essential.

But first of all, you must introduce Mike to your mother and your mother to Mike. Please, do so now.

You:

Mother: **Hauska tavata. Valitettavasti en puhu englantia. Mutta tule nyt kuitenkin istumaan ja juttelemaan.** (Nice to meet you. Unfortunately I don't speak any English. But, please, come in to talk with us.)

You:

Mike: Thank you.

Mother: **Olet varmaankin väsynyt. Milloin olet lähtenyt Cincinnatista?** (You must be tired. When did you leave Cincinnati?)

You:

Mike: *I left Cincinnati eight days ago. We have been in Finland a week now.*

Mother: **Mitä haluaisit nyt syödä?** (What would you like to eat now?)

You:

Mike: *I don't want anything, thank you. We stopped at a cafeteria on our way.*

Mother: **Missä teillä on ollut konsertteja Suomessa?** (Where have you had concerts in Finland?)

You:

Mike: *We had two in Helsinki and one in Harjulahti.*

Mother: **Miten ne onnistuivat?** (How did they succeed?/Were they a success?)

You:

Mike: *Oh, they went well. We had quite a big audience and a lot of applause.*

Mother: **Mitä soitinta sä soitat?** (Which instrument do you play?)

You:

Mike: I play the saxophone.

Mother: **Mä soitin nuorempana trumpettia. Kuinka kauan sä oot soittanu?** (When I was younger, I used to play the trumpet. How long have you been playing?)

You:

Mike: *I started playing when I was eight.*

Mother: **Kuinka paljon sun täytyy harjoitella päivässä?** (How much do you have to practise per day?)

You:

Mike: *Well, it depends. Usually an hour. But if I am busy at school, there are days when I don't play at all.*

Mother: **Kello onkin jo aika paljon. Sä haluat nyt varmaan nähdä huoneesi.** (It's getting late now. You would surely like to see your room.)

You:

Mike: *That's OK with me.*

Mother: **Huomenna iltapäivällä me viedään sut katsomaan Härmälän nähtävyyksiä. Toivottavasti sä viihdyt täällä.** (Tomorrow afternoon we will take you to see the sights of Härmälä. I hope you will enjoy your stay here.)
You:
Mike: *I'm sure I will.*
That's the end of part two.

BACKGROUND MUSIC

Part 3 Telling a Story

Later that evening you are sitting in the living room with Mike. He is watching TV, and you are reading the newspaper. In the newspaper you see a story which you think is quite amusing and you decide to tell the story to Mike. You have the story printed in the sheets given to you. Now read the story to yourself and think how you would tell it in English. You have 3 minutes preparation time. When you hear the signal, start telling the story. You have about 3 minutes, in which to tell it.

THE STORY FROM HELSINGIN SANOMAT

Maailman ihmisiä

Päivänsankari hyppäsi

Floridalainen nainen juhli 86-vuotispäiväänsä hyppäämällä laskuvarjolla pienkoneesta 2 900 metrin matkan. "Kaikki pitivät minua hulluna, mutta minua ei pelottanut hiukkaakaan", **Manya Joyce** sanoi neitsythyppynsä jälkeen. "Se oli todella ihanaa."

Joyce on taitava golfin ja tenniksenpelaaja, joka keräsi hypyllä rahaa veteraaniurheilijoiden olympialaisia varten. Hän on aina ollut kovanaamainen nainen: toimiessaan rikosreportterina *Chicago Tribune* -lehdessä 1920-luvulla hän pelasi korttia rikollispomo **Al Caponen** porukan kanssa.

Joyce ei pitänyt hyppyä mitenkään vaarallisena. Hänellä oli mestarihyppääjä seuranaan seitsenminuuttisen hypyn ajan. Guinnessin ennätysten kirjan toimituksen mukaan Joyce on vanhin laskuvarjohyppääjä, jonka kirjan toimituskunta tietää.

[An English paraphrase of the story:

Parachuting on the Birthday/A Birthday Jump

A woman in Florida celebrated her 86th birthday by parachuting from a small airplane the distance of 2 900 meters. "Everybody thought I was crazy, but I wasn't a bit afraid", said **Manya Joyce** after her virgin jump. "It was gorgeous/really wonderful."

Joyce is a skillful golf and tennis player, who performed the jump to raise money for the Olympic Games of veteran sportsmen. She has always been a tough

woman: when she acted as a crime reporter for *The Chicago Tribune* in the 1920s, she played cards with **Al Capone** and his men.

Joyce did not consider the jump dangerous at all. She was accompanied by a master parachutist during her seven-minute jump. According to the editors of the *Guinness Book of Records* she is the oldest parachutist they know of.]

BACKGROUND MUSIC

Part 4 Presenting Finnish Education

Today the American orchestra are visiting your school. They are going to visit various classes, but before they go they want to hear something about the Finnish school system. You have promised to give them a general idea about going to school in Finland and particularly about going to the senior high school. Look at the printed diagram and tell your American visitors what schools we attend in Finland. Tell them also what subjects we study during the last years. Tell them what you think is different between senior high school in Finland and in the US.

The text on the handout:
Sinulla on 3 minuuttia aikaa suunnitella, mitä ja miten haluat kertoa suomalaisesta koulusta. Yritä sanoa sanottavasi selkeästi ja johdonmukaisesti. Kun kuulet äänimerkin, sinulla on 5 minuuttia aikaa sanoa sanottavasi.

1. Kerro suomalaisen koulun rakenteesta
2. Kerro lukio-opiskelusta
   - kurssimuotoisuus
   - pakollisuus/vapaaehtoisuus
   - läksyt
   - arvostelu ja ylioppilastutkinto
3. Vertaa lyhyesti amerikkalaiseen kouluun

Translation of the text on the handout:
You have three minutes time to plan what you want to tell about Finnish school and how to do it. Try to say it briefly and consequently. When you hear the signal, you have five minutes time to present your talk

1. Tell about the structure of the Finnish school system
2. Tell about studies in the upper secondary school
   - courses
   - compulsory/voluntary subjects
   - differences between Finnish and American schools

The schema of the Finnish educational system:

Ikä	Oppilaitos	
19 -	KORKEAKOULU	AMMATILLINEN OPISTO
16 - 18	LUKIO	AMMATILLINEN KOULU TAI OPISTO
13 - 15	PERUSKOULUN YLÄASTE	
7 - 12	PERUSKOULUN ALA-ASTE	
6	ESIKOULU	
1 - 5	PÄIVÄKOTI	

The schema of the Finnish educational system in English:

Age	Type of school or institute	
19 -	UNIVERSITY OR COLLEGE	VOCATIONAL INSTITUTE
16 - 18	SENIOR SECONDARY SCHOOL	VOCATIONAL SCHOOL OR INSTITUTE
13 - 15	COMPREHENSIVE SCHOOL, HIGHER LEVEL	
7 - 12	COMPREHENSIVE SCHOOL, LOWER LEVEL	
6	PREPARATORY SCHOOL	
1 - 5	KINDERGARTEN OR NURSERY SCHOOL	

BACKGROUND MUSIC

Part 5 Reacting in Situations and Expressing Opinions

A year has passed. The American orchestra had a very successful two days in Härmälä. Afterwards you and Mike kept writing to one another, and after a year you went to see Mike in Cincinnati. Now you are there, staying in Mike's home. You have been happy to notice that it has been quite easy to speak English. Of course you are trying to be as well-behaved and polite as possible. Today is quite a busy day in your life. Let us see how you cope in some simple situations.

1. *Michael's sister Doris has come home. She is quite depressed, because she has failed in an important exam. What do you say to her?*
2. *A week ago you bought a woolen pullover. Yesterday you washed it, and it shrank to half its size. Take it back to the shop and suggest what you would like them to do about it.*
3. *For some time you have not received any money from home and you have spent almost all you had. You would now like to buy an expensive camera. Ask Mr. Stephenson, Mike's father, if he could lend you 400 dollars. Explain why you would want to buy the camera (in the Unites States) and tell him when you can pay the money back.*

4. *A classmate phones and asks you to go and see a Vietnam war film tomorrow night. You are not interested, but you like the guy and would not want to hurt his feelings. Refuse politely.*
5. *You stand in a line at a supermarket cash desk. The woman in front of you has bought at least twenty items. Your bus leaves in 5 minutes. Ask the woman politely if you can go ahead of her.*
6. *You were not lucky. You missed your bus and now you are half an hour late for your friends' dinner party. Your friend opens the door for you. What do you say?*
7. *During the evening the conversation often stops. You try to help to keep it going. In the course of the evening you ask some questions to make people talk. Ask such a question now.*
8. *You tell those present about some differences between the US and Finland. The Americans ask you what you yourself think about these things. Here are three such issues. Give your opinion about them and give reasons for your opinion.*

a. *Should there be fewer compulsory subjects at school?*
b. *Should wine be sold at supermarkets?*
c. *Should Finland accept more refugees?*

9. *The evening is over. You think that it was actually quite boring, but of course you are a polite young person. How do you thank your hosts?*

The test is over, too, and it is time for us to thank you for your cooperation. We wish you every success with your future studies of English.

8.2 The criteria

When designing the criteria of the LLOPT special consideration was given to validity and reliability. Pronunciation and fluency were chosen to be the central criteria, because they represent a domain of language proficiency that is characteristic of spoken discourse. As for reliability, both the whole test and the different subsections were made so long that it was possible to get an adequate sample. Also the number of criteria in each subsection was significant. It was thought that the assessors could reliably observe only a few criteria. In the same way the number of levels was kept low. On the other hand, some criteria were used in more than one subsection to check reliability.

Part 1 Reading a Letter Aloud

In this subtest two elements, fluency and pronunciation, are assessed using the criteria below. Scales 4-1. The performance is assessed as a whole.

1.1 Pronunciation Levels 4-1

4 points Almost all individual words correctly pronounced. Sounds are unambiguous and sufficiently well articulated for easy understanding. Appropriate word-stress, stress-timing, and rhythm. Foreign accent, though still evident, does not impair understanding.

3 points Individual words may occasionally be mispronounced. Most sounds are close to those of a native speaker and sufficiently well articulated for utterances to be understood. Foreign accent in prosodic features is quite noticeable.

2 points Many individual words may be mispronounced and some individual sounds poorly articulated. L1 interference of prosodic features is very noticeable. Phonetic inaccuracy occasionally impairs understanding, listening demands some extra concentration.

1 point Wrong pronunciation of words is common. Individual sounds are often poorly articulated. L1 interference of prosodic features is quite disturbing. Many utterances are difficult to understand. Strenuous to listen to.

1.2 Fluency (reading aloud) Levels 4-1

4 points Comfortable, natural flow of speech, not too slow and not too fast. Pauses at natural junctures and functions. Presentation easy and comfortable to listen to.

3 points Flow of speech approximately natural. Some hesitation and unnatural pauses. Presentation relatively easy to listen to.

2 points Speed may be too slow or too fast. Hesitation and/or restarts. Pauses often in unnatural places. Weak syllables often too strong/stressed. The hearer is all the time conscious of having to put some effort into the listening.

1 point Speech is disjointed and halting. Speed often too slow. Frequent hesitation and/or restarts. Presentation cumbersome to listen to.

Part 2 Interpreting

The focus of assessment is accuracy, i.e. how well and error-free the Finnish idea is expressed in English. Special attention is paid to the function of asking questions. Each question is assessed separately on a 4-1 scale. The maximum score is 28.
NB. Accuracy is not the same as a word-by-word translation.

Only the following 7 questions are assessed:

1. Milloin olet lähtenyt Cincinnatista? (*When did you leave C.?*)

2. Mitä haluaisit nyt syödä? (*What would you like to eat now?*)

3. Missä teillä on ollut konsertteja Suomessa?	*(Where have you had concerts in Finland?)*
4. Miten ne onnistuivat?	*(How did they succeed?)*
5. Mitä soitinta sä soitat?	*(What instrument do you play?)*
6. Kuinka kauan olet soittanut?	*(How long have you been playing?)*
7. Kuinka kauan sun täytyy harjoitella päivässä?	*(How much a day do you have to practise?)*

4 points Correct answer or a minor mistake for instance in prepositions.

What'(d) you like to eat now? Difficult to hear whether the *d* is there or not.
How long have you played/been playing?

To gain full points, every word need not be translated as long as the meaning of the proposition does not change. Thus, in addition to the exact translation *How many hours do you have to practice a day?* also *How many hours do you practice a day?* is accepted.

3 points One mistake of grammar (or other mistake), such as may occur for instance in irregular verbs, tenses, prepositions, etc.

When did you leave from Cincinnati?
When have you left Cincinnati?

What do you like to eat?
What'd you like to have to dinner?
What do you wanna eat now?

Where have had you concerts?
How did they went?
How was the concert (singular)?

How long have you been playing saxophone?

2 points A gross mistake of grammar or two mistakes of grammar. The score is also affected by the fact how much the meaning changes. Thus *When have you left* instead of *did you leave* causes only one missing point, but *How long you are playing* instead of *...have you played* causes two missing points.

When have you leaved Cincinnati?
When did you left from Cincinnati?

Where have you play in Finland? Where have you have concerts?

What instrument do you playing?

How long you are playing?
How long you have been playing?

Rephrasing is accepted as a correction. In the following, the first phrasing would have been given only 1 point, but adding a question raised the score into 2 points:

How they was (about the concerts)*? Were you satisfied?*

1 point Questions with wrong word order AND some other mistake, for instance the auxiliary do is missing:

When you left Cincinnati? (the auxiliary missing + wrong word order)

Where you have been ...a concert... in Finland?
When you have ...been in Finland...and have a concerts?

How they go?
What you play in orchestra?

How long you are playing?
How long you are played?

How much you train one day?

More than two mistakes:
Where you have been...a concert...in Finland?
When you have...been in Finland...and have a concerts?

Part 3 Telling a Story

In this section four aspects are assessed:

1. pronunciation: how well does the candidate pronounce? Scale 4-1.
2. quality of the listening experience (= fluency): how pleasant was the speaker to listen to? The performance is assessed as a whole, scale 4-1.
3. propositions: is the proposition transmitted understandably? Each proposition is assessed separately on a 1-0 scale. The maximum score is (12 x 1 =)12.
4. cohesion of the text: how coherently does the testee use the pronoun *she/her* instead of *he/him/his* to refer to the heroine of the story? Scale 4-1.

3.1 Pronunciation

How good is the candidate's pronunciation? The scale is, with one exception, the same as was used in Subtest 1, 4-1.

4 points Almost all individual words correctly pronounced. Sounds are unambiguous and sufficiently well articulated for easy understanding. Appropriate word-stress, stress-timing, and rhythm. Foreign accent, though still evident, does not impair understanding.

3 points Individual words may occasionally be mispronounced. Most sounds are close to those of a native speaker and sufficiently well articulated for utterances to be understood. Foreign accent in prosodic features is quite noticeable.

2 points Many individual words may be mispronounced and some individual sounds poorly articulated. L1 interference of prosodic features is very noticeable. Phonetic inaccuracy occasionally impairs understanding, listening demands some extra concentration.

1 point Individual sounds are often poorly articulated. L1 interference of prosodic features is quite disturbing. Many utterances are difficult to understand. Strain to listen to.

3.2 Quality of the listening experience

What was the listening experience like? Was it enjoyable? How easy was it to follow the story? How easy was the testee to understand? (Was it possible for the listener to listen in a relaxed way?) Scale 4-1.

4 points The speech is natural, perhaps even vivid, it is easy and pleasant to follow. It is, however, a little slower than if the testee speaks freely, because here s/he has to think of the facts. If the speaker does not know a word, she is, however, able to make herself understood. Sometimes the speaker may also appeal to the listener for assistance. The pauses are at appropriate places and of appropriate length.

3 points The speech is relatively easy to follow. There are similar characteristics as in the 4-point speech, but to a smaller extent.

2 points There are difficulties in following the speech. Several mistakes of grammar or, for instance, self-coined words make the listening experience more strenuous. There are similar characteristics as in the 1-point speech, but to a smaller extent.

1 point The speech is difficult to follow, the listener has to make an effort, and the story may seem incoherent. When uncertain, the speaker may mumble her/his words. Pauses are often too long, often also at the wrong places. The vocabulary is

insufficient, and the round-about expressions do not always make the meaning clear. Faulty pronunciation may also hamper understanding.

3.3 Propositions

Which of the following propositions were transmitted understandably? This criterion might also be called Communicative effectiveness. Scale 1-0. Will be 50 % of the final score of Subtest 3.

THE PROPOSITIONS

1. Who?	an 86-year old lady
2. Did what?	parachuted from a small plane
3. Where?	in Florida
4. How high?	2900 m
5. How long?	7 minutes
6. With whom?	a master parachutist
7. Why did she do it?	o raise money
8. How did she feel?	a. wonderful
	b. not afraid
9. What sort of person ?	a. a (criminal) reporter
	b. a sportswoman/a golf and tennis player
10. What is remarkable?	a. the oldest parachutist The Guinness Book of Records knows of
	12 propositions

3.4 Coherence of the text

How cohesively does the testee use the pronoun *she/her* instead of *he/him/his* to refer to the heroine of the story? Scales 4-1.

4 points The pronoun *she* used cohesively all the time.

3 points An occasional slip, or a couple of successive ones.

2 points Several masculine pronouns.

1 point Incohesive use of the pronouns *he/she* makes listening quite strenuous.

Part 4 Presenting Finnish Education

In this subtest two aspects will be evaluated: fluency in transactional text and transmitting information or the propositions. Fluency, 4.1, was evaluated on a 4-1 scale. In the transmission of information (4.1) the range and accuracy of information are important. To be able to give a correct description of the Finnish school system, the testee must have sufficient vocabulary. Each of the three themes, 4.2.1 the Finnish

school system, scale 4-0, 4.2.2 the Finnish senior secondary school, scale 4-0 , and 4.2.3 comparison between Finnish schools and American schools, scale 2-0, was assessed separately.

4.1 Fluency (transactional speech)

4 points Natural, comfortable speed and tempo in most contexts. Occasional groping, rephrasing, and circumlocution may occur. Pleasant, easy to listen to.

3 points Natural hesitation while organizing thoughts and some hesitation while searching for language. This may sometimes interfere with the speed of delivery but does not interrupt the general flow of language. The hearer is all the time conscious of listening to a non-native.

2 points Coherence maintained though not a constant flow of language. Hesitation while searching for language is noticeable but does not demand unreasonable patience from the listener.

1 point Hesitation demands considerable patience from the listener. Words may come one by one. Utterances often incomplete and restricted in length with long unfilled pauses between them.

4.2 Transmitting information

In the same three parts of this section similar skills were required, only the amount of information was smaller in the third part.

4.2.1 Presenting the Finnish school system

How many-sided and accurate is the picture which the description gives of the Finnish school system? In the assessment attention will be paid also to the range and accuracy of the vocabulary. Scale 4-0.

4 points The description gives - considering the shortness of time - a rather good general idea of the Finnish school system. Some term may be wrong or missing (e.g. vocational school), but most terms are correct. The linguistic quality of the speech is good, and it is easy to understand.

3 points The description gives some kind of general idea of the Finnish school system. There are some mistakes in the terms, and the attempts at circumlocution do not always succeed in giving the right picture. Most of the concepts presented in the figure are, however, transmitted in an understandable way. In addition to terms, there may be other linguistic shortcomings, but on the whole the speech is, nevertheless, understandable.

2 points The description gives some information about the Finnish school system, and some of the concepts in the figure are correctly rendered. Also a longer description only gains two points if it is linguistically quite deficient.

1 point There is some attempt at a description. The performance is short, or contains many mistakes. Some things are, however, rendered correctly.

0 point No attempt or nothing right.

4.2.2 Presenting the Senior Secondary School

Scale 4-0, like in 4.2.1. In this part, however, even a 4-point performance is often shorter than the descriptions in the previous section.

4 points In the performance some central features of the senior secondary school are mentioned. Each of the four subthemes is touched on, generally in an understandable and correct way, though not comprehensively.

3 points One of the four subthemes may not be touched on. The information given is reasonably correct, and, in spite of the potential linguistic shortcomings, mostly easy to understand.

2 points Of the subthemes given only one (extensively) or two are dealt with. The listener gets some information about the senior secondary school. Even a longer presentation only receives two points if it is linguistically very defective.

1 point There is some attempt to present the senior secondary school, or, the presentation is linguistically so defective that the speaker is unable to convey his meaning.

0 point No attempt or no correct information.

4.2.3 Comparing Finnish and American Schools

Scale 2-0.

2 points Two or more elements compared correctly.

1 point One correct comparison

0 point No correct comparison

Part 5 Reacting in Situations and Expressing Opinions

In this subtest both linguistic accuracy, including pronunciation, and sociopragmatic appropriacy are assessed. Each situation or expression of opinion is assessed separately. Scale 3-0, maximum (11x3=)33.

3 points Vocabulary, grammar, and pronunciation good, some incidental mistake is permitted. Intonation in harmony with the semantic content. Sociopragmatically appropriate, not forgetting the linguistic politeness words such as please, excuse me, and I'm sorry. The degree of politeness is in harmony with the importance of the message. In expressing an opinion, also justification is given.

If the answer is good, but not completed in the time given, it can still gain full points (sometimes too little time was allowed).

2 points There are mistakes on the linguistic side or one major single mistake (like confusing *lend* and *borrow*), OR the answer is sociopragmatically unsatisfactory. The testee has slightly misunderstood the question.

1 point There are mistakes on BOTH the linguistic AND the sociopragmatic side. Or, the answer is not understandable to a foreigner, like building the answer on the word *alko* (the Finnish alcohol monopoly). Part of the answer is indistinct mumbling.

0 point The speaker has obviously not understood the question, or her speech seems otherwise irrelevant.

Instruction for different situations:

In for instance the following cases one point was deducted:

Situation 1 Sympathy was offered quite curtly.

Situation 2 Too straightforward, of the type *I want my money back.*

Situation 3 There is no mention about paying the money back.
The request is presented as a matter of course, as if the person were borrowing only 10 dollars.

Situation 4 No suggestion of any compensatory activity.
Refusal without any explanation. Even the word *sorry* makes the refusal more acceptable. A good answer is of the type *Sorry, I can't make it today. But what about some other day?*

Situation 5 Employs an accusing tone as if the woman ahead had too many purchases.

Situation 6 The apology is too wordy with lots of explanation. (On the other hand, some explanation is necessary. It is not acceptable, either, to pass over the fact of being late as a matter of course.)
Putting the blame on the woman ahead in the line.

Situation 7 Showing boredom and/or suggesting a compensatory activity.
The question *What do you do at these parties?* easily transmits the

impression as if conversation alone were not a sufficient pastime.
Egocentricity/ethnocentricity, of the type *What do you know about Finland?*
If the testee directs her speech to only one person, e.g. to an imaginary person sitting next to her, no point is deducted.

Situation 8 Exaggerated politeness.
Suggestion of coming again.
Inviting people to visit the American host family.

9 THE EXPERIMENT

The experiment was carried out in two schools in January and December 1993. In both schools the experiment involved two tests, the ACTFL oral proficiency interview and the language laboratory oral proficiency test (the LLOPT). The principles of designing the LLOPT were explained in Chapter 7, and the structure and the criteria of the LLOPT were presented in Chapter 8. The subjects, the ACTFL interview, and the test arrangements will be described in the present chapter.

9.1 The subjects

Because the present test was designed to be used in the future with students who would have explicitly studied the speaking skill, it seemed reasonable also now to try to find students who had had some special teaching in the oral skill. With few exceptions, the teaching of the speaking skill in an ordinary Finnish upper secondary school has so far been a rather neglected area. If the students of such schools had been tested for the speaking skill, both the tester and the testees would have been left with a sense of frustration. Fortunately, the National School Board had started an experimentation on the teaching and testing of the speaking skill in 1990, and the students of the participating schools had come to their final year by 1993.

The fact that only nine schools had participated in the experiment, and even in these particular schools only a limited number of students took part, made the choice of the subjects easy. Though the number of the potential schools was thus restricted, it was natural to try to find as different subjects as possible and, therefore, to choose more than one school. Two were chosen: *Halikon lukio* (a senior secondary school in a medium-sized south-western municipality) and *Joensuun normaalikoulun lukio* (the Senior Secondary Practice School at the University of Joensuu). In the text from now on they will be called School 1 and School 2 respectively. In addition to the difference of school type, there were some other dissimilarities: Joensuun normaalikoulu is an urban school with above average students (the average comprehensive school

leaving grade of the participating students: 8.83 on a scale of 4-10; the result of the written part of the English matriculation examination 3.86/5), whereas Halikon lukio is a rural district school with average students (the average comprehensive school leaving grade of the participating students was, however, rather high, 8.59/10; the written part of the English matriculation examination 3.40/5). Joensuu lies in the eastern part of Finland, whose Carelian people are claimed to be more open and voluble than the people of the south-western part of Finland, where Halikko is situated.

In both schools two classes had participated in the three-year experiment. For part of the first year the students had been divided into smaller groups, and at the beginning one of the two classes in Joensuu had had 36 extra suggestopaedically oriented hours of teaching. The students involved in the English experiment had, at the same time, participated in the National School Board oral experimentation in another foreign language, Swedish, and some of them also in the German and/or French and/or Russian experiment, which had naturally augmented their overall oral skills. It was voluntary for the students to take part in the oral test, but they were encouraged to do so and they knew that they would receive a certificate for participating. 89% of the students involved in the oral experiment took part also in the oral test of English. Those who did not, were absent from school during one or both of the testing days or for a longer period. Among the 25 Halikko students there were 10 boys and 15 girls, in Joensuu 10 boys and 25 girls, 35 students altogether. Of the Halikko students 12% had spent a longer time (45 days or more) in an English speaking country, 28% 8-45 days. In Joensuu 11% of the students had paid a longer than 45-day visit, 26 % a 8-45-day visit.

9.2 The ACTFL oral proficiency interview

The ACTFL oral proficiency interview (or the ACTFL OPI) is a holistic measuring instrument widely used to assess oral proficiency in a number of foreign languages. It can be used to measure all ranges of speaking ability from the very beginners to students with native-like proficiency. The ACTFL interview is individually adapted to the needs and skills of each test-taker, but to ensure reliability it is standardized for the conversational procedure to follow a prescribed pattern.

9.2.1 The rating scale

Although the ACTFL assessment makes use of many criteria, the rating scale is holistic with a global description of each level (Byrnes 1989). It is based on the ACTFL view of language proficiency, which is customarily depicted as an inverted pyramid (Figure 9).

The pyramid is divided into four sections representing the four major levels of language performance: Novice, Intermediate, Advanced, and Superior. The narrow end of the pyramid shows how little command is needed to perform the simplest tasks, while the upper surface is left open to illustrate that the highest level of

proficiency has no ceiling but finally merges into native proficiency, which is never 'complete'.

At the *Novice Level*, the speaker can communicate minimally and mainly in a reactive way. She uses memorized material, single words or formulaic expressions, and can function only in the most common contexts. At the *Intermediate Level*, the interviewee is able to create with the language, to combine and recombine learned material, but can still function in quite predictable situations only. She can initiate, minimally sustain, and close a communicative task, and ask and answer simple questions. The ability to perform these tasks signifies a discourse shift from word level to sentence level.

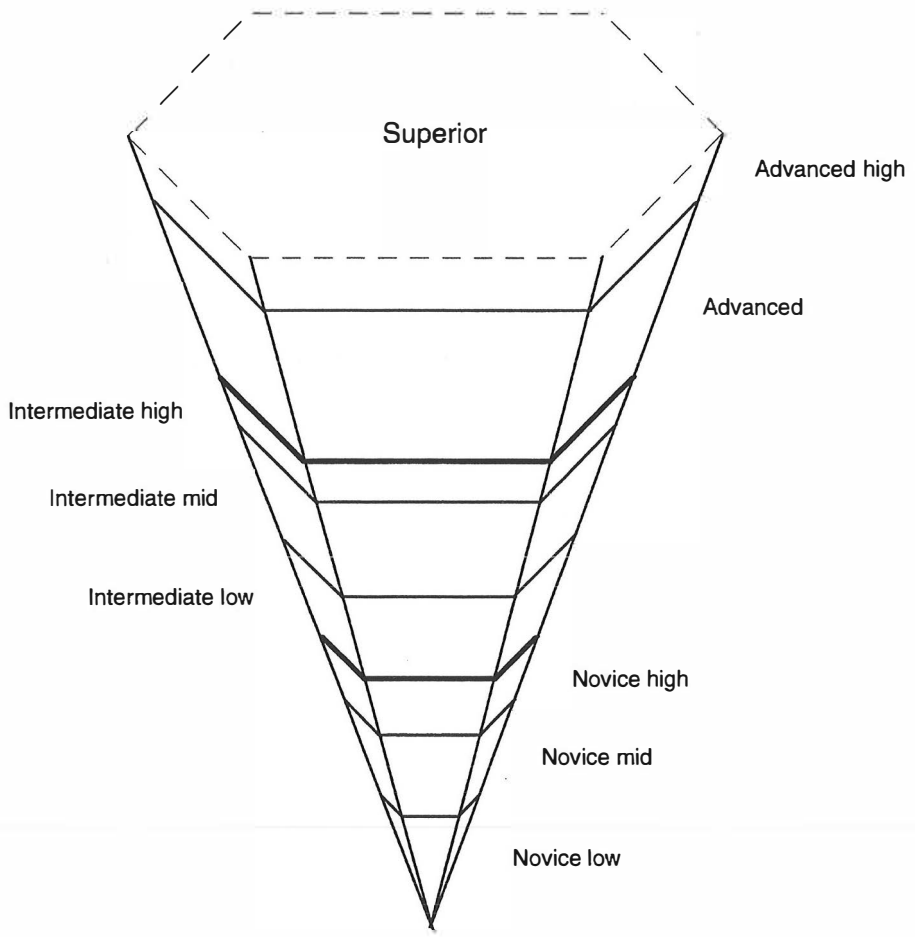

FIGURE 9 Inverted pyramid representing fhe ACTFL major ranges and sublevels of language proficiency (from Buck 1989, Illustration 2-C)

An *Advanced Level* speaker's skills are quite versatile. Her conversational proficiency is no longer mainly reactive but she is a fully participatory partner who is able to initiate, sustain, and bring to a close a wide variety of communicative tasks. She can satisfy the requirements of general school and work situations and is able to

cope with also an unforeseen turn of events. At the Advanced level, the interviewee can describe and narrate, which is only possible with the command of a paragraph type of discourse.

At the *Superior Level* the speaker's scope stretches as far as the native or bilingual proficiency, and even beyond that of the uneducated native. To be described as a Superior Level speaker a person has to be able to discuss a broad range of topics in depth by supporting opinions and hypothesizing about abstract issues. Her command of vocabulary and discourse strategies has to be native-like. However, she need not have native-like pronunciation, neither does she have to be as swift in shifting register or using cultural allusions as the original speakers of the language.

Only little is so far known about the development of interlanguage, and it has not been shown that it should advance by leaps from one major level to the next. It is only gradually that the learner begins to acquire the skills which are characteristic of the next stage. How is the cut-off set? The ACTFL guidelines rate the interviewee as belonging to the next major level/category if she can satisfy its requirements for more than half of the time. The major levels are, in turn, divided into sublevels: the Novice Level and the Intermediate Level into three, and the Advanced Level into two. The superior level is undivided. The borders between the sublevels are not as clear-cut as the thresholds between the major levels.

The sublevels included, the ACTFL interview scale consists of ten levels. It is obvious that if the candidates in a test have all studied the language for the same number of years and passed the previous tests, which is the situation at the school-leaving stage, their results will not be spread all over the scale but concentrate on some section. On the basis of the course in Siuntio (section 9.2.4) and the pilot tests, the cut-off point for passing the present test was placed between Novice Mid and Novice High. It seemed justifiable that no one who had studied English for more than nine years should be passed at less than Novice High. There was a temptation to set the cut-off even one step higher up, but as this was the first 'official' oral school leaving test ever, there was place for some lenience. At the upper end no distinction was made beyond the Advanced Level. This did not mean that the potential existence of better candidates was excluded, but there was no need and probably not sufficient testing proficiency to draw the line higher. In an FL matriculation examination even an educated native speaker only scores the highest mark available.

9.2.2 Assessment criteria

The ACTFL oral interview is a holistic instrument, in which the rating scale is divided into different levels. The level descriptions are, accordingly, global characterizations of the integrated performance, but for the benefit of the user, first and foremost the rater, the ACTFL oral interview assessment criteria offer a detailed description of each of the enabling skills. The factors that mainly guide the assessment are four: the functions or global tasks the test-takers have to perform, the context and content in which they have to move, the accuracy they are capable of showing, and the type of oral language they are able to produce.

The ACTFL term *global task* or *function* refers to the testee's ability to use the language, to perform speech acts with the language. It covers more or less the same area as, for instance, Bachman's term functional competence (Bachman 1991a). As

examples of the lowest level mention can be made of such simple tasks as listing or enumerating, whereas the other end is represented by for instance a well-structured argument. If the testee's speech is flawed, the tester is not so much concerned with her mistakes of grammar or vocabulary as with the instances of communicative failure.

The word *context* is used in its ordinary sense signifying the circumstances or settings in which a person uses language. In a setting where it is possible to function on the basis of a more or less settled script, like in a restaurant, one can come off with lower level skills. The more unpredictable the circumstances, the greater are the demands on proficiency. In different contexts the same topics can be discussed at different depth, whereas the number of *contents (topics)* increases exponentially with an increase in proficiency.

In the assessment criteria, the concept *accuracy* refers to the "acceptability, quality, and precision of the message" and covers a wide range of skills: fluency, grammar, pragmatic competence, pronunciation, sociolinguistic competence, and vocabulary. The ACTFL coverage and distribution of the terms differ, to some extent, from those used in other testing literature. When for instance *fluency* is used to signify also the cohesive devices, it comes close to Canale's (1983) textual competence and Bachman's (1991a) textual knowledge. *Grammar* (= usage of norms of morphology and syntax), *pronunciation* (= ability to reproduce segmental and suprasegmental features of the language), and *vocabulary* (= size of lexicon and adherence to norms of usage) are used in the traditional way, but *pragmatic competence* in the sense of "ability to use various discourse management devices to get the message across and to compensate for imperfect control of the language" is similar to Canale and Swain's strategic competence. The definition of *sociolinguistic competence*, finally, does not make a distinction between appropriacy of function and appropriacy of form (cf. e.g. Canale 1983), but speaks generally of the "ability to use language appropriately in different registers in various situations within a particular culture, and to use cultural references and idioms."

At the lower end of the scale "accuracy", from Novice Low till about Intermediate Mid, the speaker's command of grammar, vocabulary, and the rest of the enabling skills is often so poor that she can only be understood by a listener who is accustomed to that particular foreign accent. At this stage the responsibility for conducting the negotiation of meaning lies mainly with the listener. As the skills increase towards the upper levels of proficiency, quite elaborated skills are needed, and the responsibility for the success of the negotiation is more evenly shared.

The increase in accuracy is a quantitative as well as qualitative change. To be able to express herself more precisely, the learner needs to know and use more and more items of grammar, vocabulary, and culture. In the same way the learner's ability to handle longer and longer stretches of discourse increases. The Novice speaker is only able to produce individual words and phrases, but gradually she learns how to handle discrete sentences and later paragraphs, till she reaches the Superior Level, where even extended discourse is mastered. The *command of text types* or 'the quantity and the organizational aspects of speech produced by the interviewee' is also one of the ACTFL assessment criteria.

9.2.3 The structure of the interview

The ACTFL interview is a dynamic procedure, in which everything is interdependent, and the pattern changes all the time. In addition to linguistic and evaluative factors, also the psychological parameter has to be considered. The testee should have the feeling of participating in any natural conversation, yet the ACTFL interview is highly structured. The interviewee is taken through four phases: the warm-up, the level checks, the probes, and the role play. (Liskin-Gasparro 1989.)

The goal of the *warm-up phase* is to establish the interviewee in the FL conversational situation. The interaction consists of pleasant, small talk type of dialogue, which is meant to put the interviewee at ease and to give her time to get used to the foreign language in general and the interviewer's way of speaking in particular. To the interviewer this phase gives a chance to form a preliminary concept of the candidate's level and to find some topics for the future conversation.

The function of the *level check phase* is to find the interviewee's performance *floor*, that is the level at which the candidate is comfortable with the language, and is able to handle the functions and contexts with confidence and accuracy. The interviewer may have formed a hypothesis of the interviewee's level during the previous phase and now seeks confirmation. If she finds it with one topic, it is important to repeat the procedure with a variety of others, which proves that the candidate can sustain the level.

Once the floor has been established, also the *ceiling*, the level where the interviewee can no longer communicate without effort, must be found. For that purpose, the interviewer makes a probe into the next level trying to find out whether the interviewee can still function at that level. If the result is a breakdown, the interviewer will probably have another try with another topic, but if that fails, too, the interviewee's ceiling has been established. If the probes prove successful, the higher level is the new floor, and in order to find a new ceiling the same procedure will be repeated.

A means to confirm the decision about the level is the *role play*. An interview test with its obviously skewed power relations does not easily give the testee a chance to naturally perform all the functions mentioned in the ACTFL guidelines. That is why there are role cards suitable for three levels: for the Intermediate Level there are cards which make the testee ask questions or initiate, sustain, and close a simple situation, at the Advanced Level there is a situation with a complication, and the Superior Level candidate is asked to manage a linguistically unfamiliar situation. *The ACTFL Tester Training Manual* (Buck 1989) includes a package of role cards, but because many of them seemed rather unsuitable for the present testing situation, a new set of role cards depicting situations which a Finnish student might come across (14 Intermediate, 11 Advanced, 13 Superior; see Appendix 1) was created for the Finnish school-leavers.

If the role play has confirmed the tester's previous assessment, the test is brought back to easy everyday conversation and soon wound down. The interviewee should be left with a pleasant feeling of a successfully completed task. If the role play does not provide evidence for the tester's hypothesis of level, the interview must still be continued for some time. The time needed for the total test varies according to the level of the interviewee. To judge a Novice candidate does not take an experienced

interviewer more than ten minutes. An Intermediate candidate requires 12 to 15 minutes (an Intermediate High a little longer), while Advanced and Superior ones take between 20 and 25 minutes. If the candidate is particularly shy or reserved, a longer time may be needed.

9.2.4 Tester training

Carrying out and evaluating an ACTFL interview is a demanding task. The tester must have a certified Superior Level proficiency herself. As an interviewer and assessor, she needs three basic skills: to be able to use efficient elicitation techniques, to structure the interview adeptly, and to rate it in a reliable way. For each of the skills, the ACTFL has a rating scale from 0 to 3. To earn the highest points for elicitation, the interviewer has to create a situation which makes it possible for the candidate to reach her best performance and for the evaluators to assign an accurate rating. The interviewer must show genuine interest and friendliness and, at the same time, maintain a neutral attitude. While making the candidate feel comfortable in a natural conversation, she must, unnoticed, guide her through a highly structured interview. (Buck 1989.)

However interesting some candidates' history and opinions may be, the interviewer must not be carried away by the contents, for attending to the various phases of the interview demands full concentration. A good interviewer knows how to use her time optimally: she uses warm-up, role play, and final wind-down effectively and is able to place checks and probes in a relevant way. Every question must have a purpose also from the structural point of view. (Buck 1989.)

The ACTFL interviews are regularly taped, and every interviewer usually rates her own interviews. In addition to the eliciting and structuring skills, the interviewer must, accordingly, also be a reliable rater. It is no wonder, then, that it requires a long and thorough training for anyone to become a certified ACTFL tester. If the would-be interviewer is not a native speaker of the language, she must first of all show Superior Level language proficiency herself.

The training of the interviewers takes place in several phases. It is begun with a four- to five-day intensive workshop, during which each trainee has a chance to observe about twenty interviews and to conduct and rate several supervised ones herself. Later on she conducts first ten, then fifteen interviews autonomously and sends the tapes to the ACTFL to be rated. Those who receive a certificate have to renew it every two years. (Liskin-Gasparro 1985, 39-40.)

In Finland the first ACTFL oral proficiency testing seminar was arranged at Siuntio in August 1991 by the University of Helsinki and the America Center. It lasted five days and was conducted by two American ACTFL trainers. The present writer attended the seminar and a three-day extension course one year later. In between she conducted 10 interviews on her own and sent them to the ACTFL for appraisal. This training was, however, less than the regular ACTFL oral interview tester training.

9.3 Test arrangements

Since Joensuun normaalikoulu had started the National School Board experiment one year earlier than the rest of the schools, the test could be held in the spring term, in January 1993. The Halikko students were tested in December 1993, in a week somewhat unsuitable for them, for the six-week period with English had only started after a long period of no English.

Of the two tests, the ACTFL oral proficiency interview and the language laboratory test, the ACTFL was carried out first at both schools. The fact that the tester was a stranger and the format of the test partly unknown was considered to be a source of anxiety, which would be somewhat reduced if the socially easier test was taken first and the testees had an opportunity to make personal acquaintance of the tester. The testees' oral comment indicated that the decision was evidently right.

A period of 30 minutes was reserved for each interview. Because the interviews lasted 20 minutes on average, the interviewer had some preparation and concentration time between two candidates. The maximum number of interviews per day was nine. Although there was a one-hour lunch-break in the middle of the day, the tester's and perhaps also the testees' fatigue was clearly noticeable in the afternoon. This fact may have resulted in decreased reliability (see further 8.3.1).

In both schools the testing was carried out during five days from Monday to Friday. The language laboratory test took place on the last day in each school. In Halikko the students were tested in two sets and in Joensuu in three. For assistance and eventual emergency there were always two testers present, one operating the machine, the other dealing out handouts and checking the arrangements.

10 RESULTS

At the beginning of the study four research questions were asked concerning the development of oral tests. Below they will be dealt with each at a time.

10.1 The language laboratory test (LLOPT) as a test format

The main research task of the present study was to find out to what extent senior secondary school students' oral English proficiency can be tested validly, reliably, and efficiently using a language laboratory test. This question has to be answered at two stages: to say something about the psychometric qualities of the test, it is necessary to study its results. Thus the figures of the different parts of the test will be presented first. At the second stage a closer scrutiny will be made of the significance of these figures and of the psychometric qualities of the test.

10.1.1 The results

In the language laboratory test a student's total result could only be seen after all the subtest scores had been added up. To avoid the halo effect and to assure comparability, the samples were assessed one subtest at a time. When the raw scores were added up, the maximum total was 107 points and the mean 77.8. To compare the level reached in each subtest and subskill, the means are also indicated as percentages of the maximum (Table 5). Below an outline is first given of the results of the five subtests, and this information is compared with other language proficiency results. Different subskills, pronunciation, fluency, cohesion, transmitting information, reacting in situations, and expressing opinions are then studied separately.

The subtests The test had been planned so that the first subtest should have been the easiest and the tasks should then have become more difficult, culminating in Subtest 4. The latter part of the plan was implemented, but Subtest 1, Reading Aloud, proved

more difficult than had been expected. The easiest part turned out to be Subtest 2, Interpreting (everyday speech). The students were also especially successful in Subtest 3, Telling a Story, i.e. at conveying simple facts which were given in Finnish, whereas the other transmitting information task, Subtest 4, in which both content and form had to be created by the student, proved quite difficult. The latter task involved great cognitive loading; another explanation may be the fact that the test format was new to the students. However, also Reading Aloud proved relatively difficult, and yet it is an exercise which is very common at school.

Subtest 4, which was the most difficult part, also discriminated most among the testees. A discriminating factor was both the amount of information and the fluency of presentation. In Subtest 5, Reacting in Situations and Expressing Opinions, there was another discriminating section, the opinions.

TABLE 6 The results of the LLOPT

	Subtest	Mean	Maximum	Standard deviation	Percentage of maximum
1	Reading Aloud	5.44	8	1.42	68
	pronunciation	2.63	4	.78	66
	fluency	2.80	4	.71	70
2	Interpreting	21.61	28	5.5	77
3	Telling a Story	18.24	24	3.60	76
	pronunciation	2.97	4	.72	74
	propositions	9.80	12	1.83	82
	quality of listening	2.72	4	.86	68
	cohesion	2.73	4	1.16	68
4	Presenting Finnish Education	8.54	14	3.05	66
	fluency	2.81	4	1.01	70
	presenting education	2.62	4	1.05	66
	presenting senior high	2.58	4	1.11	65
	comparing the systems	1.01	2	.79	50
	total information	6.21	10	2.1	62
5	Situations and Opinions	23.83	33	7.34	74
	situations	17.68	24	4.96	75
	opinions	6.18	9	2.70	69
6	Total score of the LLOPT	77.8	107	18.67	73

Comparison of the LLOPT and other results The results of the LLOPT subtests were also compared with the ACTFL OPI and various other indicators of the students' language proficiency. To facilitate comparison, all results were converted into percentages of the maximum (Figure 10). From the point of view of the present study the most interesting issue was to find out how the LLOPT and the ACTFL OPI compared. It is noteworthy that although the LLOPT and the ACTFL OPI may have assessed different aspects of the language, the total results (73% versus 72%) were very close. Similarly, the grade of English in the school final report and the matriculation examination were quite the same (74%), and the teacher's grade and the students' grade very similar to each other (67% versus 66%), but both the teacher and the students had underestimated the students' oral skills. The distribution of the

scores of the different subtests in the matriculation examination may tell something about the impact of the three-year oral experiment: there was a high score in listening comprehension (an oral skill) as well as in the essay (a productive skill), while reading comprehension, the receptive skill in the written domain, and the grammar skill showed lower scores.

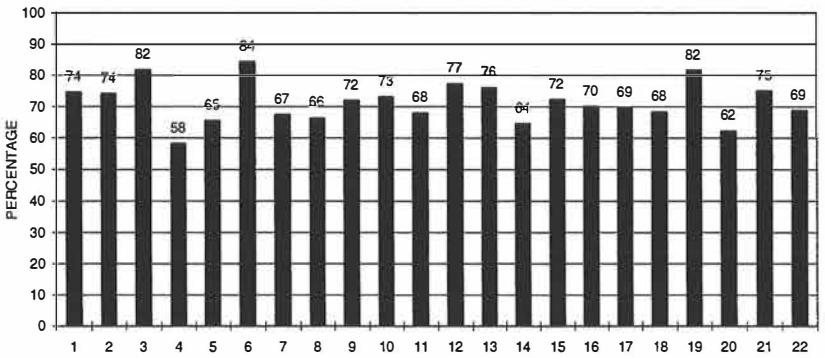

FIGURE 10 The levels reached in the different skills and subtests shown as percentages of the maximum

Column 1: Grade of English in the senior secondary school final report
Column 2: Grade of English in the matriculation examination
Column 3: Score in the matriculation examination listening comprehension
Column 4: Score in the matriculation examination reading comprehension
Column 5: Score in the matriculation examination grammar (and vocabulary) test
Column 6: Score in the matriculation examination essay test
Column 7: Teacher's grade for oral proficiency
Column 8: Student's grade for own oral proficiency
Column 9: ACTFL mean
Column 10: LLOPT mean
Column 11: LLOPT Reading Aloud
Column 12: LLOPT Interpreting (grammar)
Column 13: LLOPT Telling a Story (interpreting written text)
Column 14: Presenting Finnish Education
Column 15: Reacting in Situations and Expressing Opinions
Column 16: Pronunciation
Column 17: Fluency
Column 18: Cohesion: gender
Column 19: Transmitting information 1
Column 20: Transmitting information 2
Column 21: Reacting in Situations
Column 22: Presenting Opinions

The effect of gender Previous researchers of Finnish school-children's oral proficiency have shown significant differences between the two genders. Kristiansen (1990), who tested grade 6 elementary school L2 English, and Pasanen and Hietanen (1994), who tested junior secondary school grade 9, found that female subjects were significantly better than male subjects in both comprehension and production. Of Huttunen and Kukkonen's (1995) results in grade 6 some showed significantly better success for females than for males. Nevertheless, with age the results seem to even out. In the present study (see Figure 11) the differences in the various parts of the matriculation

examination were so small that the total grade for the examination was the same. In the two oral tests, the ACTFL OPI and the LLOPT, however, the female subjects were superior to the males. The only oral part in which there was no difference was the listening comprehension section in the matriculation examination. This even result may partly be explained by the test method factor: teachers claim that males are more skillful at handling the multiple-choice techniques.

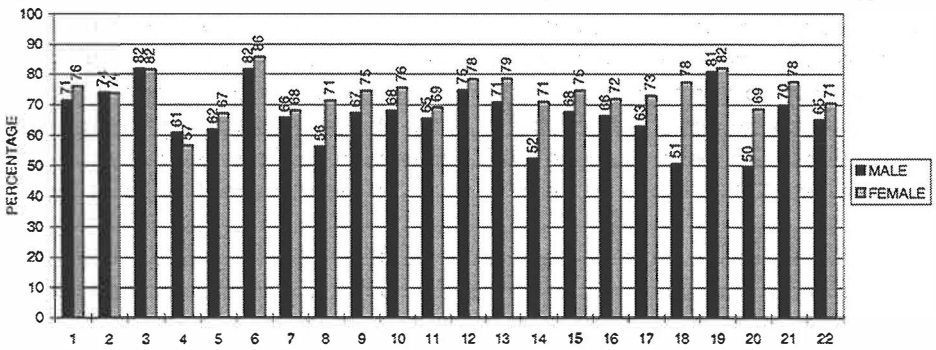

FIGURE 11 The levels reached by boys and girls shown as percentages of the maximum. For the contents of the columns see the text in Figure 10.

Pronunciation

Both pronunciation and fluency were tested in more than one subtest to see how much difference there was in different types of discourse (see Chapter 8). Pronunciation was tested in Subtests 1 and 3, and it was thought that Reading Aloud a Letter would be the easiest. However, the result of 66% was well below the total score 70.6% (weighted score). One explanation for the low result may be the fact that some of the words were unknown to the weaker students. The fact that as many as 10 students scored the maximum of 4 points indicates that the best students either knew practically all the words or were sufficiently familiar with the English rules of pronunciation to infer, for instance, where the stress should be placed in an unknown word.

In Subtest 1 the students were asked to read also the unknown words, whereas in Subtest 3, Telling a Story, they could choose the expressions that they used for the reproduction of the given content. It was hypothesized that this difference would show in the results. The criteria for the two tests were also designed on a different basis. If, however, the results of the two tests should be close to one another, it would be possible to draw the conclusion that one test of pronunciation would be sufficient. The difference between the two tests was considerable, 66 % in Subtest 1 and 74 % in Subtest 3, but in a direction that had not been expected. As the correlation between the tests was merely moderate (Table 12), it is likely that they measure the same thing only to a certain extent. The Reading Aloud Test can be expected to measure also the student's vocabulary size. It is also possible that there are some weaker students who are able to pronounce familiar material well, but to whom many words and

expressions are unfamiliar. If only pronunciation is tested, reading aloud a new text which also incorporates a number of unknown words is not the best form. But in most cases an integrated test is efficient and economic, which makes reading aloud quite suitable to be used along with other pronunciation tests. However, a shorter version of the present test ought to give the needed information.

The two schools At the end of the comprehensive school the students in school 2 had been significantly better in English than the students in school 1 (see 9.1), and they were also better at the end of the senior secondary school. Nevertheless, the means of the pronunciation tests, 70 percent of the maximum in both schools, show that this was a skill in which they were on the same level. School 1 had put more effort in practicing pronunciation than school 2.

The gender. In pronunciation the girls were superior to the boys (72% against 66%).

Fluency

The results in fluency were even one percentage unit lower than those in pronunciation. One might wonder whether the popular communicative approach pays sufficient attention to developing these skills. Also fluency was tested in two tests: Reading Aloud and Presenting Finnish Education. The two tests were chosen because of the five subtests the former was believed to be the easiest part and the latter the most difficult. The first hypothesis, however, proved to be wrong. And really: why should it have been easy to read aloud fluently a text where some words are met with for the first time? The very similar score for both tests (2,80/4 and 2,81/4, 70%) shows that there were disfluencies in both parts, but the causes might have been different. In Reading Aloud there were two primary causes: either the text had not been analyzed correctly and the pauses were in unnatural places, or there were unnecessary pauses because the speaker stopped to consider the pronunciation of the next word. In Presenting Finnish Education the speaker stopped in the wrong places, because she was either thinking of what to say or how to say it.

The criteria for the two fluency tests were the same. Fluency was also tested in a third test: in Subtest 3, Telling a Story, there was a criterion called the quality of the listening experience, and from the beginning it was designed as another criterion of fluency. In agreement with Sajavaara and Lehtonen (1980), who had maintained that fluency is something that is ultimately experienced by the listener, the present tester had wished to design and try out criteria that would emphasize the role of the listener. In this case there was a slight difference: with the result of 2.72/4, 68%, fluency in Telling a Story was 2 percentage points lower than in the two other tests. The difference can be explained in two ways: One alternative is that there may have been a real difference in fluency in the different subtests, and the figures would have been the same even when measured with one and the same description. The other alternative is that the two criterion descriptions depict a different thing. There was also another qualitative difference: the listener-oriented criteria were found easier to use. However, the actual difference in the different genres of text was so small that it would seem justifiable to assess fluency with only one subtest.

Grammar

The only part of the LLOPT where grammar and, to an even smaller extent, vocabulary were separately explored was Subtest 2, Interpreting. The only aspect of grammar that was studied was asking questions. In theory it should be a skill completely mastered by students who had studied English for more than nine years, but in practice this was not the case. It is true that the percentage, 77, was higher than that of any other subtest, but it did discriminate considerably with a standard deviation of 5.5 (mean 21.6, Table 6).

A special question concerning this subtest was how well it would correlate with the part of the matriculation examination in which grammar and vocabulary were assessed. The underlying hypothesis was that if the rules of grammar had not been automated, there might be a discrepancy so that in an oral part with time constraints even the average students would make more mistakes than in a written test of grammar. In this test, however, the average students seemed reasonably well able to construct correct questions. The fact that the correlation between Subtest 2 and the grammar part of the matriculation examination was only .57 (Table 13), lower than the correlations with some other parts of the matriculation examination (essay .76, LC .64) would, nevertheless, seem to support the hypothesis, but for a more reliable answer this question would need further investigation.

Cohesion: gender

When the LLOPT was first planned, the intention was to assess as many factors of communicative competence as possible. To test discourse competence a subtest with long turns would be suitable. In the LLOPT the tester tried to analyze textual organization in Subtest 4, Presenting Finnish Education, but in spite of several efforts it proved too difficult to compile satisfactory criteria, and the attempt was given up. One reason for the failure may be the fact that the material was already organized in the instructions. Secondly, the assessors felt that it was strenuous to pay attention to more than the two criteria: fluency and the relevance and adequacy of content.

The criterion "cohesion" was applied only in Subtest 3, Telling a Story, and used in a very limited sense referring merely to cohesion in the use of the pronouns *he/his/him* and *she/her* respectively. When speaking Indo-European languages Finns often make a mistake of gender in the third person singular, because the Finnish language does not make this distinction but has only the pronoun *hän* to refer to both sexes. Although a mistake of gender is not uncommon, the assessors were struck by the fact that in Subtest 3, when describing the adventures of Miss Manya Joyce, only 37% of the candidates managed without any mistakes of gender. When the mistake is repeated, it may be felt as both irritating and confusing, as the following excerpt from Student 48's text may show:

> Hi, listen this, Mike. This is good story. It is about a birthday hero who jumped. A woman from Florida...*he* celebrated *her* birthday. *He* was eighty-six years old then, and *he* celebrated the day by jumping from an aeroplane, and the aeroplane flew...proximately about two thousand and nine hundred meters above the sea level. "Everybody thought I was crazy, but I wasn't scared a bit", Manya Joyce said after *his* first jump. "It was really great".
>
> Joyce is a very skillful tennis-player and also a golf-player and *he* rais...by this jump *he* raised some money for the Olympics of veteran athletes. *He* has always...*she* has always

been a tough woman. As *he* was working as a crime reporter in Chicago Tribune in a nineteen-twentieth century *he* played cards with the Mob leader's Al Capone's gang.

Joyce didn't think that jump was dangerous at all, *he*...*he* did...*he* had...*she* had a master champion with *him* during the whole seven-minute jump. According to Guinness Book of Records Joyce is the oldest skydiver who...that has jumped from aeroplane.

The frequency of the error is also contradictory to Pasanen and Hietanen's results (1994, 46), according to which mistakes of gender were very rare in the nation-wide junior secondary school composition test taken at the age of 16. One explanation may be the time constraints in a spoken genre. However, after getting this research result, the present writer has paid special attention to Finnish people's use of gender in speaking English and discovered how very common it is that even the most competent speakers, including professors of English, may have an occasional lapse in this respect. For this reason cohesion as defined here may not be a valid criterion after all. Its low correlation with the other tests may also point to this conclusion.

Transmitting information

Subtest 3 and Subtest 4 tested transactional speech. The former had been planned to be the easier one, because in it all the information was given, though in Finnish. So it proved to be. Subtest 3 was the task in which the subjects received the highest percentage, 82, transmitting 9.8 propositions on an average out of 12. It is true that narration is easier than many other genres, but considering the fact that Miss Joyce's undertakings were both numerous and colorful and, accordingly, demanded either a varied vocabulary or subtle communication strategies, the candidates' achievement in this task can be described as commendable.

The other transactional task, Presenting Finnish Education, was intended to be the most demanding task of all and have a good discrimination capacity. Also this expectation materialized. The percentage, 66, was the lowest of all, and the standard deviation was 3.05 (mean 8.54). Two less persevering candidates in School 2 even gave up at this point. Total unfamiliarity with the test format may have been an explanation.

The three parts of Subtest 4, Presenting the School System, Telling about the Senior Secondary School, and Comparing the American and the Finnish School, required all both knowledge of facts and communication skills to present it with. The difference of scores between the first and the second task was small, but the third task was the most demanding of all: the score was only 1.01/2 and the percentage 50. One comparison gave one point, and for the full score only two were needed, but even that was too much for most. The fact that knowledge of culture was assessed in a test of English was obviously new to the students.

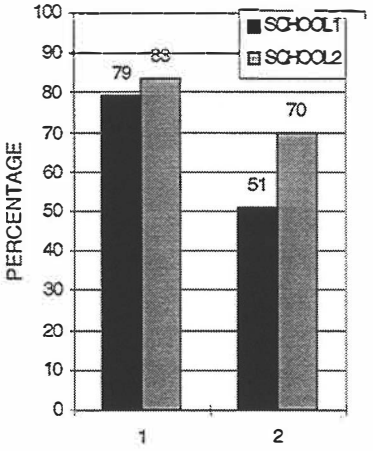

FIGURE 12 The results of Transmitting Information 1 (Subtest 3) and 2 (Subtest 4) by school (School 1 = Halikko, School 2 = Joensuu)

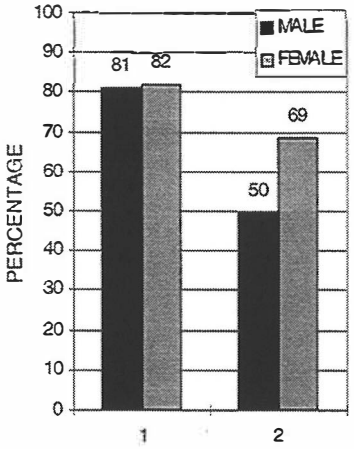

FIGURE 13 The results of Transmitting Information 1 (Subtest 3) and 2 (Subtest 4) by gender

The two tests that assessed transmitting information serve as examples of an easy and a difficult subtest. In the easy test - telling the story about the old parachutist - the standard deviation was small and the difference between the schools and the genders was also small (see Figures 12 and 13). In the difficult test - Presenting Finnish Education - the standard deviation was great and the difference between the schools and the genders greater than in any other subtest (the difference was increased by the fact that two males in School 1 did not complete the Transmitting Information 2 -section at all). This difference can be seen as an indication of the fact that in the two tasks called Transmitting Information two quite different cognitive processes were involved.

Reacting in situations and expressing opinions

Subtest 5 was designed to assess interactional and sociolinguistic skills. The test-takers were to react in different situations that they would encounter when returning their American friends' visit. To give the students a special challenge, a conversation was included in which they were asked to express an opinion on three matters. Because the matriculation examination is supposed to be a test of maturity, it was considered appropriate to include a task which gave the students a possibility to operate in the genre of argumentation. The opinions served this target well and proved to be a discriminating task with a mean of 6.18/9 (69%) and a standard deviation of as high as 2.70.

The situations varied in difficulty (for means and deviations see Table 7). The easiest one was number 3, in which the testee had to turn down a friend's invitation to go to the cinema. A blunt refusal was not awarded with a full score, but some other alternative had to be suggested. Other easy tasks were thanking for the evening and apologizing for having arrived late. In these items wordiness or excessive politeness

were penalized with a loss of points. The most difficult task was the first of the opinions, perhaps only because it was a new type of task and the students did not quite know what to do. There were difficulties also in suggesting a new topic after a silence in conversation. Some of the suggestions seemed to imply boredom (What do you usually do at these parties?), which incurred a light penalty.

TABLE 7 Reacting in Situations and Expressing Opinions (Subtest 5)

	Mean (max. 3)	Standard deviation
Situation 1: Sympathizing	2.18	1.04
Situation 2: Returning a purchase	2.06	.90
Situation 3: Asking for a loan	2.25	.81
Situation 4: Refusing an invitation	2.39	.82
Situation 5: Asking to jump the line	2.26	.86
Situation 6: Apologizing for coming late	2.29	.75
Situation 7: Opening a new topic for conversation	1.91	1.08
Opinion 1: Compulsory subjects at school?	1.87	1.12
Opinion 2: Alcohol at supermarkets?	2.13	1.07
Opinion 3: Number of refugees in Finland?	2.18	1.07
Situation 8: Thanking for the evening	2.31	.82
Whole Subtest 5	2.17	.94

Judging sociopragmatic appropriacy is a complex matter, in which assessors may interpret a tone of intonation or a nuance in the choice of words quite differently. The assessors discussed examples from the pilot tests and had many meetings before they came to an understanding of the main principles. In the end the agreement was very high with correlations between .89-.94 for the opinions and .88-.91 for the situations.

10.1.2 The psychometric qualities of the LLOPT

The test design of the LLOPT aimed at maximizing efficiency, reliability, and validity. How far this succeeded can be judged by looking at the results obtained from the test and by comparing them with the principles that it was based on.

Reliability

The great advantage of the laboratory test as compared with an interview is the *uniformity of elicitation*. The interview is an individual test, while in a language laboratory test the instructions, requests, and questions are the same to all test-takers. This signifies a great increase in reliability. Below first the elicitation and then the rating will be described and discussed.

In both the ACTFL interviews and the language laboratory test there were problems with *technology*. In the LLOPT the problems were of two kinds: those concerning the length of pauses and those concerning recording. The manufacturer had assured that the tape-recorder would automatically start recording the students' speech, and for the pilot test it was decided to let the machine take care of the pauses.

However, the system did not work in the intended way. The tape-recorder started the recording too late, and the beginning of each turn was lost. In the test proper the pauses were, therefore, controlled manually, which may have meant that they were not all equally long for each group of testees. The other breakdown of technology was even more serious: many days after the test it was discovered that a test-taker's tape-recorder had not functioned and his tape was completely empty. This had happened although a technician from the language laboratory manufacturer's had checked that everything was functioning perfectly on the test morning.

When compiling the *criteria*, the present researcher had tried to make use of her long experience in practical language testing. For dependable assessment, the capacity of the working memory had to be considered. The principle of making the scales long enough to secure maximum discrimination and short enough for the human mind to still discern accurately had been clear from the beginning, but what this meant in practice had to be found by experiment in cooperation with the other assessors. To make the *rating* as reliable as possible the pilot test samples were used for training. When the first version of the criteria had been created, the three assessors gave their marks independently and compared the results. If there was disagreement, the raters discussed the difference and listened to the tapes together. When the pilot tapes had been agreed on, the work could proceed to assess the test proper. The subtests were rated one at a time to avoid a potential halo effect.

TABLE 8 The internal consistency (Cronbach alpha) of the LLOPT subtests (**p<.01, *p<.05)

Subtest		Internal consistency
1. Reading a Letter Aloud,	pronunciation	.90**
	fluency	.84**
2. Interpreting a Dialogue,	grammar	.97**
3. Telling a Story,	pronunciation	.80**
	propositions	.92**
	quality of listening	.86**
	cohesion	.93**
4. Presenting Finnish Education,	fluency	.91**
	presenting education	.93**
	presenting senior high	.93**
	comparing the systems	.94**
5. Reacting in Situations and Expressing Opinions		.98**

To make the LLOPT as reliable as possible it had been designed to be a long test with several subtests. The use of different test formats was meant to decrease test method effects (Bachman 1990). The goal of reliability was well achieved as judged by internal consistencies. In the study of the reliability of the subtests an interesting aspect emerged. Though the test was designed to be a direct test, so that the final

score should be based on the assessors' subjective judgment instead of counting, some subtests, or rather criteria, appeared to be more judgmental than others (on assessment by counting versus judgment see Pollitt 1991). It seems that in assessment there is no distinct dividing line between the quantitative and the qualitative. When judging the qualitative tests the score is also often based on the frequency of error versus correct form. Quantifiers such as *hardly any/some/frequent/ many errors/ mistakes* are common in the criterion descriptions. In this test the most clearly countable part was Subtest 2, in which the correctness of grammar was assessed. The propositions (how many?) and cohesion (how many mistakes?) in Subtest 3 and the information (how much?) in Subtest 4 belong to the same category of the more countable. Such criteria as pronunciation, fluency, and the quality of listening belong to the other category, that of the less countable. (Subtest 5 is difficult to definitely place in any of the two categories.) The latter category is less easy to assess than the former. This division is also visible in Table 8, in which the less countable criteria are represented by a range of .80-.91 and the more countable by .92-.97.

The same division between the more or less countable categories is seen if the interrater reliability coefficients are compared. The overall interrater correlation for the LLOPT was very high: .96 between the raters A and B, .96 between A and C, and .97 between B and C. For the subtests and the different criteria the figures were somewhat lower (Table 9). There were not enough resources for testing intrarater reliability, but the comparison of the correlations in the two pronunciation estimates (Table 12) and the three fluency estimates (Table 12) can, to some extent, show also intrarater consistency in pronunciation and fluency assessment. The tables show that the interrater and intrarater correlations are moderate, but they are higher than the correlations with other criteria such as cohesion and information, which are not shown here. As for the work

TABLE 9 The range of correlation coefficients between the Raters A, B, and C in the LLOPT subtests (**p<.01, *p<.05)

Subtest or criterion	Range of interrater correlation coefficients
1. Reading Aloud a Letter	.79 - .79**
2. Interpreting a Dialogue	.91 - .94**
3. Telling a Story	.79 - .87**
4.PresentingFinnish Education	.78 - .86**
5. Situations and Opinions	.92 - .95**
Pronunciation, Subtest 1	.74 - .77**
Pronunciation, Subtest 3	.48 - .70**
Fluency, Subtest 1	.62 - .67**
Quality of Listening, Subtest 3	.62 - .79**
Fluency, Subtest 4	.75 - .80**
Propositions, Subtest 3	.76 - .82**
Cohesion, Subtest 3	.78 - .88**
Information, Subtest 4	.88 - .90**
Situations, Subtest 5	.88 - .91**
Opinions, Subtest 5	.89 - .94**

of the three raters, Tables 18 and 19 (Appendix 4) show that Rater B was on a different track from the other raters in pronunciation assessment (Subtest 3) and Rater C in fluency (Subtest 1). These low figures show also in Table 8. The most reliable rater seems to have been Rater A. The two raters would have needed more practice in pronunciation and fluency assessment.

The correlations within a subtest are somewhat higher than those between the subtests, which may indicate that the subtests are, under the same/different name, tapping a slightly different quality.

On the whole, it can be said that the language laboratory test has reliability figures that, for the most part, are sufficiently high for high-stakes testing. The technical infallibility of language laboratories is, however, not yet hundred per cent sure.

Validity

In designing the LLOPT an attempt was made to concentrate as much as possible on the *a priori* aspects of validation (cf. Weir 1990). It was believed that if the *a priori* validation is neglected, the *a posteriori* validation would not be of great use. For *construct validation* the test was derived from the concept of communicative competence and the nature of spoken language, and in the operationalization of the construct the principles of a communicative test were applied (see section 4.2.1). An attempt was made to present the tasks in a realistic context and to formulate them so that they should be relevant to the test-takers.

The extensive content domain analysis showed how diverse the area of oral proficiency is. To give sufficient samples of such a rich domain, different test formats and different scales were used. However, although it was possible to make the candidates process the language in real time, the main weakness of a language laboratory test seemed to have remained: having a tape-recorder as an interlocutor is not real interaction with genuine negotiation of meaning. But to what extent is any test situation ever authentic interaction, as Alderson points out (1981, 48; for the concept of authenticity see Shohamy & Reves 1985; van Lier 1996; Widdowson 1979). However, in the language laboratory the test-takers seemed really engulfed in conversation. Particularly in Subtest 5, Reacting in Situations and Expressing Opinions, they really set themselves into comforting, complaining, excusing, and persuading. The language laboratory was full of vigor and enthusiasm.

If the LLOPT is considered a proficiency test rather than an achievement test, *content validity* is of less importance. The national senior secondary school curriculum is so vague that not even criterion descriptions could be based on it. However, the many senior secondary school teachers who, either through the National Board of Education experiment (see The Introduction) or in inservice training, have come into contact with the test have attested that it agrees with the skills and contents taught in the senior secondary school.

The study was originally constructed so that the ACTFL OPI should be used as a *concurrent validity* criterion of the LLOPT, but in the course of the research program it became, however, clear that the ACTFL OPI also had its imperfections and its construct of oral proficiency was only partly shared by that of the LLOPT. (see Figure 1; for critique of the ACTFL concept of proficiency see 6.1). Accordingly, it was no

longer desirable that the ACTFL OPI and the LLOPT should have a hundred per cent correlation. After this discovery the ACTFL OPI still remained a criterion of concurrent validation, but no longer a *sine qua non*. The question rose as to what would have been the optimal correlation of the two tests. Both are extensive oral tests and therefore there should have been substantial common coverage, but, on the other hand, the LLOPT had been created in the hope that it should perhaps measure features that an interview did not tap. The test results (Table 11) showed that the correlation varied from .64 to .88 (disregarding coherence, the odd criterion), which was quite close to what one might have wished it to be. Sociolinguistic competence (Subtest 5), pronunciation, and particularly fluency had high correlations with the ACTFL, whereas the tests assessing the transmission of information had lower correlation figures.

In addition to the ACTFL OPI there was other information about the students' English proficiency against which the LLOPT could be compared: the grade of English in the matriculation examination, the grade of English in the school final report, the teacher's grade of the student's oral proficiency, and the student's own grade of it. The two grades that, in theory, could have had perfect correlation with the LLOPT were the teacher's and the student's assessments. The grade in the school final report, which covered also the speaking skill, should have correlated more than the matriculation examination, in which the only oral part was listening comprehension. On the other hand, the school final report may, to some extent, also reflect such characteristics as cooperation and active interest, which are not directly part of language proficiency. Table 11 shows that the teacher's grade of oral proficiency has about the same correlation figures as the matriculation examination and the school final report, whereas the student's grade has conspicuously lower figures. This result seems to indicate that it is not easy for teachers or, even less, for students to assess oral proficiency without a comprehensive test. However, the correlation might have been higher if the students had been assessed by the same or similar measuring instruments in their ordinary work at school. Now they were tested for instance on personal opinions and long narration and presentation, which had not been used by the teachers.

Another way of estimating the need for an oral test, and also its validity, is to see what grade the students at different levels of the matriculation examination would have got if the examination had been oral (= the LLOPT). The LLOPT scores were divided into six categories equivalent of those in the matriculation examination as shown in Table 10. Less than half of the students would have received the same grade in both tests, 27% would have received a higher grade in the LLOPT, and another 27% would have received a lower grade. Twenty-five percent would have achieved one grade higher and only one student two grades higher. Of those who performed less well in the oral test, 22% would have lost one grade, 3% two grades, and one student as many as three grades. This was a boy who gave up trying when he had come as far as Subtest 4, Presenting Finnish Education. However, also in the ACTFL oral test his performance was below average. On the whole it can be said that the best students were good also in the oral test, while there was more variation among the weaker performers. In the LLOPT three students would have failed, while everybody passed in the matriculation examination.

TABLE 10 Grades achieved in the matriculation examination compared with the LLOPT grades

	Grade in the language laboratory test						
Grade in matriculation examination	87-107 (l)	76-86 (m)	65-75 (c)	54-64 (b)	43-53 (a)	0-42 (i)	Row total
laudatur (l)	18	5					23
magna cum laude (m)	5	4	3		1		13
cum laude (c)	1	4	3	1	1		10
lubenter (b)			4	2	2	1	9
approbatur (a)				2	1	2	5
improbatur (i)							
Column total	24	13	10	5	5	3	60

For the further development of the LLOPT it was also interesting to see how well the different subtests correlated with the other measures, although here again it was difficult to tell what the optimum correlation should be. If a subtest had a very high correlation with an existing test, it would signify that the new test would be superfluous. A low correlation, on the other hand, could signify two quite opposite things: it could, for one thing, mean that it did not measure language proficiency at all or, on the other hand, that it tapped something - maybe important - that was not measured by any other instrument. The fact that the criteria that assessed transmitting information had the lowest correlation figures with the other subtests and criteria probably shows that it is a skill which is not measured by any other test.

TABLE 11 The correlations of the LLOPT with some other measures of (oral) proficiency (** p <.01, * p <.05)

Subtest	Teacher's grade	Student's grade	ACTFL OPI	Matriculation examination	School final report
1. Reading Aloud	.78 **	.57 **	.83 **	.78 **	.80 **
pronunciation	.75 **	.53 **	.78 **	.76 **	.76 **
fluency	.75 **	.55 **	.81 **	.73 **	.76 **
2. Interpreting	.72 **	.67 **	.80 **	.81 **	.78 **
3. Telling a Story	.78 **	.65 **	.80 **	.79 **	.79 **
pronunciation	.77 **	.53 **	.78 **	.74 **	.74 **
propositions	.62 **	.55 **	.64 **	.57 **	.68 **
quality of listening	.83 **	.61 **	.81 **	.81 **	.78 **
cohesion	.33 *	.33 *	.35 **	.34 **	.45 **
4. Presenting Finnish Education	.51 **	.65 **	.66 **	.52 **	.62 **
fluency	.64 **	.66 **	.75 **	.58 **	.65 **
total information	.49 **	.64 **	.60 **	.45 **	.57 **
5. Situations and Opinions	.72 **	.68 **	.78 **	.71 **	.72 **
LLOPT total	.79 **	.76 **	.88 **	.80 **	.83 **

Face validity Although face validity is not a central psychometric quality, it is an important motivating factor for the testees. In this experiment the students had an opportunity to express their views about the test in a questionnaire. For this part the questionnaire was unstructured so that the students could write freely. The reaction was overwhelmingly positive: *mukava, hyvä, kiva, nasta, ihan yes* were adjectives that the students used to express the fact that they had liked the test. The Finnish equivalents to *practical, relevant, real-life language* were expressions which showed that the test-takers had experienced the test as authentic. There was one boy who commented on all parts as being *boring (tylsä)* and a girl who found everything *difficult*, but apart from that negative comments were very rare.

It appeared that the five parts had been experienced in the way that the tester had intended. Subtest 1 and particularly 2 had been perceived as easy (mentioned by 27 students), Subtest 3 also as easy but somewhat more difficult, Subtest 4 as difficult, and Subtest 5 also as difficult but by equally many as easy. Some students pointed out that it was a good idea to place reading aloud first, because "you could do it without thinking" and "it gave you a feeling of knowing and strength" (*osaaminen ja vahvuus*). As many as 18 students remarked that Subtest 1, Reading Aloud, was a good test of pronunciation.

It was interesting to see how an easy test, such as Subtest 2, was also perceived as a pleasant experience. Twenty-seven students (n=60) mentioned that it was easy, and thirty that it was *good, practical, pleasant, useful* or something like that, and the only fact that was criticized (by three students) was that you might forget what you had to interpret. However, also when a subtest was experienced as difficult, like Subtest 4, some students could admit that they were themselves to blame: their own vocabulary was insufficient, or that the test was, after all, useful. In Subtest 4 the students had to know some facts and terms and plan and organize their message. Though it was considered difficult, the students did not really complain. Perhaps it was obvious to them that they should be able to explain something about their daily life and surroundings in English. Because the matriculation examination is a test of maturity, the students can be expected to know something about culture, too.

The most controversial part was Subtest 5, which provoked the most comments. It really seemed to divide the testees. The great majority described it as *relevant, practical, really useful, many-sided, sensible, cool*, and so on, but there were a few who were of the opposite opinion. They commented on the situations as being *unnatural* and *unrealistic*. This was the only test in which one could perhaps see differences between the two regions: most negative comments came from the west. Perhaps it was easier for the lively Carelians to think of something proper to say, whereas some western students complained of too short a time to answer. "Opinions are easy for those who have opinions","it is difficult to form an opinion quickly", "even in Finnish I wouldn't have known what to say". The last type of comment was also given in Subtest 4, Presenting Finnish Education. However, not a single student offered the commonplace comment "Shouldn't this be a language test (and not one of thinking and opinions)?". Mere L2 vocabulary and grammar are not enough if there are no thoughts to be expressed. Besides, students want realistic tests, and in real life it is not uncommon that when the appropriate comment comes to the mind, the moment to voice it is already gone.

Many individual comments on the details of the test were useful for its further development. When the test was over, students in Joensuu came to the tester quite excited and asked whether the matriculation listening comprehension test could be made on these lines. These comments can be seen as expressions of the face validity, but they may also reflect some more essential validity. If today's students are supposed to be the legitimate experts of their own learning, would it not be possible to think that they are the legitimate experts of their own testing, too?

The washback validity The washback validity of the LLOPT can only be speculated about. It is obvious that a multiform test produces more diversified practice in the classroom than a single test format. If the basic elements of the LLOPT are used in the future final examination, one could wish that skills like good pronunciation and fluency, sociolinguistic appropriacy, and presentation would be practiced in schools. In the case of a language laboratory test one could wish that introducing it as a final test would also contribute to the increased use of the language laboratory as a multipurpose FL learning instrument, a valuable opportunity that has been greatly neglected.

Efficiency and usability

In a cost efficient test the quantity and quality of the information produced is maximum in comparison to the cost. Mass testing is costly, but ultimately the actual expenses depend on the required quality of the results. What will the intended oral proficiency test be used for? Does it have to have the same reliability as, for instance, the matriculation examination, or is the oral test there only to make the teaching of speaking more credible? For the latter purpose a more modest test will suffice, but even in this case the test-takers must be able to rely on the test being fair. If important decisions are based on the results, a longer, multiform examination is needed. To cite Hughes: "Accurate information does not come cheaply" (1989, 37). In addition, the role of the matriculation examination is growing, because the recommendation of the Ministry of Education is that universities should more and more base their intake on the results of the matriculation examination.

In any test both material and manpower costs are involved. It is apparent that a language laboratory test is more time and cost efficient than a face-to-face test even in the case of more elementary tests where the tester can give the grade immediately. The software cost of the LLOPT is comparable to that of the ACTFL OPI except for the fact that the LLOPT cannot so far be recorded on the video. However, both video- and audiotapes are reusable. The main material cost is, of course, the investment in the AAC-type language laboratories. To a layman in technology it would seem natural that if the test is carried out as part of a school final examination, it has to take place simultaneously all over the country. Even though the test-takers can be tested one group after another, which is the case at present in the matriculation listening comprehension test, there is a limit to how long the participants can be kept enclosed in a room waiting for their turn. If the test lasts 45 minutes plus 15 minutes for the arrangements, and it is estimated that the testees can wait three hours, the number of laboratory booths needed in a municipality is the number of the testees divided by four. This would mean a considerable expense to communities lacking the sufficient

equipment. However, professional testers claim that modern tehcnology could offer other possibilities.

In addition to the material costs there are the human expenses. The labor cost and the need for expertise in the LLOPT are, however, much smaller than the expense entailed by the interview. Real expertise is needed for creating a new version of the test twice a year, but this involves only a few people, probably no more than are involved in the present matriculation examination. The cost of teacher training would be reasonable.

The rating costs depend on the length and complexity of the sample. To assure reliability and validity, the present version of the LLOPT was made both long and multiform. An essential issue in the further development of the test would be research into the question how long a sample would still be reliable and how many subtests and criteria would be needed. It is possible to gain some understanding of the matter by looking at the figures in Table 12, which shows the correlations between various subtests and criteria. The figures show how closely to one another the different pronunciation and fluency tests measured a quality. Though the correlations within a subtest are higher than those between the subtests, even the latter are sufficiently high to justify the testing of both pronunciation and fluency in only one subtest. Table 12 shows which parts seem to measure the same quality/factor.

TABLE 12 The internal correlations of the LLOPT subtests and criteria. The majority of the figures have a correlation of $p<.01$. Those marked with * have a correlation of $p<.05$, and in the two figures in brackets the correlation is not significant.

Subtest/Criterion	1	1.1	1.2	2	3	3.1	3.2	3.3	3.4	4	4.1	4.2	5	5.1	5.2
1. Reading Aloud															
1.1. pronunciation	.96														
1.2. fluency	.95	.82													
2. Interpreting	.80	.75	.77												
3. Telling a story	.83	.78	.80	.79											
3.1. pronunciation	.87	.81	.84	.79	.87										
3.2. propositions	.59	.53	.59	.65	.75	.59									
3.3. quality of listening	.81	.77	.79	.74	.90	.82	.58								
3.4. cohesion	.41	.40	.38	.38	.69	.40	.27*	.52							
4. Presenting Finnish Education	.60	.57	.58	.59	.64	.65	.48	.58	.35						
4.1. fluency	.67	.59	.70	.62	.69	.66	.50	.67	.41	.85					
4.2. total information	.53	.51	.49	.53	.56	.59	.43	.50	.30*	.98	.72				
5. Situations and opinions	.70	.67	.68	.68	.68	.69	.59	.69	(.23)	.81	.80	.75			
5.1. situations	.61	.57	.59	.62	.62	.61	.52	.63	(.15)	.78	.76	.72	.97		
5.2. opinions	.77	.72	.75	.71	.71	.73	.63	.70	.34	.78	.78	.70	.92	.81	
6. LLOPT total	.86	.81	.83	.85	.85	.85	.68	.83	.45	.87	.85	.81	.91	.85	.91

The criterion that has the smallest correlation with the other parts is cohesion of gender (3.4). Other parts that have a smaller correlation with the rest are those having to do with transactional language: 3.2 propositions, 4.1 fluency of transmitting information, and 4.2 (the amount of) information. The common factor in these three parts is the significance of vocabulary. This result seems to indicate that an oral proficiency test should include one section which measures transmitting information.

On the other hand, the mutual correlations of the pronunciation and fluency criteria are so good that only one measurement of each would suffice. The criteria in the reading aloud task correlate well with the others so that this easy to design and easy to assess test type is usable.

Correlating the LLOPT subtests and criteria with the present matriculation examination and its different parts shows to what extent the existing test and the language laboratory test measure the same qualities (Table 13).

The part of the matriculation examination that correlates best with the oral test is the essay. This may be explained by the fact that both are productive tests. However, it is remarkable that even reading aloud should correlate so well with it. If economizing is aimed at, interpreting could be left out, for it does not seem to measure anything that would not be assessed in the written test. However, as a test form it is authentic and practical, and it could well be used if it were made more difficult and/or judged by other criteria. The correlation of subtests 4 and 5 with the matriculation examination is moderate, which could be interpreted to testify that these tests assess language proficiency but perhaps features that are not tested in the written test.

TABLE 13 Correlations of the LLOPT with the aural and written parts of the matriculation examination (** p< .01, * p< .05)

Subtest or criterion	Matriculation examination				
	Total	LC	RC	Grammar	Essay
1. Reading a Letter Aloud	.78 **	.64 **	.55 **	.45 **	.77**
pronunciation	.75 **	.58 **	.61 **	.37 **	.72 **
fluency	.73 **	.64 **	.44 **	.49 **	.75 **
2. Interpreting	.81 **	.64 **	.42 **	.57 **	.75 **
3. Telling a Story	.79 **	.69 **	.46 **	.51 **	.78 **
pronunciation	.74 **	.60 **	.49 **	.44 **	.74 **
propositions	.68 **	.61 **	.37 **	.46 **	.59 **
quality of listening	.81 **	.73 **	.50 **	.48 **	.70 **
cohesion	.34 **	.28 *	.12	.29 *	.43 **
4. Presenting Finnish Education	.52 **	.47 **	.10	.54 **	.61 **
fluency	.58 **	.56 **	.20 **	.48 **	.63 **
total information	.45 **	.44 **	.06	.52 **	.56 **
5. Situations and Opinions	.71 **	.68 **	.34 **	.53 **	.67 **
situations	.51**	.51 **	.25 *	.39 **	.48 **
opinions	.67 **	.58 **	.34 *	.50 **	.67 **
LLOPT total	.80 **	.71 **	.39 **	.60 **	.80 **

Another viewpoint on estimating how the subtests predict the overall score in the test was acquired by using *regression analysis*. Several analyses were run, using both stepwise selection of the predictors and entering the variables in a predetermined manner.

Stepwise regression analysis uses both forward and backward selection for entry in the regression equation. In forward selection, the first variable considered for entry is the one with the largest positive (or negative) correlation with the dependent variable. If the first variable selected for entry meets a statistical test, the same

procedure is applied to the second candidate for inclusion, and so forth. When forward selection enters variables in the equation, backward selection eliminates potential predictors on similar grounds.

At first the LLOPT sum total was used as the dependent variable. This introduces, of course, technical correlation since subtests also contribute to the sum total. The purpose of the analysis is not so much to estimate the exact explanatory power of different subtests as to find out their *relative* explanatory power. The predictors were the sum scores for the five subtests: Reading Aloud (Subtest 1), Interpreting (Subtest 2), Telling a Story (Subtest 3), Presenting Finnish Education (Subtest 4), Situations and Opinions (Subtest 5). It will be remembered that the criteria were somewhat different in the different subtests. The results of the analysis are to be seen in Table 14.

TABLE 14 Results of the stepwise regression analysis with the LLOPT sum total as the dependent variable

Model	Variables Entered	Removed	R	R Square	Adjusted R Square	Standard Error of the Estimate
1	Test 5	-	.923	.852	.849	7,2463
2	Test 2	-	.984	.967	.966	3,4332
3	Test 3	-	.993	.987	.986	2,2104
4	Test 4	-	.999	.998	.998	0,7837
5	Test 1	-	1.000	1.000	1.000	4,60E-07

The table shows that Test 5, a test that consists of eight situational responses and three opinions and is scored in terms of accuracy and appropriacy on a scale of 0-3, was the best predictor of the overall score. It explained about 85% of the variance in the final score. Test 2, a test where the situation is one of simulated interpretation for a non-English mother, interpreting her 10 questions in concrete terms, was the second best predictor. It added about 10 percentage points to the explanatory power and raised it to about 96%. The rest of the test, for obvious reasons, made only negligible contributions to the prediction. The importance of Tests 5 and 2 is confirmed by the internal consistency figures in Table 8.

The stepwise regression analysis suggests that fairly advanced learners' overall proficiency in speaking can be measured quite effectively with a limited number of situational response tasks. Asking a limited number of questions given in L1 and embedding them in a situational context raises the predictability of overall performance to a very high level.

Using *forced entry*, it is possible to get an idea of what proportion a certain predictor makes when it is entered in the regression equation last. In the forced entry the rest of the predictors are allowed to predict as much as they possibly can of the variation in the dependent variable. This rigorous analysis showed that Test 5, in the very least, makes about 10 percentage point contribution to the prediction. Using the same method, the rest of the tests make almost no contribution as the last predictor. This analysis confirms the significance of Test 5, which requires the learners to cope in everyday situations by responding appropriately or presenting opinions in verbally described situations.

Another regression analysis was run with the same five predictors but with ACTFL OPI as the dependent variable. This was done in order to provide a kind of replication of the prediction exercise within the LLOPT framework. In this case, only a stepwise regression analysis was performed. It appeared that tests 3 and 4 could not be entered since they failed to satisfy the traditional entry criteria. The prediction model is summarized in Table 15.

Some differences are visible if we compare the stepwise model. The order of predictors is different. In case of the LLOPT, by far the best predictor was Test 5 with Test 2 as making a substantial contribution to the explanation of variance in the overall score. In predicting performance on the OPI, by far the best predictor was Test 1, and Test 5 made a substantial further contribution to prediction. The power of prediction is about 10 percentage points lower with the OPI as the dependent variable, but still remarkably high, close to 80%. The fact that predictability was even higher within the LLOPT is partly explained by technical reasons: the final score is made up of the weighted sum of the predictors.

TABLE 15 Results of the stepwise regression analysis with the ACTFL OPI as the dependent variable

Model	Variables		R	R Square	Adjusted	Standard Error of the
	Entered	Removed			R Square	Estimate
1	Test 1	-	.822	.675	.670	.7219
2	Test 5	-	.871	.759	.751	.6274
3	Test 2	-	.887	.787	.775	.5961

The analysis with the OPI as the dependent variable gives support to the conclusion that a language laboratory test of oral skills lends itself quite well to the purposes of oral testing and that only a limited set of tasks are needed for an adequate estimate of oral proficiency.

To sum up, both the correlation figures and the regression analyses show that the LLOPT subtests are efficient instruments of assessing oral language proficiency. In planning the future use of the test, it is useful to know that also the simplest sections like Subtest 1, Reading Aloud, proved to give sufficiently valid information. The question of how many tests to use in a potential school-leaving test depends on many factors such as the available resources, the needed power of discrimination, and the desired washback validity.

10.2 The ACTFL OPI as a validating instrument

Research question 2 asked whether the language laboratory test can be validated by means of the ACTFL oral proficiency interview. When the LLOPT test was designed, the ACTFL oral proficiency interview was chosen as the validity instrument with which the new test would be compared. At that stage the writer believed that the ACTFL OPI would be an unquestionably valid measure of oral proficiency. Literature on oral testing and over 70 interviews has, however, revealed some new aspects of not only the ACTFL but also of the interview in general. (Researchers' views on ACTFL validity have been reviewed in Section 6.1).

The administration of the interviews was particularly revealing, because it showed how vulnerable the reliability, and consequently the validity, of the interview may be. As I have not come across any literature about a nonnative tester performing a demanding interview like the ACTFL, I will describe my experiences of the procedure below. After that I will report on how the ACTFL measured the oral proficiency of the target population.

Reliability and validity

The crucial point of the interview is the elicitation of a ratable sample. The ACTFL executors are right in keeping up the standard by a strict system of certification, because the task of the interviewer is demanding. The interviewer's position resembles that of a theater director. The text and the actors (interviewees) should be in focus, but it is the director (interviewer) who is responsible for everything to work smoothly and for the overall gain of the spectators (assessors). In one respect the interviewer's task is more demanding than the director's: in a play the text is the starting-point, whereas in an interview also the text has to be created. To perform all that, the interviewer has to keep several aspects in mind: the working memory is constantly on trial. Among the crucial factors that the interviewer has to keep in mind are for instance the creation of a positive climate, the choice of relevant topics, and the observation of the smooth flow of conversation. A nonnative tester is often also concerned about her own L2 speech. In addition, in the administration of the ACTFL OPI it is important to follow the proper stages.

For any conversation to succeed the *affective balance* is of utmost importance. This is even more the case in an oral interview, where the test situation and the skewed distribution of power between the participants are apt to upset the delicate balance. An alleviating factor in this particular test was the awareness that the results of the test would not have any serious consequences for the students' future. On the other hand, the students - particularly the ones tested early - did not know what the test was going to be like, which may have increased the nervousness. It was the tester's task to try to put the students at ease, to convey the feeling that what the student was saying was worthwhile and interesting, and to leave her with a feeling of success. To be able to accomplish this, the tester herself had to master her own states of mind and feelings, such as nervousness in the first interviews.

To find the suitable *topics* for every test-taker was an important factor contributing to the right atmosphere. The interviewer had to find the right balance between freedom and control. The interviewee had to feel free to talk about subjects that interested her, but at the same time the interviewer had to ensure that she got a wide sample, which gave a many-sided and sufficiently extensive picture of the interviewee's speaking skill. A quick search through the commonest subjects: family, home, hobbies, summer jobs, and trips abroad often gave a clue. A narrow course between something relevant and something not too personal had to be kept.

The discussions with the more taciturn testees were not totally unproblematic. Sometimes the age gap or the different sex or both made matters worse: the two parties' knowledge of the world was too different. In this respect the tester's task of interviewing was different from that of a journalist: when a journalist has to make an interview about a totally unknown subject, he can prepare himself. In testing, on the other hand, the topic comes up quite unexpectedly. With the two favorite subjects,

sports and pop music, this was, however, not the case: they occurred regularly, but the interviewer was every time equally ignorant. Nevertheless, instances of the opposite also occurred: there were topics and/or opinions that were so interesting that the interviewer forgot her role and was carried away by the conversation. This was sometimes to the disadvantage of the interview: too much time was spent with one particular interviewee on one particular topic.

Candidates of different temperaments presented the problem of *treating everybody fairly*, which was not automatically the same as treating everybody equally. How long a pause should the interviewer tolerate, and/or should she give help and encouragement? No general rule could be followed, but each case had to be decided individually. It seemed natural to give a candidate of apparently slow temperament more time than to the others, but surely a limit had to be set. When a linguistic breakdown occurs, it is an instinctive act for a language teacher to offer help, and every now and then the missing word did slip from the tester's lips. There were two kinds of candidates to whom the tester appeared to offer help more easily than to the others: on the one hand, the very slow ones, who seemed to try one's patience in the extreme, and on the other hand, the very good ones, whose minor instances of being lost for a word only appeared like an occasional slip of the tongue. The need of help seemed as much a matter of personality as of L2 proficiency: a very reticent young sportsman did need more encouragement than a verbose would-be actress[5].

The interviews convinced the present tester of the fact that it was impossible to hold the conditions constant. If a great deal (some scholars say most, see 4.1) of communication takes place nonverbally, how can the tester control for example encouragement given by her eyes or facial expressions? In addition, the *physical conditions* such as the time of the day appeared to have a great impact. During the very first interview in the morning neither the interviewer nor the interviewee seemed to be fully awake, and the best interviews were usually the second or third in the morning. As the day advanced, problems of inattention and failing memory became more frequent, and the last candidates of the day were clearly in a worse position than the ones in the morning. With the slow candidates, whose speech came sluggishly word by word, the interviewer sometimes caught herself losing track and being carried away by her own thoughts. Similarly, an everyday topic (family, home, way to school, etc.) demanded extra concentration. It was sometimes difficult to remember whether some particular question had been asked of this particular candidate, and/or what she had told at the beginning of the interview.

Another factor that may have had an influence on test reliability was the fact that *the interviewer was speaking a foreign language*. The test fatigue, which affected the interviewer's alertness, also had consequences on her L2 speech. Later in the afternoon there were occasions when it was suddenly difficult to retrieve even a familiar word. Carrying out an interview is, however, a semantically very demanding challenge to a nonnative speaker. Theoretically, it is the interviewer who

---

[5] Different temperaments posed the question of assessment: are we assessing language proficiency or personality if we give a lower score to a slow candidate than to an entertaining one? In my opinion it is unnecessary to separate the two. In real-life communication, too, people prefer a witty companion to a boring one. This view is shared by e.g. Huhta (1994) and Pasanen and Hietanen (1994).

has control of the course of the interview, but, in practice, the interviewee may give the conversation an unexpected turn, and any topic may suddenly come up. The interviewees may have the most unusual hobbies, or they have, for instance, seen a very strange film. It is common knowledge to any nonnative speaker - teachers not excluded - that when a subject comes up unexpectedly, it may be difficult to retrieve a missing word as quickly as necessary. The needed schema is simply not available. When a testee missed a word, she usually appealed to the tester, who could, of course, ask the student to explain the matter in other words, but there were occasions when such procedure did not seem natural. Sometimes the interviewer would have needed that very word to carry on the conversation, particularly to ask a follow-up question. Though a teacher must often admit that she does not know every word, such a confession seemed less proper in a testing situation.

In a structured interview like the ACTFL OPI the *interviewer's working memory* is under higher pressure than in an unstructured test. In addition to the strains discussed above, the interviewer has to keep in mind things like whether she has had enough samples of the past tense, whether the floor has been established properly, whether there have been a sufficient number of examples of description and narration, etc. The move to the next probe depends on the interviewer's estimation of the testee's level of proficiency, which means that assessment has to go on all the time.

A further aspect of the reliability of the interview is its *technical standard*. A performance recorded on a videotape is more reliable than one on an audiotape, because it is more natural to observe for example pauses or small talk when one sees the expressions and gestures. However, cost and practicality decided the choice of audiotapes.

It is natural that the quality of tape-recorders and tapes has an effect on reliability. The two visits to the participating schools left the tester with the impression that the technical equipment at schools is far from satisfactory. The tester had brought her own audiotapes, but the school had promised to provide the tape-recorder. Though no specially high quality was demanded, there were great difficulties in getting a satisfactory recording in the first school. There was exactly the same problem in the second school, but this time the tester had brought her own tape-recorder.

Concerns like those discussed above are meant to show how exacting it is to conduct a structured interview and, accordingly, how frail test reliability and validity are. It seems strange that, in reporting research results, the interviewers' competence is seldom or ever mentioned. However, the present writer does not believe that her problems would have been unique or even out of the ordinary. The ACTFL results, which will be dealt with in the next section, will at least show that it was possible to elicit interviews of different levels and to give them mainly unanimous assessments.

Results of the ACTFL OPI

Of the two tests, the ACTFL oral proficiency interview and the language laboratory test, the ACTFL was carried out first at both schools. The fact that the tester was a stranger and the format of the test partly unknown was thought to be a source of anxiety, which would be somewhat reduced if the socially easier test was taken first

and the testees had an opportunity to get personally acquainted with the tester. The testees' oral comments indicated that the decision was right.

The ACTFL interview scale consists of ten levels if the sublevels (for level descriptions see Chapter 9) are counted as well. It is obvious that if the candidates in a test have all studied the language for a certain number of years and passed the previous tests, their results will not be spread all over the scale but concentrate on some section. On the basis of the course in Siuntio (section 8.2.4) and the pilot tests, the cut-off points were placed between Novice Mid and Novice High at the lower end and between Advanced and Advanced Plus at the upper. It seemed justifiable that no one who had studied English for more than nine years should be passed at less than Novice High. There was a temptation to set the lower cut-off point even one step higher up, but as this was the first "official" oral school leaving test ever, there was place

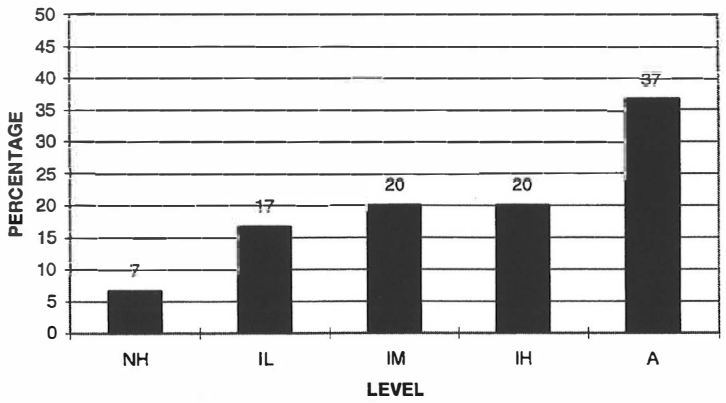

FIGURE 14 The percentile distribution of the ACTFL oral interview (OPI)

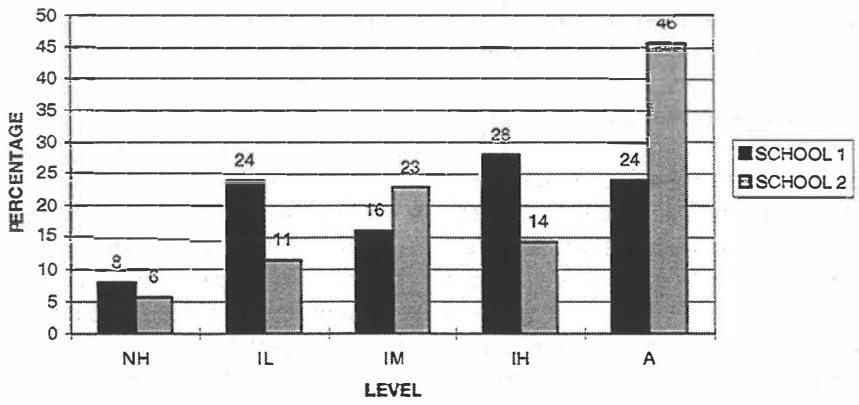

FIGURE 15 The percentile distribution of the ACTFL oral interview (OPI) by school (School 1 = Halikko, School 2 = Joensuu)

for some lenience. Setting the upper cut-off at the Advanced Level did not mean that the potential existence of better candidates was excluded, but there was no need and no testing proficiency to draw the line higher.

For comparability with the other tests the different levels were given the following numerical equivalents: Novice Mid 0, Novice High 1, Intermediate Low 2, Intermediate Mid 3, Intermediate High 4, Advanced 5. The mean of the test was 3,7/5, (72 % of the maximum) and the standard deviation 1,28. The distribution of the different levels is shown in Figure 14. A certain skewness of the distribution was to be expected even on the basis of such early figures as the average of the marks in English at the end of the comprehensive school (see section 9.1).

Another expected result was the distribution of the figures in the two schools, which also corresponded to the early results and was repeated in the figures of the language laboratory test (Figure 15).

Because a marked difference in the English language proficiency between boys and girls had recently been shown at the end of the junior high school (Pasanen & Hietanen 1994) and also at the end of the primary school (Huttunen & Kukkonen 1995), the writer wanted to see whether there was a difference also in Halikko and Joensuu. The results were similar to those of the two other studies (Figure 16), but the level of significance was only suggestive. The fact that the interviewer was a woman with less shared knowledge and interest with boys than girls may be one factor in explaining the superiority of girls.

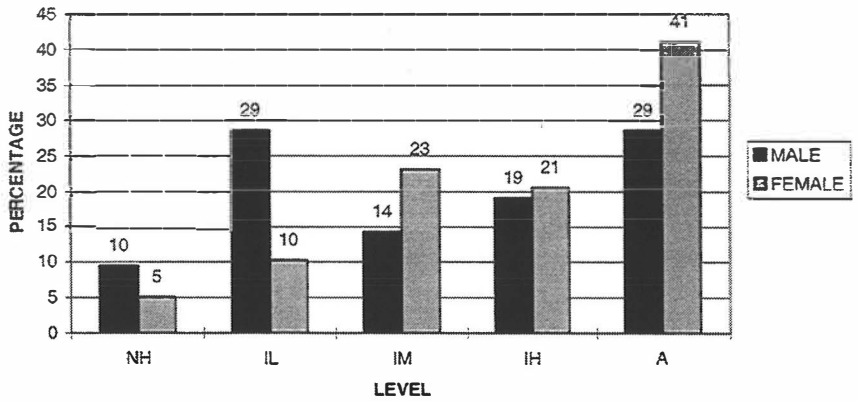

FIGURE 16 The percentile distribution of the ACTFL oral interview (OPI) by gender

On the whole the ACTFL interview results were remarkably good, particularly if compared with such foreign results as those of Carroll (1967) and Lafayette (see 6.1). However, after nine and a half years' study of English, with a popular subject and motivated students, powerful mass media and entertainment exposure, and the advanced teaching methods in the two experimental schools, it would have been surprising if the achievement had been poor.

The question of validating the language laboratory test by means of the ACTFL interview cannot be answered unambiguously. As the reciprocal correlation of the tests was as high as .88, higher than the correlation of each test with any other

indicator, like the results in the matriculation examination, it is likely that the tests assessed more or less the same trait, evidently oral L2 proficiency. Similarly, Luoma (1997) reports a correlation of .85 between the interview test and the language laboratory test in the new Finnish national language certificate (cf. 6.3). However, mere correlation is not a sufficient proof of validity (Shohamy 1988, 1994; Weir 1988, 30; Luoma 1997 found that - in spite of the high correlation - the discourse produced in each test was different), particularly not in a case like this, when the validity of the validator test (here the ACTFL) is disputed. On the other hand, a perfect test has not yet been created, so any test validation by means of another test is somewhat questionable.

10. 3 Attitudes towards learning and testing spoken language

Reseach question 3 addresses the learners' attitudes towards the learning and testing of speaking. Since the final examination with its high stakes is a highly emotional matter, it seemed natural that, for test development purposes, attitudes should be assessed. Because the testees had just participated in a three-year experiment of developing oral proficiency, it was supposed that here the alleged Finnish communication apprehension would be less manifest than elsewhere. The results are presented in Table 16.

The fact that the survey was conducted immediately after the language laboratory test, which many students experienced as difficult, may have affected the results. Nevertheless, the attitudes towards both learning and testing the spoken language were strongly positive. The majority of the students experienced speaking practice as pleasant (90%) and easy (92%). They also thought that not only in EFL lessons should oral practice be increased (88%), but also in mother tongue lessons (73%).

There was a marked difference between the two schools. As far as speaking and practicing were concerned, both groups expressed a positive attitude, but when it came to the two tests, the Halikko students were clearly more reserved. In Joensuu positive attitudes were indisputable: the great majority of the Joensuu students experienced both the ACTFL and the LLOPT as easy (86 respective 65) and pleasant (80 respective 74), though in a choice situation only 22% would rather have taken part in the LLOPT, as opposed to 54% who would have chosen the ACTFL OPI. In Halikko only 46% thought that the ACTFL was easy, but 77% still regarded it as a pleasant experience. For the LLOPT the corresponding figures were only 27% and 35%. The cause of the difference can only be guessed: the Joensuu students may have been more accustomed to interviews or work in the language laboratory, or they simply experienced the tests as easier because they were more proficient.

One should be cautious about generalizing anything concerning the attitudes towards the interview or the language laboratory test. When speaking about *the* interview or *the* language laboratory test one should remember that the circumstances vary from case to case. Factors like the content of the subtests and tasks in the language laboratory test or the voice of the reader and the personality of the

interviewer in the interview are important variables, which may explain part of the variation in the attitudes.

TABLE 16 Results of the student attitude questionnaire. N = 60 (Halikko 26, Joensuu 34). The results are indicated in percentages. The first figure indicates the Halikko result, the second Joensuu, and the figure underneath the total. Scale: ++ I quite agree, + I agree, ? I do not know, - I disagree, -- I do not agree at all.

Scale of attitudes	++	+	?	-	--
5. I like to speak English outside school	15/17	27/34	15/29	35/20	8/0
	17	32	22	27	3
6. I do not like to speak English at school	4/0	23/9	8/0	31/37	35/54
	2	15	3	33	47
7. I do not like to speak to unknown people	0/0	12/3	8/6	31/17	50/74
even in Finnish	0	7	5	24	64
8. There should be more oral practice in	39/57	42/37	8/0	8/6	4/0
English lessons	48	40	3	7	2
9. There should be more oral practice in	31/34	39/43	12/11	19/11	0/0
mother tongue lessons	32	42	12	15	0
10. My English teacher speaks too little English	12/3	8/17	12/9	50/23	19/49
	7	12	10	35	37
11. Oral practice is very unpleasant	0/0	8/3	8/3	35/31	50/63
	0	5	5	33	57
12. Oral practice is very difficult	0/0	12/0	4/3	50/49	35/49
	0	5	3	48	43
14. The language laboratory test was easy	4/14	23/51	15/17	50/11	8/9
	10	38	15	28	8
15. Participating in the language laboratory test	8/31	27/43	8/11	42/9	15/6
was a pleasant experience	22	35	10	23	10
16. The interview was easy	27/23	19/63	31/9	15/6	8/0
	25	43	20	10	3
17. Participating in the interview was a pleasant	35/37	42/43	4/14	19/3	0/3
experience	37	43	8	10	2
18. There should be an oral part in the English	12/37	46/40	15/11	12/9	15/3
matriculation examination	25	43	13	11	8
19. Having an oral part in the matriculation	42/69	46/29	8/3	4/0	0/0
examination would augment oral practice at	57	37	5	2	0
school					
20. There should be a language laboratory test in	12/43	31/26	15/14	27/14	15/3
the matriculation examination	28	28	15	20	8
21. There should be an oral interview in the	39/40	27/46	8/3	15/9	12/3
matriculation examination	38	38	5	12	7
22. I would rather participate in the language	12/11	8/11	8/23	12/34	62/20
laboratory test than in the interview	12	8	17	25	38

Before a compulsory oral test is introduced, it would seem advisable to find out what the target group's attitudes toward it are. Special weight should be laid on the attitudes of such students - as these here - who have personal experience of oral tests. There were 44% of the students in Halikko who would have introduced an LLOPT in the matriculation examination and 69% in Joensuu, while the attitudes towards an interview were more positive: its introduction was supported by 66% in Halikko and

86% in Joensuu. The questionnaire was evidently answered on purely experiantial basis without paying any attention to the cost or other practical matters.

10.4 Impact of staying abroad on the results

A possible compulsory oral test at the end of secondary education has been opposed on the grounds that it would accentuate social inequality. It has been claimed that well-off people would have money to send their children to English-speaking countries, and they would then do better than poorer people's children. Research question 4 addresses this issue. Though the population in the present study was too small to make any generalizations, it seemed interesting to try to find out to what extent this contention would be true about these students.

It appeared that not many students had spent any longer time abroad. The answers were grouped into three categories: those who had spent 7 days or less, those between 8 and 45 days and those over 45 days. It was believed that a period of 7 days or less would not have had any essential influence on language proficiency, so it was marked as 0. The results can be seen in Table 17.

TABLE 17 Length of stay in an English-speaking country

Length of stay	0-7 days	8-45 days	Over 45 days	N
Halikko	15	7	3	25
Joensuu	22	9	4	35
Total	37	16	7	60

To see the connection between the time spent in an English speaking country and the success in the oral tests, an investigation was made into the length of stay (Table 20, Appendix 4). And truly, it did appear that many of the best students had spent a considerable stretch of time abroad, although it also appeared that it was quite possible to achieve a good result without any stay abroad. Naturally a visit to an English-speaking country does not affect the oral skill only: the same students had succeeded well also in the written tests. It is also difficult to tell which was the cause and which the effect. Were the students good because they had been abroad, or did the students who had a special talent for English also have a particularly strong desire to visit an English-speaking country? The same uncertainty about the cause and effect is true about activities that involved the use of English. However, most good students told that they did something extra to learn and keep up English (Table 20, Appendix 4).

11 CONCLUSIONS

In this chapter there will first be given a summary of the results depicted in the previous chapter. The results will then be surveyed in relation to the tasks of this study, and some conclusions and recommendations will be presented.

A test is a mirror of teaching. In the two oral tests, the ACTFL interview and the language laboratory test, the students' standard of proficiency was high. There were a great number of good results, but there was also remarkable dispersion. In the interview as many as 37% of the students gained the highest score and in the language laboratory test 40%, but in the latter test 5% of the students also failed. That there were no failures in the interview was a technical decision more than a state of facts. The subtests of the LLOPT showed what the strengths and weaknesses of the students were: they were good at interpreting everyday dialogue and at reproducing a story, but their pronunciation and fluency were not as good as one might have expected, and they had difficulties with a longer presentation. This result may indicate that the presently most popular FL method, the communicative approach, has its deficiencies as any other method: it stresses interactional skills but pays less attention to developing the more technical skills such as pronunciation and fluency and does not favor presentation.

The subjects in the test had had more training in speaking than the rest of the age group, which makes also their all-round results in the matriculation examination worth looking at. The very fact that the LLOPT correlated highly with other measures (with the ACTFL .88, with the school report .83, and with the matriculation examination .80) shows that the language laboratory test gives a reliable picture of the general standard. In the matriculation examination the high correlation with the essay (.80) and the listening comprehension (.71) seems to indicate that practicing speaking develops the productive skills as a whole and also the other oral skill, listening. The low correlation, .39, with the reading comprehension test may show that less attention had been paid to the reading skill.

The psychometric information confirmedthat the language laboratory test is a reliable measure of oral proficiency. Among factors contributing to the reliability were the length of the test, the same input to all subjects, and the creation and weighting of the criteria on the basis of the shared knowledge of several experts. A

weak point in reliability is the technical vulnerability. Conducting the ACTFL oral proficiency interview, on the other hand, showed that at the elicitation stage it was difficult to keep all the variables constant. Particularly in the case of advanced candidates the linguistic demands combined with the situational complexity present so demanding a challenge to a nonnative interviewer that both reliability and validity may suffer.

As far as validity is concerned critics accuse a language laboratory test of not being authentic. In this particular test a special effort was made to compensate for this defect. Emphasis was put on the a priori validity - for instance the extensive domain specification, the communicative frame, and the variety of subtests -, and the students' verbal comments in the attitude test were very positive, even more so than the percentile information of the likes and dislikes (57% considered the language laboratory test a pleasant experience, and 56% thought that the matriculation examination should include one). In the actual test situation the students devoted themselves to the simulated conversations (e.g. Task 5) in a most whole-hearted way.

The language laboratory test was originally chosen as a testing instrument because of its alleged efficiency. The expectation was cinfirmed: a 45-minute session yielded a 20-minute sample of student speech, and a further analysis of the material proved that less than that could be enough. However, the absolute precondition for an LLOPT is the existence of a language laboratory. The number of participating students divided by four is the minimum prerequisite (see p 130).

The main research task of this study was to find out whether a language laboratory test is a suitable instrument to assess school-leavers' oral L2 proficiency. In the background there were, however, two other questions that were not explicitly stated. With Hellgren's work in mind (section 6.3), there was a suspicion that perhaps an oral test is unnecessary, after all. The second issue concerned the superiority of the interview over the language laboratory test, that is, whether the interview would, nevertheless, be the better format for the present purpose.

The domain analysis at the beginning of this study suggests that the case for testing oral proficiency is strong. The accelerating internationalization brings more and more people into close contacts, and the needs analyses show that the greatest public demand is for the speaking skill. Besides, the oral skill is different from the writing skill. The characteristic features of spoken language such as pronunciation, intonation, fluency, appropriacy, and conversational strategies can only be assessed with an oral test. More than half of the participating students would have gained a different score in the spoken test from what they gained in the present matriculation examination. The measures taken to develop the spoken language also need testing for feedback. Furthermore, the literature reviewed in this study has shown that testing has a powerful washback on teaching and learning. If we continue to test only the written language, we develop mainly writing proficiency. If the aim of the matriculation examination is also to develop teaching and learning in line with current needs, a speaking test is highly desirable, indeed - one could argue - necessary. Testing oral proficiency is particularly important at the end of the secondary school: at present no country can afford to do without many-sidedly bilingual students and professionals. The question that this study tried to answer concerns the best way of carrying out the necessary testing.

The study showed that a language laboratory test is an adequate means of assessing senior secondary school-leavers' oral proficiency. Because the number of the subjects was small and they were not selected with a view to statistical representativeness, it is, however, not feasible to draw any definitive conclusions. As for the superiority of the interview or the language laboratory test, judgments should be particularly cautious. First of all, one cannot generalize and speak of *the* interview or *the* language laboratory test. Different language laboratory tests need not have anything else in common except the technical means, the channel for transmitting the test. The same applies to an interview: rather than speak of the interview in general it is much more pertinent to speak of the ACTFL OPI, which is a well-described test. Nevertheless, even when the specification goes as far as speaking of the ACTFL OPI and the LLOPT, it is not possible to compare the two tests in isolation but only with the view of the purpose of the test.

If an oral test is needed for a language which the students have studied only a few years and in which they are likely to be at an elementary or lower intermediate level, both a language laboratory test and an interview seem appropriate. In a language laboratory test a great variety of test formats are available: the test may vary from a simple reading aloud or picture description task to a demanding presentation. The language laboratory test is, therefore, quite flexible and can be used at different levels. Also the ACTFL OPI is, in principle, usable at any level from novice to superior, but when it comes to mass use in a country with few competent testers, the matter is more complicated. At lower levels the ACTFL OPI type of interview can be used by an ordinary FL teacher with appropriate training, but to conduct an advanced to superior level interview is merely linguistically quite demanding and needs thorough training and great skill to be reliable. This, again, involves costs which are usually beyond the means of the organizing institution. In addition, Stansfield and Kenyon (1992b) suggest that the ACTFL OPI might not be as good as the SOPI for candidates at the advanced level and those above it.

Another important factor in deciding what test format to use is the weighting given to the examination. Is it still going to be a voluntary test, for which a special certificate is given, or will it be made a regular part of the FL matriculation examination, which then has the same import as any other part? The first case may contribute to inequality between students: progressive, enthusiastic teachers will prepare their students for the oral test, whereas their less ambitious colleagues cannot be bothered. Such a situation might further improve the lot of the students in the densely populated urban areas. For mainstream Finnish FL teaching and learning such a state would mean little progress, nor would the test format be of major significance. In the latter case, however, if the oral test is made compulsory, the test format will matter. If far-reaching decisions about the students' future are based on the results, the examination has to be highly reliable, which, with the present resources in Finland, means that the L2 examination has to be a language laboratory test. Another alternative would be an ACTFL type of interview, but then the testers would have to be specially trained and only a fraction of the students could be tested each year.

If a compulsory language laboratory test is chosen, it will have to meet certain criteria. The present study was too limited to give any definite guidelines, but it can offer some suggestions. The first criterion is self-evident: the test should be founded

on the curriculum. It is true that the current senior secondary school curriculum is quite broad to base any detailed descriptions on, but what it does imply is the fact that language teaching in Finland should be communicative. Accordingly, the school-leaving test should also be communicative and meet the criteria of a communicative test (cf. section 7.3.1).

The models of communicative competence and the oral skill domain specification carried out in this study have shown how multifaceted oral proficiency is. It is, therefore, natural that it cannot be validly tested with any one test format but needs several subtests. Because the language laboratory test is flexible, different, also new formats can be incorporated into it. If beneficial washback is aimed at, it might be advisable to agree on a pool of test formats, a certain number of which could be chosen for each year's examination. This pool could contain formats that would test the most desirable skills, which would ensure beneficial washback. A judicious choice of criteria would influence teaching and learning.

The introduction of a nation-wide oral examination for over half of the age-group is an important economic issue. One of the motives of the present study was to find a maximally cost-effective solution. On the basis of the evidence here it would seem that a language laboratory test could be made simpler than the test designed for this experiment. The regression analysis showed that even one subtest, Reacting in Situations and Expressing Opinions, gives valuable, perhaps sufficient information. From this point of view the present test might be simplified or shortened. This could be done in three ways: the test could consist of fewer subtests, the subtests could be shorter, and fewer criteria could be used.

The choice of the subtests would have to be made considering, for example, the ease of designing and assessment and the desired washback effect. All the three test types: reading aloud, interpreting, and reacting in situations which proved powerful predictors in the regression analyses give the test designer plenty of choice. Used in combination, they would each have an important share to contribute, particularly interpretation and reacting in situations. One of the important results of this study was the discovery of the fact that simple and easy subtests, which have not necessarily been conspicuous in the international testing literature, may give valuable results. One of such tests is interpretation from L1 to L2. Reacting in situations, on the other hand, is also a test format which comes as close to natural interaction as it is possible to get in the language laboratory. However, if also a skill in transactional speech is considered an important objective of oral teaching, the test should occasionally include some story-telling or oral presentation, too. An oral presentation seems also to have significant differentiating power. If the matriculation examination is one of several levels, an oral presentation would be appropriate at the higher level(s).

As for the length of the subtests, in the LLOPT the reading aloud and the oral presentation were unnecessarily long. It would be a useful task for further experimental research to find out what the minimum sufficient length would be. A 10-15 minute sample of student speech might well be enough.

In addition to the number and length of the subtests, the third possibility to cut down expenses would concern the number of criteria. It would seem quite natural to use pronunciation and fluency as criteria in an oral test, but they need not be rated in

every subtest. The comparison of the results in the various subtests appeared to signify that pronunciation and fluency could be assessed in only one subtest.

The Federation of Foreign Language Teachers in Finland has taken a negative attitude towards an oral section in the school-leaving examination (Tuomiharju 1993). The main reason seems to be the teachers' fear that a new test would add to the amount of work, possibly also a fear of the new. It is true that secondary school L2 teachers' work-load is heavy, and there is no need to increase it. If, however, the rest of Europe is following the Council of Europe's suggestion and going in for oral testing and the accompanying development of the speaking skill, Finland can hardly afford to fall behind. The decision-makers will evidently have to find a compromise. One alternative would be to have the oral test of the matriculation examination at an earlier stage, say a year before the written parts. Students could then at different stages of their studies pay more attention to developing some particular skill(s). In the matriculation examination the student might be able to choose to take either an oral or an aural or a written test or any combination of the three.

If, on the other hand, all the four skills are tested at the same time, some simplification of the existing matriculation examination will be necessary. The separate subtest that assesses grammar and vocabulary might be abolished and the equivalent criteria be incorporated in the other subtests. The writing skill is at present tested both with open answers and with an essay test. In real life there are few, if any, situations in which a non-professional person would need the L2 essay writing skill. It might be possible to simplify the testing of writing and to test the students' L2 production and presentation in the oral form. Except for the initial training, this arrangement would not increase the work of either the teachers or the Matriculation Examination Board nor the manpower cost of assessment. However, any changes in the matriculation examination should always be weighed carefully. Testing should not become an end in itself but should ultimately be an aid to learning and teaching. Each of the four language skills is necessary and should be tested in proportion to its importance to learners. On the other hand, not everything that is useful can be tested.

If a language laboratory test were adopted mainly for added test reliability, it would be unfortunate if the technical standard could not meet the expectations. Reliability of assessment can be increased by teacher education and landmark tests, but the technical development is in the hands of the manufacturers. Before introducing a language laboratory test, the administrators would need to have absolute certainty that every student's production can be recorded reliably. To a recent inquiry the manufacturers answered that progress should have been made since the present tests were carried out in 1993.

Because the material collected for this study was so rich, it would have been possible to pose and answer further questions. One can but hope that it is possible to continue in the future. It would, for instance, be interesting to know how native speakers (from, perhaps, different parts of the English-speaking world) would rate the material, such as the test of reacting in situations, which the Finnish raters found difficult to agree about. In that particular subtest it would have been possible to look closer into the language that the students used to sound polite. Another research task would be a discourse analytic comparison of the language in the interview and the language laboratory test, which might give useful material for the comparison of the two test formats. The present researcher's failure to find appropriate criteria for

analyzing discourse competence should not prevent others from making an attempt. Furthermore, the whole area of fluency and pronunciation, particularly the prosodic features, would need closer investigation. It is actually striking how little research on Finnish learners' suprasegmental features was found.

To summarize, learning and speaking a foreign language is a live field with constant development. What seems a good test today, will probably be found inadequate tomorrow. Also at present there are many unanswered questions, like the relationship of an L2 speaking test to an L3 or L4 or even L1 speaking test. In a country where every child has to learn at least two foreign languages and the majority of them for the minimum of eight to ten years, the investment in language teaching is considerable. The question arises, though, whether the present resources are optimally spent. If the country can afford to pay the salary of 7000 language teachers, one should think that it would also be able to maintain a handful of researchers. Holland is a leading country in foreign language teaching and learning, but it does have a national language testing institute, CITO, whose work has been of both national and international significance. It is true that we have research at universities, but the connection to language learning at school or to developing the matriculation examination or other common tests is only accidental. What private industry with 7000 employees could afford not to invest substantially in research? Resources spent on continuous, systematic development of teaching and testing would, no doubt, repay themselves in efficient and rewarding learning.

BIBLIOGRAPHY

Abercombie, D. 1956. Teaching pronunciation. In A. Brown (Ed.) 1991. *Teaching English pronunciation. A book of readings*. London: Routledge, 87-95.

Adams, C. & Munro, R. R. 1978. In search of the acoustic correlates of stress: fundamental frequency, amplitude and duration in the connected utterance of some native and nonnative speakers of English. *Phonetica* 35, 125-56.

Ahlroos, P. & Muilu, M. 1989. Communication objectives in English language teaching. A study of the hidden curriculum. Unpublished Pro Gradu thesis. Jyväskylä: University of Jyväskylä. Department of English.

Albrechtsen, D., Henriksen, B. & Faerch, C. 1980. Native speaker reactions to learners' spoken interlanguage. *Language Learning* 30, 365-96.

Alderson, J. C. 1981. Reaction to Morrow paper (3). In J. C. Alderson & A. Hughes (Eds.) *Issues in language testing*. ELT Documents 111. London: The British Council.

Alderson, J. C. 1993. The state of language testing in the 1990s. In A. Huhta, K. Sajavaara & S. Takala (Eds.) *Language testing: new openings*. Jyväskylä: University of Jyväskylä. Institute of Educational Research, 1-19.

Alderson, J. C. & Wall, D. 1993. Does washback exist? *Applied Linguistics* 14 (2), 115-29.

Allen, G. D. 1975. Speech rhythm: its relation to performance universals and articulatory timing. *Journal of Phonetics* 3, 75-86.

Anastasi, A. 1986. Evolving conceptions of test validation. *Annual Review of Psychology* 37, 1-15.

Anderson, J. 1985 (1980). *Cognitive psychology and its implications*. 2nd ed. San Francisco: Freeman.

Anderson-Hsieh, J., Johnson, R. & Koehler, K. 1992. The relationship between native speaker judgements of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning* 42, 529-555.

Anderson-Hsieh, J. & Koehler, K. 1988. The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning* 38, 561-92.

Angoff, W. H. 1988. Validity: An evolving concept. In H. Wainer & H. I. Braun. *Test validity*. Hillsdale, NJ: Erlbaum, 19-32.

Arevart, S. & Nation, P. 1991. Fluency improvement in a second language. *RELC Journal* 22, 84-94.

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

Bachman, L. F. 1988. Problems in examining the validity of ACTFL oral proficiency interview. *Studies in Second Language Acquisition* 10 (2), 149-64.

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

Bachman, L. F. 1991a. *Communicative language test development*. Mimeo distributed at the symposium Language Testing in the 1990s. University of Jyväskylä, August 30-31, 1991.

Bachman, L. F. 1991b. *Designing and developing language tests*. Mimeo distributed at the symposium Language Testing in the 1990s. University of Jyväskylä, August 30-31, 1991.

Bachman, L. F. 1991c. What does language testing have to offer? *TESOL Quarterly* 25, 671-704.

Bachman, L. F. & Palmer, A. S. 1981. The construct validation of the FSI Oral Interview. *Language Learning* 31, 67-86.

Bachman, L. F. & Palmer, A. S. 1982. The construct validation of some components of communicative proficiency. *TESOL Quarterly* 16, 449-65.

Bachman, L. F. & Palmer, A. S. 1983. *Oral interview test of communicative proficiency in English.* Urbana, IL: Photo-offset.

Bachman, L. F. & Palmer, A. 1996. *Language testing in practice.* Oxford: Oxford University Press.

Bachman, L. F. & Savignon, S. J. 1986. The evaluation of communicative language proficiency: a critique of the ACTFL Oral Interview. *Modern Language Journal* 70, 380-90.

Barnwell, D. 1987. Oral proficiency testing in the United States. *British Journal of Language Teaching* 25, 35-42.

Barnwell, D. 1989 'Naive' native speakers and judgements of oral proficiency in Spanish. *Language Testing* 6, 152-63.

Baxter, J. 1980. How should I speak English? American-ly, Japanese-ly, or internationally? In A. Brown (Ed.) 1991. *Teaching English pronunciation. A book of readings.* London: Routledge, 53-71.

Beaman, K. 1984. Coordination and subordination revisited: syntactic complexity in spoken and written narrative discourse. In D. Tannen (Ed.) *Coherence in spoken and written discourse. Advances in discourse processes xii.* Norwood, NJ: Ablex, 45-80.

Beatens Beardsmore, H. 1979. The recognition and tolerance level of bilingual speech. *Working Papers on Bilingualism* 19, 116-28.

Beattie, G. W. 1981. Language and nonverbal communication - the essential synthesis? *Linguistics* 19, 1165-83.

Bennett, T. L. 1977. An extended view of verb voice in written and spoken personal narratives. In E. O. Keenan & T. L. Bennett (Eds.) *Discourse across time and space. Southern California Occasional Papers in Linguistics* 5.

Bentahila, A. & Davies, E. 1989. Culture and language use: a problem for foreign language teaching. *IRAL* 27 (2), 99-112.

Bhatia, V. K. 1993. *Analysing genre: language use in professional settings.* London: Longman.

Blum-Kulka, S., House, J. & Kasper, G. (Eds.). 1989. *Cross-cultural pragmatics: requests and apologies. Advances in discourse processes xxxi.* Norwood, N.J.: Ablex.

Blum-Kulka, S. & Ohlstain, E. 1986. Too many words: Length of utterance and pragmatic failure. *Studies in Second Language Acquisition 8,* 165-180.

Brown, A. 1988. Functional load and the teaching of pronunciation. In A. Brown (Ed.) 1991. *Teaching English pronunciation. A book of readings.* London: Routledge, 87-95.

Brown, G., Anderson, A., Shillcock, R. & Yule, G. 1984. *Teaching talk. Strategies for production and assessment.* Cambridge: Cambridge University Press.

Brown, G. & Yule, G. 1983a. *Discourse analysis.* Cambridge: Cambridge University Press.

Brown, G. & Yule, G. 1983b. *Teaching the spoken language*. Cambridge: Cambridge University Press.

Brown, P. & Levinson, S. C. 1987 (1978). *Politeness: some universals in language usage*. Revised edition. Cambridge: Cambridge University Press.

Brumfit, C. 1984. *Communicative methodology in language teaching. The roles of fluency and accuracy*. Cambridge: Cambridge University Press.

Buck, K. (Ed.) 1989. *The ACTFL Oral Proficiency Interview. Tester training manual*. Yonkers, NY: The American Council on the Teaching of Foreign Languages.

Burgoon, M. & Ruffner, M. 1978. *Human communication*. New York: Holt, Rinehart and Winston.

Butterworth, B. 1975. Hesitation and semantic planning in speech. *Journal of Psycholinguistic Research* 4, 75-87.

Bygate, M. 1987. *Speaking*. Oxford: Oxford University Press.

Byrnes, H. 1989. The rating scale. In K. Buck (Ed.) *The ACTFL oral proficiency interview. Tester training manual*. Yonkers, NY: The American Council on the Teaching of Foreign Languages. Chapter 2.

Callamand, M. 1987. Analyse des marques prosodiques de discours (Analysis of the prosodic discourse markers). *Etudes de Linguistique Appliquée* 66, 49-70.

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

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

Carrell, P. L. & Konneker, B. H. 1981. Politeness: comparing native and nonnative judgements. *Language Learning* 31, 17- 30.

Carroll, J. 1961. Fundamental considerations in testing for English language proficiency of foreign students. In *Testing the English language proficiency of foreign students*. Washington, DC: Center for Applied Linguistics, 30-48.

Carroll, J. 1967. Foreign language proficiency levels attained by language majors near graduation from college. *Foreign Language Annals* 1, 131-51.

Carroll, J. 1980. *Testing communicative performance: an interim study*. New York: Pergamon.

Chafe, W. L. 1982. Integration and involvement in speaking, writing and oral literature. In D. Tannen (Ed.) *Spoken and written language. Advances in discourse processes ix*. Norwood, NJ: Ablex, 35-53.

Chalhoub-Deville, M. 1995. Deriving oral assessment scales across different tests and rater groups. *Language Testing* 12, 16-33.

Chomsky, N. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Clark, H. H. & Clark, E. V. 1977. *Psychology and language*. New York: Harcourt, Brace, Jovanovich.

Clark, J. L .D. & Clifford, R. T. 1988. The FSI/ILR/ACTFL proficiency scales and testing techniques. Development, current status, and needed research. *Studies in Second Language Acquisition* 10, 129-47.

Cohen, A. D. & Olshtain, E. 1981. Developing a measure of sociocultural competence: the case of apology. *Language Learning* 31, 113-34.

Conein, B. 1986. Conversation et interaction sociale: analyse de séquences d'offre et d'invitation (Conversation and social interaction: analysis of sequences of offering and invitation). *Langages* 81, 111-20.

Cook, G. 1989. *Discourse.* Oxford: Oxford University Press.

Coulmas, F. 1981. Introduction: conversational routine. In F. Coulmas (Ed.) 1981. *Conversational routine. Explorations in standardized communication situations and prepatterned speech.* The Hague: Mouton, 1-18.

Coulthard, M. 1985. (2nd ed.). *An introduction to discourse analysis.* London: Longman.

Cowie, A. P. 1989 (4th ed.). *Oxford advanced learner's dictionary of current English.* Oxford: Oxford University Press.

Cronbach, L. 1988. Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.) *Test validity.* Hillsdale, NJ: Erlbaum, 3-17.

Crown, C. L. & Feldstein, S. 1985. Psychological correlates of silence and sound in conversational interaction. In D. Tannen, & M. Saville-Troike. (Eds.) *Perspectives on silence.* Norwood, NJ: Ablex, 31-54.

Crystal, D. 1991 (1980). *A dictionary of linguistics and phonetics.* (3rd ed.) Oxford: Basil Blackwell.

Cumming, A. &. Berwick, R. (Eds.) 1995. *Validation in language testing.* Clevedon, OH: Multilingual Matters.

Currie, K. L. & Yule, G. 1982. A return to fundamentals in the teaching of intonation. In A. Brown (Ed.) 1991. *Teaching English pronunciation. A book of readings.* London: Routledge, 270-5.

Dalton. P. & Hardcastle, W. J. 1977. *Disorders of fluency and their effects on communication.* London: Edward Arnold.

Dandonoli, P. & G. Henning. 1990. An investigation of the construct validity of the ACTFL Proficiency Guidelines and oral interview procedure. *Foreign Language Annals* 23 (1), 11-22.

Davies, A. 1990. *Principles of language testing.* Oxford: Basil Blackwell.

Davies, E. E. 1987. A contrastive approach to the analysis of politeness formulas. *Applied Linguistics* 8 (1), 75-88.

Davies, N. F. 1982. Training fluency: an essential factor in language acquisition and use. *RELC Journal* 13, 1-13.

Dijk, T. A. van 1977. Context and cognition: knowledge frames and speech act comprehension. *Journal of Pragmatics* 1, 211-32

Drazdrauskiene, M.-L. 1981. On stereotypes in conversation, their meaning and significance. In F. Coulmas (Ed.) *Conversational routine. Explorations in standardized communication situations and prepatterned speech.* The Hague: Mouton, 55-68.

Eisenstein, M. 1983. Native reactions to non-native speech: a review of empirical research. *Studies in Second Language Acquisition* 5 (2), 160-76.

Eisenstein, M. & Bodman, J. W. 1986. 'I very appreciate': expressions of gratitude by native and non-native speakers of American English. *Applied Linguistics* 7 (2), 167-79.

Ellis, A. & Beattie, G. 1986. *The psychology of language & communication.* New York: Guildford.

Ellis, R. 1985. *Understanding second language acquisition*. Oxford: Oxford University Press.

Ellis, R. 1987. *Second language acquisition in context*. Englewood Cliffs, NJ: Prentice-Hall.

Els, T. van & De Bot, K. 1987. The role of intonation in foreign accent. *Modern Language Journal* 71, 147-55.

Enkvist, N. E. 1990. Success concepts. Keynote address at the NORDTEXT Symposium, Hanasaari, May 1990. In A.-C. Lindeberg., N. E. Enkvist & K. Wikberg (Eds.) *Nordic research on text and discourse*. NORDTEXT Symposium 1990. Turku: Åbo Academy, 17-26.

Faber, D. 1986. Teaching the rhythms of English: a new theoretical base. In A. Brown (Ed.) 1991. *Teaching English pronunciation. A book of readings*. London: Routledge, 245-58.

Faerch, C., Haastrup, K. & Phillipson, R. 1984. *Learner language and language learning*. Copenhagen: Gyldendals Sprogbibliotek.

Faerch, C. & Kasper, G. 1983. *Strategies in interlanguage communication*. London: Longman.

Faerch, C. & Kasper, G. 1984. Pragmatic knowledge: rules and procedures. *Applied Linguistics* 5 (3), 214-23.

Faerch, C. & Kasper, G. 1987. Procedural knowledge as a component of foreign language learners' communicative competence. *Aila Review* 4, 7-23.

Feyereisen, P., Pillon, A. & de Partz, M.-P. 1991. On the measures of fluency in the assessment of spontaneous speech production by aphasic subjects. *Aphasiology* 5 (1), 1-21.

Fillmore, C. J. 1979. On fluency. In C. J. Fillmore, D. Kempler & W. S.-Y. Wang (Eds.) *Individual differences in language ability and language behavior*. New York: Academic Press, 85-101.

Framework curriculum for the senior secondary school 1994. Helsinki: National Board of Education.

The framework of the Finnish national foreign language certificate. 1995. Helsinki: National Board of Education

Frederiksen, J. R. & Collins, A. 1989. A systems approach to educational testing. *Educational Researcher* 18 (9), 27-32.

Gardner, R. 1984. Discourse analysis: implications for language teaching with particular reference to casual conversation. State of the art article. *Language Teaching* 17 (2), 102-17.

Görding, E. & L. Eriksson. 1991. On the perception of prosodic phrase patterns. In Department of Linguistics, *Working Papers* 38. Lund: University of Lund, 45-70.

Goldman-Eisler, F. 1968. *Psycholinguistics: experiments in spontaneous speech*. London: Academic Press.

Grice, H. P. 1975. Logic and conversation. In P. Cole & J. Morgan (Eds.) *Syntax and semantics 3: speech acts*. New York: Academic Press, 41-58.

Gumperz, J. J. 1977. Sociocultural knowledge in conversational inference. In M. Saville-Troike (Ed.) *28th Annual Round Table Monograph Series on Language and Linguistics*. Washington, DC: Georgetown University Press.

Gumperz, J. J. 1978. The conversational analysis of interethnic communication. In E. Lamar Ross (Ed.) *Interethnic communication.* Athens, GA: University of Georgia Press.

Gumperz, J. J. 1979. The sociolinguistic basis of speech act theory. In J. Boyd & S. Ferra (Eds.), *Speech acts ten years after.* Milan, Italy: Versus.

Gumperz, J. J. 1982. *Discourse strategies.* Cambridge: Cambridge University Press.

Gumperz, J. J. 1987. Foreword. In P. Brown & S. C. Levinson. *Politeness: some universals in language usage.* Revised edition. Cambridge: Cambridge University Press, xiii-xiv.

Gutknecht, C. 1978. Intonation and language learning: the necessity for an integrative approach. In A. Brown (Ed.) 1991. *Teaching English pronunciation. A book of readings.* London: Routledge, 259-69.

Hackman, D. J. 1978. Rhythm in Finnish and English. In E. Görding, B. Bruce & R. Bannert (Eds.) *Nordic Prosody. Papers from a symposium.* Lund, Sweden: University of Lund. Department of Linguistics, 263-9.

Hall, E. T. 1977. *Beyond culture.* Garden City, NY: Doubleday.

Halliday, M. A. K. 1985. *Spoken and written language.* Oxford: Oxford University Press.

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

Hammerley, H. 1991. *Fluency and accuracy.* Cleveland, OH: Multilingual Matters.

Harlow, L. L. 1990. Do they mean what they say? Sociopragmatic competence and second language learners. *Modern Language Journal* 74, 328-51.

Hatch, E. & Long, M. H. 1980. Discourse analysis, what's that? In D. Larsen-Freeman (Ed.) *Discourse analysis in second language research.* Rowley, MA: Newbury House, 1-40.

Hellgren, P. 1982. *Communicative proficiency in a foreign language, and its evaluation.* Research report 2. Department of Teacher Education. Helsinki: University of Helsinki.

Hembree, R. 1988. Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research* 58, 47-77.

Henning, G. 1992. The ACTFL oral proficiency interview: validity evidence. *System* 20, 365-72.

Higgs T. V. & Clifford, R. 1983. The push toward communication. In T.V. Higgs (Ed.) *Curriculum, competence and the foreign language teacher.* Skokie, IL: National Textbook Company, 57-79.

Hiltunen, R. 1992. *Kansainvälinen englanti.* Virkaanastujaisesitelmä Turun yliopistossa 6.5.1992. [International English. Inauguration lecture at the University of Turku, May 5, 1992.]

Hirvonen, P. (1) 1971a, (2) 1971b, (3) 1973a, (4) 1973b, (5) 1974. *Englannin kielen taidon mittaaminen lukion päättyessä: 1. Lähtökohdat. 2. Kuuntelukoe. 3. Lukukoe ja kirjoituskoe. 4. Puhumiskoe. 5. Kielitaitokoe.* [Measuring English language proficiency at the end of senior secondary school: 1. Point of departure. 2. Listening comprehension. 3. Reading comprehension and writing. 4. Speaking. 5. General proficiency.]. Publications de l'Association Finlandaise de Linguistique Appliquée 1, 4, 7, 8, 9. Turku: AFinLA.

Holmes, J. & Brown, D. F. 1987. Teachers and students learning about compliments. *TESOL Quarterly* 21 (3), 523-46.

Horwitz, E. K., Horwitz, M. B. & Cope, J. 1986. Foreign language classroom anxiety. *Modern Language Journal* 70, 125-32.

House, J. & Kasper, G. 1981. Politeness markers in English and German. In F. Coulmas (Ed.) *Conversational routine*. The Hague: Mouton, 157-185.

Hughes, A. 1989. *Testing for language teachers*. Cambridge: Cambridge University Press.

Huhta, A. 1990. Suullisen kielitaidon arviointi ja arviointikriteerit [Assessment of the oral skill and the assessment criteria]. A mimeo. Jyväskylä: University of Jyväskylä. Language Centre for Finnish Universities.

Huhta, A. 1993. Suullisen kielitaidon arviointi [Assessing oral language proficiency]. In S. Takala (Ed.) *Suullinen kielitaito ja sen arviointi* [Nature and assessment of oral language proficiency]. Jyväskylä: University of Jyväskylä. Institute for Educational Research. Publication series B. Theory into Practice 77, 143-225.

Huhta, A. 1994. Suullisen kielitaidon arviointiasteikkojen validius [The validity of rating scales for testing oral proficiency]. An unpublished licentiate thesis. Jyväskylä: University of Jyväskylä. Department of Applied Linguistics.

Huhta, A., Sajavaara, K. & Takala, S. 1993. Recent developments in national examinations in Finland. In A. Huhta, K. Sajavaara & S. Takala (Eds.) *Language testing: new openings*. Jyväskylä: University of Jyväskylä. Institute for Educational Research, 136-59.

Hurley, D. S. 1992. Issues in teaching pragmatics, prosody, and non-verbal communication. *Applied Linguistics* 13 (3), 259-81.

Huttunen, I. & Kukkonen, L. 1995. *Englannin kielen oppimistuloksia peruskoulun 6. luokan valtakunnallisessa kokeessa* 1994 [Results of the 1994 nation-wide EFL test in the 6th grade]. Helsinki: National Board of Education.

Hymes, D. 1972. On communicative competence. In J. B. Pride and J. Holmes (Eds.) *Sociolinguistics*. Harmondsworth: Penguin, 269-93.

Jaakkola, H. 1997. *Kielitieto kielitaitoon pyrittäessä. Vieraiden kielten opettajien käsityksiä kieliopin oppimisesta ja opettamisesta [Language knowledge and language ability: Teachers' conceptions of the role of grammar in foreign language learning and teaching]*. Jyväskylä: University of Jyväskylä. Jyväskylä Studies in Education, Psychology and Social Research 128.

Jakobovits, L. & Gordon, B. 1979. Language teaching vs. the teaching of talk. *International Journal of Psycholinguistics* 6 (4), 5-22.

James, E. 1976. The acquisition of prosodic features using a speech visualizer. *International Review of Applied Linguistics in Language Teaching* 14, 227-43.

Janicki, K. 1982. *The foreigner's language in a sociolinguistic perspective*. Uniwersytet im. Adama Mickiewicza W Poznaniu. Seria filologigia angielska nr 17.

Johansson, S. 1978. *Studies of error gravity*. Native reactions to errors produced by Swedish learners of English. Gothenburg, Sweden: Acta universitatis gotheburgensis.

Järvinen, H. 1988. *Relative clauses in three different genres of present-day English*. Unpublished licentiate thesis. Department of English. Turku: University of Turku.

Kachru, B.B. 1976. Models of English for the Third World: white man's linguistic burden or language pragmatics? In A. Brown (Ed.) 1991. *Teaching English pronunciation. A book of readings*. London: Routledge, 31-52.

Kachru, B.B. 1992. World Englishes: approaches, issues and resources. State of the art article. *Language Teaching* 25, 1-14.

Kallmeyer, W. 1988. Konversationsanalytische Beschreibung. In U. Ammon, N. Dittmar & K. J. Mattheier (Eds.) *Sociolinguistics. Soziolinguistik*. Berlin: Walter de Gruyter, 1095-1107.

Kallmeyer, W. & Schütze, F. 1976. Konversationsanalyse. *Studium Linguistik* 1, 1-28.

Karppinen, M. & Sarkkinen, R. 1995. *Ruotsin kielen oppimistuloksia peruskoulun 9. luokan valtakunnallisessa kokeessa* 1994 [Results of the nation-wide Swedish FL test 1994 at the end of comprehensive school]. Helsinki: National Board of Education.

Karlsson, F. 1977. Morphotactic structure and word cohesion in Finnish. In K. Sajavaara & J. Lehtonen (Eds.) *Contrastive papers. Jyväskylä contrastive studies 4*. Jyväskylä: University of Jyväskylä. Reports from the Department of English, 59-74.

Karlsson, F. 1983. *Suomen kielen äänne- ja muotorakenne*. Helsinki: WSOY.

Keenan, E.O. 1976. The universality of conversational implicature. *Language in Society* 5, 67-80.

Keenan, E.O. & Bennett, T.L. (Eds.) 1977. *Discourse across time and space*. Southern California Occasional Papers in Linguistics 5.

King, R.D. 1967. Functional load and sound change. *Language* 43, 831-52.

Klatt, D.H. 1975. Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics* 3, 129-40.

Kohonen, V. 1987. *Towards experiential learning of elementary English* 1. A theoretical outline of an English and Finnish teaching experiment in elementary English. Tampere: University of Tampere. Reports from the Department of Teacher Training in Tampere A8.

Koponen, M. 1990. Glottal boundary markers and aspects of disfluency in Finnish English. Unpublished pro gradu thesis. Jyväskylä: University of Jyväskylä. Department of English.

Korpijaakko-Huuhka, A.-M. & Moore, K. 1992. Kuinka objektiivista on subjektiivisuus? Kuinka puheterapeutti arvioi puheen sujuvuutta? [How objective is subjective judgement? How does a speech therapist assess fluency?]. *Puheterapeutti* 2, 9-18.

Koskinen, M.-L. 1994. *Pienten ja keskisuurten yritysten kielitaito* [Language proficiency in small and medium-sized firms]. Helsinki: Fintra.

Kramsch, C. 1986. From language proficiency to interactional competence. Modern *Language Journal* 70, 366-72.

Kreckel, M. 1981. Tone units as message blocks in natural discourse: segmentation of face-to-face interaction by naive, native speakers. *Journal of Pragmatics* 5 (5), 459-76.

Kristiansen, I. 1990. *Nonverbal intelligence and foreign language learning*. Helsinki: University of Helsinki. Department of Education. Research bulletin 73.

Kristiansen, I. 1992. *Foreign language learning and nonlearning*. Helsinki: University of Helsinki. Department of Education. Research bulletin 82.

Lakoff, R. 1973. The logic of politeness; or, minding your p's and q's. In J. Corum (Ed.) *Papers from the Ninth Regional Meeting of the Chicago Linguistic Society*. Chicago: Chicago Linguistic Society, 292-305.

Lantolf, J. P. & Frawley, W. 1985. Oral-proficiency testing: a critical analysis. *Modern Language Journal* 69, 337-45.

Lantolf, J. P. & Frawley, W. 1988. Proficiency: understanding the construct. Studies in *Second Language Acquisition* 10, 181-95.

Larsen-Freeman, D. & Long, M. H. 1991. *An introduction to second language acquisition research*. London: Longman.

Laver, J. D. 1981. Linguistic routines and politeness in greeting and parting. In F. Coulmas (Ed.) *Conversational routine*. Explorations in standardized communication situations and prepatterned speech. The Hague: Mouton, 289-304.

Lazaraton, A. 1992. The structural organization of a language interview: a conversation analytic perspective. *System* 20, 373-86.

Leech, G. N. 1977. *Language and tact*. Trier: L.A.U.T. paper 46.

Leech, G. N. 1982. *English grammar for today*. A new introduction. London: Macmillan.

Leech, G. N. 1983. *Principles of pragmatics*. London: Longman.

Leeson, R. 1975. *Fluency and language teaching*. London: Longman.

Lehtonen, J. 1977. Contrastive phonetics and the analysis of speech communication. In K. Sajavaara & J. Lehtonen (Eds.) *Contrastive papers. Jyväskylä Contrastive Studies* 4. Jyväskylä: University of Jyväskylä, Reports from the Department of English, 31-44.

Lehtonen, J. 1978a. On factors affecting the pitch level of speech. In E. Görding, G. Bruce & R. Bannert (Eds.) *Nordic prosody*. Papers from a symposium. Lund, Sweden: University of Lund. Department of Linguistics, 55-63.

Lehtonen, J. 1978b. On the problems of measuring fluency. In M. Leiwo & A. Räsänen (Eds.) *AFinLA yerbook* 1978. Publications de l'Association Finlandaise de Linguistique Appliquée (AFinLA) 23. Jyväskylä: AFinLA, 53-68.

Lehtonen, J. 1979. Speech rate and pauses in the English of Finns, Swedish-speaking Finns, and Swedes. In R. Palmberg (Ed.) *Perception and production of English: papers on interlanguage*. AFTIL 6: Publications of The Department of English. Turku: Åbo Akademi.

Lehtonen, J. 1981. Problems of measuring fluency and normal rate of speech. In *Proceedings of the 5th congress of AILA*, Montreal, August 1978. Quebec: Les Presses de l'Université Laval, 322-331.

Lehtonen, J. 1985. Sprechangst und Fremdsprachenunterricht. *Finlance* 4, 141-51.

Lehtonen, J. 1990. Foreign language acquisition and the development of automaticity. In H. W. Dechert (Ed.) *Current trends in European second language acquisition research*. Cleveland, OH: Multilingual Matters.

Lehtonen, J. & M. Koponen. 1977. Signalling of morphophonological boundaries by Finnish speakers of English: Preliminary findings. In K. Sajavaara & J. Lehtonen (Eds.) *Contrastive papers. Jyväskylä Contrastive Studies* 4. Jyväskylä: University of Jyväskylä. Reports from the Department of English, 75-87.

Lehtonen, J. & Sajavaara, K. 1985. The silent Finn. In D. Tannen & M. Saville-Troike (Eds.) *Perspectives on silence*. Norwood, NJ: Ablex, 193-201.

Lehtonen, J., Sajavaara, K. & May, A. 1977. *Spoken English.* The perception and production of English on a Finnish-English contrastive basis. Jyväskylä: Gummerus.

Lehtovaara, J. 1978. *Peruskoulun kolmasluokkalaisten englannin kielen ääntämistaidosta* [On the EFL pronunciation skills of comprehensive school third-graders]. Julkaisusarja A: tutkimusraportti n:o 12. Tampere: University of Tampere. Publications of the Department of Education.

Lennon, P. 1990. Investigating fluency in EFL: a quantitave approach. *Language Learning* 40, 387-417.

Levelt, J. M. 1989. *Speaking.* From intention to articulation. Cambridge, MA: MIT Press.

Levinson, S. C. 1983. *Pragmatics.* Cambridge: Cambridge University Press.

Lindblad, T. 1992. Oral tests in Swedish schools: a five-year experiment. *System* 20, 279-92.

Linn, R., Baker, E. & Dunbar, S. 1991. Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher* 20, 5-21.

Liskin-Gasparro, J. E. 1985. The ACTFL proficiency guidelines: a historical perspective. In T.V. Higgs (Ed.) *Teaching for proficiency, the organizing principle.* Lincolnwoods, IL: National Textbook Company.

Liskin-Gasparro, J. E. 1989. Structure of the OPI. In K. Buck (Ed.) 1989. *The ACTFL Oral Proficiency Interview.* Chapter 4. *Tester training manual.* Yonkers, NY: The American Council on the Teaching of Foreign Languages.

Liskin-Gasparro, J. E. 1989. Structure of the OPI. In K. Buck (Ed.) *The ACTFL Oral Proficiency Interview. Tester training manual.* Yonkers, NY: The American Council on the Teaching of Foreign Languages. Chapter 4.

Local, J. K. 1985. Phonology for conversation - phonetic aspects of turn delimitation in London Jamaican. *Journal of Pragmatics* 9 (2/3), 309-30.

Loundsbury 1954. See Lennon 1990, 393.

Lowe, P., Jr. 1983. The ILA oral interview: origins, applications, pitfalls, and implications. *Unterrichtspraxis* 16 (2), 230-44.

Lowe, P. & Clifford, R. T. 1980. Developing an indirect measure of overall oral profiency. J. R. Frith (Ed.) *Measuring spoken language proficiency.* Washington DC: Georgetown University Press.

Lowe, P., Jr. & Liskin-Gasparro, J. E. 1986. *Testing speaking proficiency: the oral interview.* Q & A. ERIC clearinghouse on languages and linguistics.

Ludwig, J. 1982. Native speaker judgements of second language learners' efforts at communication: A review. *Modern Language Journal* 66, 274-83.

Luoma, S. 1997. Comparability of a tape-mediated and a face-to-face test of speaking. A triangulation study. An unpublished licentiate thesis. Jyväskylä: University of Jyväskylä. Centre for Applied Language Studies.

Luoma, S. & Takala, S. 1993. Aikuisten kielitaitotutkinto [The language proficiency examination for adults]. In S. Tella (Ed.) *Kielestä mieltä - mielekästä kieltä* [Toward meaningful language]. Ainedidaktiikan symposiumi 5.2.93. Osa 2. Helsinki: University of Helsinki. Research Reports from the Teacher Education Department 118, 212-25.

Lyons, J. 1977. *Semantics.* Cambridge: Cambridge University Press.

McLaughlin, B. 1990. Restructuring. *Applied Linguistics* 11 (2), 113-28.

McNamara, T. 1996. *Measuring language performance.* London: Longman.
Madsen, H. S. 1982. Determining the debilitative impact of test anxiety. *Language Learning* 32, 133-43.
Madsen, H. S., Brown B. L. & Jones R. L. 1991. Evaluating student attitudes toward second-language tests. In E.K. Horwit & D.J. Young (Eds.) *Language anxiety.* Englewood Cliffs, NJ: Prentice Hall, 65-86.
Magnan, S. S. 1987. Rater reliability of the ACTFL Oral Proficiency Interview. *Canadian Modern Language Review* 43 (3), 525-37.
Magnan, S. S. 1988. Grammar and the ACTFL Oral Proficiency Interview: discussion and data. *Modern Language Journal* 72, 266-76.
Maley, A. 1987. Foreword. In R. Nolasco & L. Arthur, *Conversation.* Oxford: Oxford University Press, 3.
Manes, J. & Wolfson, N. 1981. The compliment formula. In F. Coulmas (Ed.) *Conversational routine.* Explorations in standardized communication situations and prepatterned speech. The Hague: Mouton, 115-132.
Maude, G. 1980. Role playing in foreign language learning. *Tempus* 8, 10-12.
Maurice, K. 1983. The fluency workshop. *Tesol Newsletter* 3, 29.
Meredith, R. A. 1990. The oral proficiency interview in real life: sharpening the scale. *Modern Language Journal* 74, 288-96.
Messick, S. 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: Erlbaum, 33-45.
Messick, S. 1989. Meaning and values in test validation: the science and ethics of assessment. *Educational Researcher* 19, 5-11.
Mikes, G. 1977 (1946). *How to be an Alien.* Harmondsworth: Penguin.
Moisio, R. & Valento, E. 1976. *Testing Finnish school-children's learning of English consonants. Jyväskylä Contrastive Studies 3.* Jyväskylä: University of Jyväskylä. Reports from the Department of English.
Moore, K. 1990. On prosodic elements in television and radio sports narrations. *Nordic Prosody* 5, 1-17.
Moore, K. 1991a. Speech rate, phonation rate, and pauses in cartoon and sports narrations. In R. Aulanko & M. Leiwo (Eds.) *Studies in logopedics and phonetics 2.* Helsinki: University of Helsinki. Publications of the Department of Phonetics. Series B: Phonetics, logopedics and speech communication 3, 135-143.
Moore, K. 1991b. A taxonomy of pauses in Finnish. In R. Aulanko & M. Leiwo (Eds.) *Studies in logopedics and phonetics 2.* Helsinki: University of Helsinki. Publications of the Department of Phonetics. Series B: Phonetics, logopedics and speech communication 3, 145-150.
Morrow, K. 1979. Communicative language testing: revolution or evolution? In C. Brumfit & K. Johnson (Eds.) *The communicative approach to language teaching.* London: Oxford University Press, 143-57.
Moss, P. 1992. Shifting conceptions of validity in educational measurement: implications for performance assessment. *Review of Educational Research* 62, 229-258.
Moss, P. 1994. Can there be validity without reliability? *Educational Researcher* 23, 5-12.

Munby, J. 1978. *Communicative syllabus design: a sociolinguistic model for defining the content of purpose-specific language programmes.* New York: Cambridge University Press.

Munro, M.J. & Derwing T.M. 1994. Evaluations of foreign accent in extemporaneous and read material. *Language Testing* 11, 253-266.

Newbrook, M. 1986. Received pronunciation in Singapore: a sacred cow? In A. Brown (Ed.) 1991. *Teaching English pronunciation. A book of readings.* London: Routledge, 72-86.

Nolasco, R. & Arthur, L. 1987. *Conversation.* Oxford: Oxford University Press.

Norris, C. B. 1991. Evaluating English oral skills through the technique of writing as if speaking. *System* 19, 203-16.

North, B. 1993. Transparency, coherence, and washback in language assessment. In K. Sajavaara, S. Takala, R. D. Lambert & C. A. Morfit (Eds.) *National foreign language planning: practices and prospects.* Jyväskylä: University of Jyväskylä. Institute for Educational Research, 157-93.

O'Connor, D. J. & Arnold, G.F. 1959. *Intonation of colloquial English.* London: Longman.

O'Donnell, R.C. 1974. Syntactic differences between speech and writing. *American Speech* 49, 102-10.

Oksaar, E. 1988. *Kulturemtheorie.* Ein Beitrag zur Sprachverwendungsforschung. Hamburg: Joachim Jungius Gesellschaft der Wissenschaften.

Oller, J. W., Jr. 1979. *Language tests at school.* London: Longman.

Oller, J. W., Jr. 1983. A consensus for the '80s? In J. W. Oller, Jr. (Ed.) *Issues in language testing research.* Rowley, MA: Newbury House, 351-6.

Olynyk, M., d'Anglejan, A. & Sankoff, D. 1990. A quantitative and qualitative analysis of speech markers in the native and second language speech of bilinguals. In R. C. Scarcella, E. S. Andersen & S. D. Krashen (Eds.) *Developing communicative competence in a second language.* Rowley, MA: Newbury House, 139-55.

Olynyk, M., Sankoff, D. & d'Anglejan, A. 1983. Second language fluency and the subjective evaluation of officer cadets in a military college. *Studies in Second Language Acquisition* 5 (2), 213-249.

Owen, M. 1990. Language as a spoken medium: conversation and interaction. In N. E. Collinge (Ed.) *An encyclopaedia of language.* London: Routledge, 244-80.

Pasanen, U.-M. 1977. *Päättökokeen uudistuksen heijastuminen lukion vieraan kielen opetukseen opettajien arvioimana.* [How the matriculation examination reform is reflected in the senior secondary school FL teaching, as assessed by teachers]. Jyväskylä: University of Jyväskylä. Research Reports 57. Department of Education.

Pasanen, U.-M. & Hietanen, A. 1994. *Peruskoulun 9. luokan valtakunnallinen englannin kielen koe ja sen kehittäminen.* [The nation-wide EFL test at the end of comprehensive school and its development]. Jyväskylä: University of Jyväskylä. Research Report 3. Department of Education.

Pawley, A. & Syder, F. H. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.) *Language and communication.* London: Longman, 191-226.

Pennington, M. & Richards, J. C. 1986. Pronunciation revisited. TESOL Quarterly 20 (2), 207-25.

Piirainen, Marsh, A. 1995a. *Face in second language conversation*. Jyväskylä: University of Jyväskylä.

Piirainen-Marsh, A. 1995b. Politeness and facework. *Tempus* 7, 6-8.

Pimsleur, P., Hancock C. & Furey P. 1977. Speech rate and listening comprehension. In M. Burt, H. Dulay & M. Finocchiaro (Eds.) *Viewpoints on English as a second language*. New York: Regents. Sivut ?

Pohjala, K. 1995. Suullisen kielitaidon opettamisesta ja arvioinnista lukiossa. In I. Huttunen, R. Paakkunainen & K. Pohjala, *Suullisen kielitaidon opetus ja arviointi lukiossa.* [The teaching and assessing of oral proficiency in senior secondary school]. Oulu: University of Oulu. Reports from the faculty of education 95, 11-14.

Politzer, R. 1978. Errors of English speakers as perceived by German natives. *The Modern Language Journal* 62, 253-61.

Pollitt, A. 1991. Giving students a sporting chance: assessment by counting and by judging. In J. C. Alderson. & B. North (Eds.). *Language testing in the 1990s*. London: Modern English Publications.

Prodromou, L. 1995. The backwash effect: from testing to teaching. *ELT Journal* 49 (1), 13-25.

Raffaldini, T. 1988. The use of situation tests as measures of communicative ability. *Studies in Second Language Acquisition* 10 (2), 197-216.

Raith, J. 1984. Die Funktion und Relevanz prosodischer Systeme im Interaktionsprozess. *Die neueren Sprachen* 83 (5), 513-44.

Rehbein, J. 1987. On fluency in second language speech. In H. W. Dechert & M. Raupach (Eds.) *Psycholinguistic models of production*. Norwood, NJ: Ablex, 97-105.

Reves, T. 1991. From testing research to educational policy: a comprehensive test of oral proficiency. In Alderson, J. C. & North, B. *Language testing in the 1990s*. London: Modern English Publications, 178-88.

Richards, J. 1980. Conversation. *TESOL Quarterly* 14 (4), 413-32.

Richards, J. 1983. Communicative needs in foreign language learning. *ELT Journal* 37 (2), 111-20.

Richards, J., Platt, J. & Weber, H. 1985. *Longman dictionary of applied linguistics*. Harlow: Longman.

Richards, J .C. & Schmidt, R. W. 1983 Conversational analysis. In J. C. Richards & R. W. Schmidt (Eds.) *Language and communication*. London: Longman, 117-54.

Riggenbach, H. 1991. Toward an understanding of fluency: a microanalysis of nonnative speaker conversations. *Discourse processes* 14, 423-41.

Rivers, W. M. & Temperley, M. S. 1978. *A practical guide to the teaching of English as a second or foreign language*. New York: Oxford University Press.

Ryan, E. B. 1983. Social psychological mechanisms underlying native speaker evaluations of non-native speech. *Studies in Second Language Acquisition* 5 (2), 148-59.

Sajavaara, K. 1977. Contrastive linguistics past and present and a communicative approach. In K. Sajavaara & J. Lehtonen (Eds.) *Contrastive papers. Jyväskylä*

constrasitve studies 4. Jyväskylä: University of Jyväskylä. Reports from the Department of English.

Sajavaara, K. 1987. Second language speech production: factors affecting fluency. In H. W. Dechert & M. Raupach (Eds.) *Psycholinguistic models of production*. Norwood, NJ: Ablex, 45-65.

Sajavaara, K. 1988. Cross-linguistic and cross-cultural intelligibility. In P. H. Lowenberg (Ed.), *Georgetown University Round Table on Languages and Linguistics 1987* (GURT '87). Language spread and language policy: issues, implications, and case studies. Washington, DC: Georgetown University Press, 250-64.

Sajavaara, K. & Lehtonen J. 1978. Spoken language and the concept of fluency. In L. Lautamatti & P. Lindqvist (Eds.) *Focus on spoken language*. Language Centre News. Special Issue 1. Jyväskylä: University of Jyväskylä. Language Centre for Finnish Universities, 23-57.

Sajavaara, K. & J. Lehtonen. 1980. The analysis of cross-language communication: prolegomena to the theory and methodology. In H. W. Dechert & M. Raupach (Eds.) *Towards a cross-linguistic assessment of speech production*. Frankfurt a. M.: Peter D. Lang.

Sajavaara, K. & Lehtonen, J. 1985. The silent Finn. In D. Tannen & M. Saville-Troike. (Eds.) *Perspectives on silence*. Norwood, NJ: Ablex, 193-201.

Sajavaara, K. & Lehtonen, J. 1997. The silent Finn revisited. In A. Jaworski (Ed.) *Silence: interdisciplinary perspectives*. Berlin: Mouton de Gruyter, 263-283.

Saleva, M. 1993. Ceterum censeo: suullisen taidon opetusta olisi lisättävä [Ceterum censeo: more emphasis should be paid to the speaking skill]. In S. Takala (Ed.) *Suullinen kielitaito ja sen arviointi* [The speaking skill and how to assess it]. Jyväskylä: University of Jyväskylä. Institute for Educational Research. Publication series B. Theory into Practice 77, 1-14.

Savignon, S. J. 1985. Evaluation of communicative competence: The ACTFL Provisional Proficiency Guidelines. *Modern Language Journal* 69, 129-134.

Schaffer, D. 1983. The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics* 11 (3), 243-57.

Schaffer, D. 1984. The role of intonation as a cue to topic management in conversation. *Journal of Phonetics* 12 (4), 327-44.

Scollon, R. 1985. The machine stops. In D. Tannen & M. Saville-Troike (Eds.) *Perspectives on silence*. Norwood, NJ: Ablex.

Scollon, R. & Scollon, S. 1983. Face in interethnic communication. In J. Richards & J. Schmidt (Eds.) *Language and communication*. London: Longman.

Searle, J. 1969. *Speech acts*. Cambridge: Cambridge University Press.

Seelye, H. N. 1974. Teaching culture. Skokie, IL: National Textbook Company.

Selting, M. 1988. The role of intonation in the organisation of repair and problem handling sequences in conversation. *Journal of Pragmatics* 12 (3), 293-322.

Shohamy, E. 1982a. Affective considerations in language testing. *The Modern Language Journal* 66, 13-17.

Shohamy, E. 1982b. Predicting speaking proficiency from cloze tests. *Applied Linguistics* 3, 161-71.

Shohamy, E. 1983. The stability of oral proficiency assessment on the oral interview testing procedures. *Language Learning* 33, 527-40.

Shohamy, E. 1984. Does the testing method make a difference? The case of reading comprehension. *Language Testing* 1, 147-159.

Shohamy, E. 1988. A proposed framework for testing the oral language of second/foreign language learners. *Studies in Second Language Acquisition* 10, 165-179.

Shohamy, E. 1992. Discourse validation of direct vs. semi direct oral tests. Paper presented at the 14th Language Testing Research Colloquium, Vancouver, February 27th - March 1st, 1992.

Shohamy, E. 1993a. The effect of the elicitation mode on the language samples obtained on oral tests. Paper presented at the Language Testing Research Colloquium, Cambridge, England, August, 1993.

Shohamy, E. 1993b. The exercise of power and control in the rhetorics of testing. In A. Huhta, K. Sajavaara & S. Takala (Eds.) *Language testing: new openings.* Jyväskylä: University of Jyväskylä. Institute of Educational Research, 23-38.

Shohamy, E. 1993c. The power of tests: the impact of language tests on teaching and learning. NFLC Occasional Papers, June 1992, 1-19.

Shohamy, E. 1994. The validity of direct versus semi-direct oral tests. *Language Testing* 11, 99-124.

Shohamy, E. & Reves, T. 1985. Authentic language tests: where from and where to? *Language Testing* 2, 48-59.

Shohamy, E., Reves, T. & Bejarano, Y. 1986. Introducing a new comprehensive test of oral proficiency. *ELT Journal* 40 (3), 212-20.

Shohamy, E., Shmueli, D. & Gordon, C. 1991. The validity of concurrent validity of a direct vs. semi direct test of oral proficiency. Paper presented at the 13th Language Research Colloquium, Educational Testing Service, Princeton, March 21-23, 1991.

Shohamy, E. & Stansfield, S. W. 1990. The Hebrew speaking test: an example of international cooperation in test development and validation. *Aila Review* 7. Standardization in Language Testing, 79-90.

Spolsky, B. 1973. What does it mean to know a language; or how do you get someone to perform his competence? In Oller, J.W., Jr. & Richards, J.C. (Eds.) *Focus on the learner: pragmatic perspectives for the language teacher.* Rowley, Mass.: Newbury House, 164-176.

Spolsky, B. 1986. A multiple choice for language testers. *Language Testing* 3, 147-158.

Spolsky, B. 1993. Testing and examinations in a national foreign language policy. In K. Sajavaara, S. Takala, R. D. Lambert & C. A. Morfit (Eds.) *National foreign language planning: practices and prospects.* Jyväskylä: University of Jyväskylä. Institute for Educational Research, 194-211.

Stansfield, C. W. 1989. *Simulated oral proficiency interviews.* ERIC Digest. Washington, DC: ERIC Clearinghouse on Languages and Linguistics.

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

Stansfield, C. W. 1990b. Some foreign language test development priorities for the last decade of the twentieth century. *Foreign Language Annals* 23, 395-401.

Stansfield, C. W. & D. M. Kenyon. 1992a. The development and validation of a simulated oral proficiency interview. *Modern Language Journal* 76, 129-41.

Stansfield, C. W. & Kenyon D. M. 1992b. Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System* 29, 347-64.

Starkweather, C. W. 1987. *Fluency and stuttering*. Englewood Cliffs, NJ: Prentice-Hall.

Sternberg, R. J. 1985. *Beyond IQ: a triarchic theory of human intelligence*. New York: Cambridge University Press.

Sternberg, R. J. 1988. *The triarchic mind: a new theory of human intelligence*. New York: Viking.

Strevens, P. 1974. A rationale for teaching pronunciation: the rival virtues of innocence and sophistication. In A. Brown (Ed.) 1991. *Teaching English pronunciation. A book of readings*. London: Routledge, 96-103.

Strevens, P. 1981. What is 'Standard English'? *RELC Journal* 12 (2), 1-9.

Suomi, K. 1976. *English voiceless and voiced stops as produced by native and Finnish speakers. Jyväskylä Contrastive Studies* 2. Jyväskylä: University of Jyväskylä. Reports from the Department of English.

Suontausta, T. 1993. *Katsaus kielitaidon tarvetutkimuksiin*. Muistio [A survey of literature on foreign language training needs. An unpublished memo]. Jyväskylä: University of Jyväskylä. Language Centre ofr Finnish Universities.

Suullisen kielitaidon kokeen työryhmän raportti [Report of the oral proficiency test working group]. 1989. Helsinki: National Board of Education.

Swales, J.M. 1990. *Genre analysis*. English in academic and research settings. Cambridge: Cambridge University Press.

Takala, S. 1977. *Piirteitä vieraiden kielten opetuksesta* [Some aspects of FL teaching]. Jyväskylä: University of Jyväskylä. Institute for Educational Research. Bulletin 94.

Takala, S. 1983. The domain of writing. In S. Takala & A. Vähäpassi. *On the specification of the domain of writing*. Jyväskylä: University of Jyväskylä. Institute for Educational Research. Bulletin 333, 1-61.

Takala, S. 1993. Suullisen kielitaidon testaaminen osana kielten ylioppilastutkintoa? [Should the matriculation examination include a speaking test?] In S. Takala (Ed.) *Suullinen kielitaito ja sen arviointi* [The speaking skill and how to assess it]. Jyväskylä: University of Jyväskylä. Institute for Educational Research. Publication series B. Theory into Practice 77, 27-33.

Takala, S. & Saari, H. 1979. *Englannin kielen opetus ja koulusaavutukset Suomessa 1970-luvun alussa: IEA:n kansainväliseen yhteistoimintaan perustuva vertaileva ja kuvaileva tutkimus* [The teaching of English in Finland at the beginning of the 1970's: a comparative and descriptive study based on the IEA data]. Jyväskylä: University of Jyväskylä. Institute for Educational Research.

Tannen, D. 1982. *Spoken and written language: exploring orality and literacy*. Norwood, NJ: Ablex.

Tannen, D. 1984. The pragmatics of cross-cultural communication. *Applied Linguistics* 5 (3), 189-95.

Tannen, D. 1985. Silence: anything but. In D. Tannen & M. Saville-Troike, M. (Eds.) *Perspectives on silence*. Norwood, NJ: Ablex, 93-111.

Tauroza, S. & Allison, D. 1990. Speech rates in British English. *Applied Linguistics* 11 (1), 90-105.

Taylor, D. S. 1981. Non-native speakers and the rhythm of English. In A. Brown (Ed.) 1991. *Teaching English pronunciation. A book of readings*. London: Routledge, 235-44.

Thomas, J. 1983. Cross-cultural pragmatic failure. Applied Linguistics 4 (2), 91-111.

Thompson, I. 1991. Foreign accents revisited: the English pronunciation of Russian immigrants. *Language Learning* 41, 177-204.

Tiittula, L. 1992. *Puhuva kieli.* Suullisen kielen erityispiirteitä. [Spoken language. Special features of spoken language.] Loimaa: Finn Lectura.

Tuomiharju, L. 1993. Testaaminen ei lisää suullista kielitaitoa [Testing does not improve the oral skill]. *Tempus 4*, 5.

Underhill, N. 1987. *Testing spoken language: a handbook of oral testing techniques*. Cambridge: Cambridge University Press.

Valdman, A. 1987. Introduction. In A. Valdman (Ed.) *Proceedings of the symposium on the evaluation of foreign language proficiency*. Bloomington: Indiana University, xiii-xxiv.

Vanderplank, R. 1993. 'Pacing' and 'spacing' as predictors of difficulty in speaking and understanding English. ELT Journal 47 (2), 117-24.

van Lier, L. 1996. *Interaction in the language curriculum. Awareness, autonomy & authenticity*. London: Longman.

Varadi, T. 1990. Disfluency phenomena in L2 speech. Paper read at the 9th World Congress of Applied Linguistics, Thessaloniki, Greece, 21st April, 1990.

Varonis, E. M. & Gass S. M. 1982. The comprehensibility of nonnative speech. *Studies in Second Language Acquisition* 4, 114-36.

Vollmer, H. J. 1983. *Spracherwerb und Sprachbeherrschung: Untersuchungen zur Struktur von fremdsprachenfachigorientierten Sprachlehr-/lernforschung*. Tübingen: Günter Narr. (Tübingen Beiträge zur Linguistik).

Wardhaugh, R. 1985. *How conversation works*. Oxford: Basil Blackwell.

Weir, C. 1988. *Communicative language testing*. (2nd rev. ed.) Exeter: University of Exeter. Exeter Linguistic Studies, Volume 11.

Weir, C. 1989. Approaches to language test design: a critical review. Mimeo, distributed on the British Council course The testing of oral interaction: principles and practice, Reading, 29 March - 11 April 1989.

Weir, C. 1990. *Communicative language testing*. New York: Prentice-Hall.

Wesche, M. B. 1983. Communicative testing in a second language. *Modern Language Journal* 67, 41-55.

West, M. 1952. Examinations in a foreign language. *English Language Teaching* 6 (2), 60-3.

Widdowson, H.G. 1979. *Explorations in applied linguistics*. Oxford: Oxford University Press.

Wiik, K. 1965. *Finnish and English vowels*. Turku: University of Turku. Annales universitatis turkuensis, Series B: 94.

Wine, J. D. 1980. Cognitive-attentional theory of test anxiety. In I. G. Sarason (Ed.) *Text anxiety: Theory, research and applications*. Hillsdale, NJ: Erlbaum, 349-385.

Wolfson, N. 1981a. Compliments in cross-cultural perspective. *TESOL Quarterly* 15 (2), 127-24.

Wolfson, N. 1981b. Invitations, compliments and the competence of the native speaker. *International Journal of Psycholinguistics* 8 (4), 7-22.

Wolfson, N. 1983a. An empirically based analysis of complimenting in American English. In N. Wolfson & E. Judd (Eds.) *Sociolinguistics and language acquisition. Rowley*, MA: Newbury House, 82-95.

Wolfson, N. 1983b. Rules of speaking. In J. Richards & J. Schmidt (Eds.) *Language and communication*. London: Longman, 61-87.

Wong Fillmore, L. 1979. Individual differences in second language acquisition. In C. J. Fillmore, D. Kempler & W. Wang (Eds.), *Individual differences in language ability and language behavior*. New York: Academic Press, 203-228.

Yli-Renko, K. 1985. *Lukion saksan kielen opetuksen tavoitteet* [The aims of German language teaching in the upper secondary school in Finland]. Helsinki: University of Helsinki. Research Bulletin 38. Department of Teacher Education.

Yli-Renko, K. 1988. *Assessing foreign language training needs of adults*. Helsinki: University of Helsinki. Research Bulletin 67. Department of Education.

Yli-Renko, K. 1989a. *Intercultural communication as an aim of English language teaching*. Helsinki: University of Helsinki. Research Bulletin 69. Department of Education.

Yli-Renko, K. 1989b. *Suullinen kielitaito ja sen mittaaminen lukion päättövaiheessa* [The oral language proficiency and its assessment at the end of senior secondary school]. Helsinki: University of Helsinki. Research Bulletin 72. Department of Teacher Education.

Yli-Renko, K. 1991. Suullisen kielitaidon oppiminen lukiossa: oppilaiden näkökulma [Learning oral language proficiency in the senior secondary school: the students' view]. In K. Yli-Renko & L. Salo-Lee, *Oral communication proficiency and its learning in senior secondary school*. Turku: University of Turku. Department of Education, Series A, 147, 25-58.

Young, D. J. 1986. The relationship between anxiety and foreign language oral proficiency ratings. *Foreign Language Annals* 19, 439-45.

APPENDIX 1

The ACTFL OPI Role Cards

The ACTFL oral interview role cards: INTERMEDIATE[1]

You are at a little Greek restaurant on Rhodes. There is no written menu. Ask the waiter such questions that you will be able to order a meal.

You are at a travel agency in London and you want to take a tour of the city. Ask 4-5 questions about the tour.

You are in London and want to go to see a friend who is lying at hospital after a motor accident. You go to a florist's to buy her some flowers. Explain the situation to the florist, ask some questions about the flowers, and buy some.

You and your friend have just arrived in England. You want to rent a car and go to a rental company. Ask some questions about the rental conditions.

I come from Turku. Ask me some questions about the place.

I spend my summers in Kustavi. Ask me some questions about the place.

I have just come back from a trip abroad. Ask me questions about the trip.

You are a reporter for the school journal. I am Jari Kurri, and you have arranged to interview me. Ask me questions about things you want to know so that you can write an article for your journal.

Phone to your friend and ask her/him to come to the cinema with you. Try to come to an agreement about the film, the time you meet and what you are going to do after the cinema.

Your neighbor is going on a holiday. She asks you to look after her house while she will be away. Ask her four to five questions about what you need to know and do.

You are in your friend's home waiting for her/him. You are left alone with her/his grandmother and try to keep her company while you are waiting. Begin a conversation with her and go on with it.

You go to an agency in London and want to rent a one-room flat. Explain to them what sort of place you are looking for. Ask them some questions about the flat that they are offering to you.

You have promised to go baby-sitting for a young couple who has moved to your neighborhood. They are just leaving, but before they go, you have time to ask them some questions about the children and your duties tonight.

You are at an English post office. You want to send a book to Finland.

---

1 The cards were given to the testees in a Finnish version

Ask the clerk some questions to find out how you can send it fast and cheaply. Find also out whether you have to pay some duty.

The ACTFL oral interview role cards: ADVANCED

There have been housebreakers in your home. Call the police, explain what has happened, describe what you saw, and find out what you must do.

You have just had a good meal at a restaurant. Now you notice that you have left your wallet at home. Explain the situation to the waiter and try to come to an agreement about paying the bill.

You have bought a pair of shoes. When you try them on at home, you discover that they do not seem to fit well and that you do not actually like their style either. Go back to the shop and try to get your money back.

You are taking your jacket back to the dry cleaner's. Explain to the shop assistant why you are bringing the jacket back and why she has to clean it again without charging you. Make it also clear to her why she had better do the job well.

You borrowed your friend's car and had a little accident with it. Phone your friend, tell what happened, describe what the car looks like now, and offer to make up for what happened.

You were present at a little motor accident. No one was seriously injured. Phone the police, describe how the accident happened, describe the present situation, and find out what you have to do now.

You stayed on at your job after the others had left. As you went to a shop to buy something to drink, you locked yourself out by mistake. You have no identification on you. Go to the caretaker, explain the situation, and ask him to come and let you in.

You have foreign guests at your summer place. They have never been to a sauna, but they want to take it by themselves. Explain to them how to go about it.

You are in a foreign city. Your wallet was stolen two days ago. You go to report it to the police station. Explain where and how it happened and why you come to report it so much later.

You were walking in the street as a woman walking in front of you at the pedestrian crossing was hit by a sports car. Tell me what you saw and what you did then.

Last night you took your English guest to a foreign restaurant. You were very dissatisfied with the place. Phone now to the owner and complain politely about, for instance, slow and impolite service, dirty tableware, cold and tasteless food, loud music, price.

The ACTFL oral interview role cards: SUPERIOR

You are being interviewed for a public opinion poll about the question of energy. Tell the interviewer your opinion and also your attitude to the need of a fifth nuclear power plant.

Your school is being visited by a group of Pakistani students. One of them asks you about the Finns' attitudes to overseas aid for the less developed countries (pro and contra). Tell her about different opinions, present and justify also your own opinion.

You are an exchange student in the USA. You belong to the school choir, and it is an important and pleasant hobby to you. The choir rehearses twice a week. Today it is the third time within two weeks that you cannot go to the rehearsal. Explain this to the choirmaster and try to convince him that you are, nevertheless, a responsible and reliable member.

At a party you meet a Greek student. He has been traveling about in Finland and claims that the youth in Finland is given too much freedom (for instance, in matters like money, accommodation, alcohol, sex). Explain to him the Finnish way of thinking, present your own opinion and justify it.

You have been invited to some friends. There you meet some Americans. One of them asks you about the equality of the sexes in Finland and whether there still are some injustices. Answer to him and justify your opinion.

EITHER
You talk about the membership of the EU with your Norwegian friend. Discuss the pros and the cons and justify your own opinion.
OR
You are talking with your French friend about becoming more international versus preserving your own national culture. Which is more important? Can you do both at the same time? Present and justify your opinion.

There are some American-Finnish relatives visiting your family. There is a discussion about unemployment. One of the visitors claims that almost all of those who receive unemployment benefit are only lazybones who live at other people's expense. Present and justify your opinion about unemployment in Finland.

During the English lesson there is a panel discussion about women going into military service or doing some compensatory work. You have been asked to present the opening address and an introduction to the subject. After this, state and justify your own opinion.

You are at an international language school in England. Today's subject is social welfare. The pupils in your group all tell about social security in their countries. It is now your turn. Tell about social security in Finland. Present and justify your opinion of whether you think it is insufficient or just right or excessive.

At a Scandinavian youth meeting the present theme is young people's unemployment. Should Finnish society do more for the young unemployed?

If so, what? Where should one get the money? Present your opinion to the Nordic members of your group.

You are at an international language school. The theme for the discussion today is the different family structure in different countries. In some societies families are big, representing different generations whereas other societies have mainly a small nuclear family. Tell about the Finnish situation now and before. Mention some good and bad sides of both types. Explain which you prefer yourself and why.

Your school has arranged a joint project in geography/science/history/English with the theme of environmental protection. The difficulty of finding a balance between economic growth and environmental protection is discussed. Describe the problem, take a stand, and justify your opinion.

Women's magazines are full of weight curves and low-calorie diets. Some people think that trying to lose weight has become the new hysterics in welfare societies. Express and justify your opinion about the matter in the English debating club.

APPENDIX 2

The Attitude Test

UNIVERSITY OF TURKU
Department of Teacher Education in Turku

Oral Testing Project

The aim of this questionnaire is to collect information for developing a speaking test. The answers to the questions are of great importance for the compiling of the test. It would therefore be very valuable that you should answer all questions carefully. Your answer will be dealt with absolute confidence. Your name is used only for the purpose of linking the information which you have given with your performance in the speaking test. Thank you for your cooperation!

Maija Saleva
Researcher

NAME_____

1. Sex a. female b. male

2. How many years have you studied English at school?_____

3. Have you studied English on a language course abroad?
   a. No b. Yes
   If you answered yes, how many weeks altogether have you studied abroad?
   _____

4. Have you been to an English-speaking country except on a language course?
   a. No b. Yes.
   If you answered yes, how many weeks or months altogether?
   _____months, _____weeks.

5. Have you had any activities outside school in which you have used your knowledge of English (e.g. studies at a 'summer gymnasium' or a 'summer university', reading books or magazines, etc.)?

a. No
b. Yes What?/How often (much)?

In the following some claims are presented, which you should react to. Please, circle the alternative which is nearest to your opinion. Use the alternative ? (= difficult to say only when you really do not have any opinion about the matter. The alternative answers are the following:

++ = I absolutely agree
+ = I agree
? = difficult to say
- = I do not agree
-- = I absolutely disagree

6. Outside school I like to seek opportunities to speak English.
++ + ? - --

7. At school I do not like to speak English, because I am afraid of pronouncing incorrectly or making other mistakes.
++ + ? - --

8. I do not like to talk to strange people even in Finnish.
++ + ? - --

9. There should be more speaking exercises in the English lessons.
++ + ? - --

10. Also in the Finnish lessons more time should be spent on developing the speaking skill.
++ + ? - --

11. My English teacher speaks too little English in the lessons.
++ + ? - --

12. I find speaking exercises very unpleasant.
++ + ? - --

13. I find speaking exercises very difficult.
++ + ? - --

14. If time spent on speaking exercises is increased, should practicing some other skill be diminished correspondingly?
a. No b. Yes.
If you answered yes, tell what should be diminished (you can also circle several letters):

a. writing exercises
b. listening exercises
c. reading texts
d. studying grammar
e. studying vocabulary
f. studying culture (life and customs in English-speaking countries)
g. something else. What?_____

15. I find the language laboratory test easy.
++ + ? - --

16. Participating in the language laboratory test was a pleasant experience.
++ + ? - --

17. I find the English interview easy.
++ + ? - --

18. Participating in the English interview was a pleasant experience.
++ + ? - --

19. I think the matriculation examination should contain an oral test of English.
++ + ? - --

20. If the matriculation examination contained an oral test, the teaching of English in the senior secondary school would be more efficient.
++ + ? - --

21. The oral test of English in the matriculation examination should be a language laboratory test.
++ + ? - --

22. The oral test of English in the matriculation examination should be an interview.
++ + ? - --

23. If in the matriculation examination I should participate in either a language laboratory test or an interview, I would rather participate in the language laboratory test.
++ + ? - --

24. I should like to comment on the different subtests in the language laboratory test in the following way:

a. Reading aloud a letter

_____

_____

b. Interpreting mother's words

c. Interpreting the newspaper article

d. Presenting the Finnish school system

e. Reacting in various situations, expressing opinions

25. The English mark in my latest school report

26. For oral skills in English I should give myself mark

27. For this test I should give myself mark

28. For the oral interview I should give myself mark

APPENDIX 3

Transcription of Student 49'S LLOPT Test

A transcription of student NN's speech in subsections 0, 2, 4, and 5

Subtest 0

About nine or ten years.

No, it's not easy, but it's not a difficult, either.

English.

Well, at my sparetime I...I do all kinds of sports...I am a junior hockey teach coach...I run... I play bandy and also I read and listen to music... and study, of course.

Well, I like all kinds of music, classic music, pop music, disco music...er...I don't' any...I don't hate any kind of music.(70 words)

Subtest 2

Hi Mike. This is my mother, X, and Mother, this is Mike.

It's very nice to meet you. I don't speak any English but you can come here to sit and talk...with me.

You must be tired. When did you leave Cincinnati?

What would you like to eat now?

Where actually in Finland...do you have...played some concerts?

How good was the success?

What instrument do you play?

When I was young, I played the trumpet. How long have you...how long have you played?

How many times...have you practice...do you have to practice in day?

Oh, look at the time...mm...you probably want to see your room.

Tomorrow afternoon we will take you to see some of the sights of Härmälä. (*The wish is lacking.*) (127 words.)

Subtest 3

See 10.1.1 Transmitting information

Subtest 4

Well, I will tell you something about the structure of the Finnish school system. When you're...one years...to fif...five years old, you go to kindergarden, then you go to some kind of prep school when you're six years old...That prepares you for the basic school...er...Then you go to basic school lower grade which last...about five years...then when you are thirteen years old, you go to basic school's higher level. When the higher level ends, you are usually fifteen years old. And then you can choose...er...what you can do next. Basic school is an...is an obligation...to go to basic school, but now when you have finished it you have a chance to decide between college...er...or a...or any kind of different school. They last about three years, then you can decide if you want to go to college or some other place...mm...it's quite free to choose where do you want to go...and...er...this college or education for ours is based on a curse...curses...we have about ten...ten curse...curses which we have to do and...it is an obligation to do it if you are...have chosen to study in a coach...or ...we get a lot of homework here but if you have chosen this school you have to do it...you have to do the homeworks. After every curse...course...we have a...after every course...after every course we get a number...and after we have finished...kind...this college...and we have had all our numbers, we have done all our courses...then we have a final...final exam...and then we are ready to go another place.

I think that a Finnish system is...er...more complicated than the American system. We have to study many years if we want to get a good job and a good education...we have many alternatives, we have to choose between them...but...I am not so familiar with the American system...school system...and I really can't say much about it...but I have to say that Finnish system is quite good although it's a very long...it's very long time be....you can get a professor if you want to study...yourself....a good education...a good education one it depends...If you prefer to do something with a minor paycheck then you can quite...er...quite quick...er...you can quite quick get a job for yourself. (363 words)

Subtest 5

1. Well, don't be sad...let's go and have a pizza or hamburger...and then we can get married...Let's go to Las Vegas...play same games...then we go to see the priest and... let's get married, have children.

2. Well, this pullover was a bad choice for me. I can tell you where you can put this pullover...Give me a new one...or I call the police...that's a promise.

3. Oh, please mister, I beg you, can you give me four hundred dollars because I want to buy a camera. I don't have any souvenirs to bring home after a year in here...and I want to take a few photographs of you...and of the countryside here. I don't have any money right now but could you lend me that four hundred dollars...I will pay you back.

4. Well, the film you succested, it's very good film and I like Vietnam films...believe me, but I have seen the film. I saw it last week. I'm sorry, but I don't want to see it again...but...er...let's go and see another film at another time. OK?

5. Excuse me, lady, but I am in a very, very big hurry...er...I'm...My wife is giving a birth in about half an hour and I really got to get in the hospital. Do you understand me? Thank you.

6. Well...hi...Is the party over yet? I missed the bus and I couldn't get here in time but I'm sure that you spent...that you left me some left-overs and some bottle of ...bottles of beer. Is there...OK, thanks.

7. Well, where do you come from? What's the life here in Cincinnati? What do you think about Finland? Are we Finnish a quiet or something else? Tell me your opinions... about Finland...

8. Well, I think that there should be a fewer compulsory subjects at school, because if we have a right to choose what to do...what we want to study, it's best for all of us because if we have a right motivation, then we can get a good results.

9. Well, I think that wine should be sold at supermarkets. In Finland it's not allowed to sell wine supermarket but here in America things are much more free. You can buy a bottle of red wine anytime and it's...I think that it's better that way... because you don't have to go to local liquor store to get it, you can get it in your neighborhood...from your neighborhood.

10. Yes, well. perhaps Finland should take more refugees...because we have so...so few of them now and , for example here in America there are refugees from all areas of this planet of ours. And we in Finland...we of course have enough room for it...so I want them to come.

11. Thanks, Mike, it was a wonderful evening...wonderful...just wonderful, I remember this evening all my life. I really enjoyed our conversations and the film we looked at...it was nice...best party I have ever, ever been. (489 words, total 1217 words)

APPENDIX 4

Tables 18-20

TABLE 18 The correlations of the pronunciation ratings of raters A, B, and C in the LLOPT Subtests 1 and 3

	Pron, 1, Rater A	Pron, 1, Rater B	Pron, 1, Rater C	Pron, 3, Rater A	Pron, 3, Rater B	Pron, 3, Rater C
Pron, 1, Rater A	1.00					
Pron, 1, Rater B	.77	1.00				
Pron, 1, Rater C	.74	.75	1.00			
Pron, 3, Rater A	.73	.73	.71	1.00		
Pron, 3, Rater B	.55	.55	.51	.54	1.00	
Pron, 3, Rater C	.62	.61	.69	.70	.48	1.00

TABLE 19 The correlations of the fluency ratings of raters A, B, and C in the LLOPT Subtests 1, 3, and 4

	Flu, 1, Rater A	Flu, 1, Rater B	Flu, 1, Rater C	Qual, 3, Rater A	Qual, 3, Rater B	Qual, 3, Rater C
Flu, 1, Rater A	1.00					
Flu, 1, Rater B	.67	1.00				
Flu, 1, Rater C	.62	.64	1.00			
Qual, 3, Rater A	.66	.66	.53	1.00		
Qual, 3, Rater B	.61	.66	.53	.64	1.00	
Qual, 3, Rater C	.68	.63	.53	.79	.62	1.00
Flu, 4, Rater A	.62	.60	.45	.51	.51	.60
Flu, 4, Rater B	.65	.64	.54	.59	.50	.66
Flu, 4, Rater C	.58	.57	.40	.50	.58	.56

	Flu, 4, Rater A	Flu, 4, Rater B	Flu 4, Rater C
Flu, 4, Rater A	1.00		
Flu, 4, Rater B	.79	1.00	
Flu, 4, Rater C	.80	.75	1.00

TABLE 20 Length of stay in an English-speaking country

Identity	School	Gender	LLOPT score and placement (max. 107)	ACTFL	Final report	Matr. exam	Months in an English-speaking country	English in spare-time
55	J	F	103.7/1	A	10	l	11	+
36	J	F	102.3/2	A	10	l	2.5	+
10	H	F	102 /3	A	10	l	3.25	+
57	J	F	101.7/4	A	10	l	12	+
38	J	F	100 /5	A	10	l	0	
43	J	F	99.7/6	A	9	l	1	
2	H	F	98 /7-8	IH	10	l	0	+
51	J	F	98 /7-8	A	10	l	0	+
29	J	F	97.7/9	A	10	l	2	+
17	H	F	96.7/10-11	A	10	l	.75	+
49	J	F	96.7/10-11	IH	9	l	1.5	
19	H	F	96 /12-13	A	10	l	1	
34	J	F	96 /12-13	A	10	l	.75	
33	J	F	95.3/14	A	10	m	.75	+
14	H	F	93.7/15	A	9	m	12	
31	J	F	92.7/16	IH	9	c	.75	
53	J	M	91.7/17	A	10	l	0	
16	H	M	91 /18	A	9	l	0	
12	H	M	90.7/19-20	IH	8	m	3	+
18	H	F	90.7/19-20	A	9	l	0	+
7	H	M	88.3/21	IH	9	l	.75	
40	J	F	88/22-23	A	9	m	1.25	
28	J	M	88/22-23	A	10	l	0	+
58	J	F	87.7/24	IH	9	m	.75	+
25	H	F	85/25	IH	9	l	0	+
41	J	M	84.7/26-27	A	10	l	0	
60	J	M	84.7/26-27	A	10	l	0	
39	J	F	83.3/28-29	A	9	m	0	
54	J	F	83.3/28-29	IM	8	c	1	
23	H	M	82/30	IH	9	l	0	+
42	J	M	81.7/31	IM	9	m	1	
45	J	F	81/32	IH	9	m	0	

32	J	F	80.7/33	IM	9	c	0	
26	J	F	79/34	A	9	l	0	+
22	H	F	78/35	IH	7	c	0	
15	H	M	77.3/36	IL	7	c	0	
27	J	F	76.3/37	IH	9	l	0	+
50	J	F	75.7/38	IL	8	m	0	
44	J	F	74.7/39	IH	8	m	0	
56	J	M	72/40	IM	9	m	0	
37	J	M	71.7/41	A	10	l	0	
59	J	F	71.3/42	IM	6	b	0	+
21	H	M	71/43	IH	8	l	0	
4	H	M	69.3/44	IM	8	c	0	
13	H	F	68.3/45	IM	8	b	1	+
35	J	F	67.7/46	IM	8	c	0	
30	J	M	66/47	IH	9	c	0	
11	H	M	61/48	IL	7	b	0	
3	H	F	58.3/49	IL	7	a	.5	
1	H	M	5.3/50	IL	7	b	1	
5	H	F	56/51	IL	5	a	0	
24	H	M	54/52	IM	7	c	1.5	+
48	J	F	51.3/53	IL	6	b	0	+
52	J	M	51/54	NH	6	c	0	
9	H	M	50.7/55	IM	8	m	0	
47	J	M	46.3/56	IL	8	b	0	
6	H	F	43.7/57	NH	6	a	0	
46	J	F	40.3/58	NH	7	b	0	
20	H	F	34.3/59	IL	6	a	0	
8	H	F	32.3/60	NH	7	a	0	

YHTEENVETO

Punnitaan puhetta. Kielistudiokoe lukion vieraan kielen suullisen taidon päättökokeena.

Tausta ja tavoitteet

Viime vuosikymmenien tutkimukset kielitaidosta ovat osoittaneet, että samalla kun kielitaidon tarve yleensä lisääntyy, korostuu erityisesti tarve käyttää kieltä suullisesti. Euroopan neuvosto on ehdottanut, että vieraiden kielten opetuksessa painotettaisiin erityisesti suullista taitoa ja että kaikkiin tärkeisiin kielikokeisiin sisällytettäisiin suullinen koe. Monissa Euroopan maissa sekä ensimmäisen että toisen asteen koulutuksen päättökokeisiin sisältyykin vieraan kielen suullinen koe.

Myös Suomessa suoritetut tutkimukset ovat osoittaneet suullisen taidon tarpeen jatkuvasti kasvavan. Suomen liittyminen Euroopan unioniin on lisännyt tätä tarvetta entisestään. Jo vuonna 1988 kouluhallitus asetti työryhmän, jonka tehtävänä oli tutkia mahdollisuuksia suullisen kokeen järjestämiseen vieraiden kielten ylioppilastutkinnon yhteyteen ja kehittää siihen tarvittavia toimenpiteitä. Tehtävänsä toteuttamiseksi työryhmä järjesti vuosina 1990-1994 suullisen kielitaidon kehittämiseen ja arviointiin tähtäävän kokeilun. Kokeilusta saatiin myönteiset tulokset, ja opetushallitus (entinen kouluhallitus) onkin jatkossa pyrkinyt tehostamaan suullisen taidon opettamista. Kirjelmissään kouluille ja vuonna 1994 julkaistuissa lukion opetussuunnitelman valtakunnallisissa perusteissa opetushallitus on painottanut suullisen taidon opettamisen ja testaamisen kehittämistä. Myös opetusministeriö ja ylioppilastutkinnon kehittämistyöryhmä (1993) ovat esittäneet samansuuntaisia ajatuksia. Vuodesta 1995 lähtien lukion päättävillä oppilailla on ollut mahdollisuus suorittaa vapaaehtoinen suullinen koe useimmissa lukiossa opetettavissa kielissä.

Asettaessaan suullisen kielitaidon opettamista ja testaamista kehittelevän työryhmän kouluhallitus esitti, että käynnistettävään kokeiluun tulisi liittää tutkimusta. Yhtenä tutkimuksen keskeisenä tehtävänä oli laatia ja kokeilla suullisia koetyyppejä, joita jatkossa voitaisiin käyttää lukion oppilaiden arviointiin. Tehtävä kiinnosti minua, koska olin pitkään toiminut kieltenopettajien kouluttajana ja siinä työssä huomannut, että suullista taitoa ryhdyttäisiin harjoittelemaan laajalti vasta kun päättökoe uudistuisi.

Jotta olisi mahdollista testata suullista kielitaitoa, olisi ensin selvitettävä, mitä kielitaito ja erityisesti suullinen kielitaito on. Tätä tarkoitusta varten tutustuin kielitaitoa koskeviin tutkimuksiin ja totesin, että ei ole olemassa mitään yleisesti hyväksyttyä käsitystä tämän laajan ja monimuotoisen ilmiön olemuksesta. Tutkimuksen viitekehykseksi valitsin laajimmin tunnetun ja osittain empiirisessä tutkimuksessa todennetun Bachmanin kielitaidon mallin, joka pääpiirteiltään vastaa yleisesti käytettyä termiä kommunikatiivinen kompetenssi.

Perehtyessäni suullisen kielitaidon kuvauksiin lähdin liikkeelle olettamuksesta, että kun oppilaat päättävät lukion, heidän vieraiden kielten taitoaan testataan myös muiden kuin suullisen taidon osalta. Suullista koetta laadittaessa olisi tällöin tärkeä tietää, mitkä ovat ne piirteet, jotka ovat luonteenomaisia vain suulliselle taidolle ja joita siis tällaisessa kokeessa nimenomaan tulisi testata. Ratkaiseva ero suullista ja kirjallista tekstiä tuotettaessa on niiden laatimiseen käytettävä aika: kirjoitettua tekstiä tuotettaessa on enemmän mahdollisuutta harkintaan ja tarkistuksiin, mutta suullinen teksti on tuotettava aikapaineen alaisena. Puhuttu tuotos eroaakin kirjoitetusta oikeastaan kaikkien kielenkäytön alueiden - sanaston, syntaksin, morfologian, diskurssirakenteen ja pragmatiikan - puolesta. Puhutun diskurssin kuvauksessa kiinnitettiin tässä tutkimuksessa eniten huomiota tuottamisolosuhteiden vaikutukseen, puheen eri lajeihin, erityisesti keskusteluun, vakiintuneiden sanontojen merkitykseen ja kohteliaisuuden vaatimuksiin. Vain puheelle ominaisina piirteinä käsiteltiin ääntämistä, prosodisia piirteitä ja nonverbaalista viestintää. Sekä puhetta että kirjoitusta kuvattaessa voidaan puhua sujuvuudesta, mutta puhutun kielen sujuvuudella on omat kriteerinsä ja erittäin suuri merkitys puheen laatua arvioitaessa.

Kun lähdin kehittämään uutta suomalaisen lukion päättökokeeksi sopivaa puhumiskoetta, oli tietysti luonnollista aluksi kartoittaa, mitä koetyyppejä on olemassa ja millaisia kokemuksia niiden käytöstä on saatu. Varsin pitkälle kehittynyttä kielitaitoa, kuten Suomessa A1-kielenä opiskellun englannin taito, on kansainvälisesti yleisimmin totuttu arvioimaan haastattelulla. Menetelmää on kuitenkin arvosteltu sen yksipuolisuudesta ja varsinkin kalleudesta. Haastattelutestin suorittamiseksi ei esimerkiksi jouduta kouluttamaan vain arvioitsijoita vaan myös suuri joukko haastattelijoita. Myös haastattelutestin validiuteen ja reliaabeliuteen on suunnattu kritiikkiä. Haastattelun rinnalle on viime vuosina kehitetty kielistudiossa toteutettava SOPI-testi (simulated oral proficiency interview). Kielistudiotestin puutteena on pidetty tilanteen keinotekoisuutta, mutta vastapainona on mahdollisuus käyttää useampia koemuotoja ja siten saada monipuolisempaa ja luotettavampaa tietoa kielitaidosta. Vähemmän edistynyttä kielitaitoa on arvioitu myös pari- ja ryhmäkeskusteluilla, joissa validius ja reliaabelius ovat kuitenkin varsin alttiina tilannetekijöille.

Alkaessani laatia suullista koetta oli laajimpana tavoitteena selvittää, olisiko mahdollista laatia koe, jolla voitaisiin testata koko lukion päättävä ikäluokka - mikäli mahdollista samanaikaisesti. Parhaimman mahdollisuuden tuntui tarjoavan kielistudiokoe, joten oli luonnollista ryhtyä kehittämään testimateriaalia tältä pohjalta. Tärkein tehtävä oli vastata kysymykseen, voitaisiinko lukion päättävien oppilaiden vieraan kielen taitoa arvioida tällaisella kokeella pätevästi, luotettavasti ja tehokkaasti. Vertailukohdan saamiseksi haluttiin myös katsoa, miten hyvin oppilaiden taitoa voitiin arvioida haastattelutestillä. Lisäksi tutkittiin oppilaiden asennoitumista vieraan kielen puhumiseen ja testaamiseen sekä kysymystä, parantaako kohdemaassa

oleskelu tuntuvasti oppilaan vieraan kielen suullista taitoa. Jos näin olisi, suullinen koe saattaisi lisätä oppilaiden eriarvoisuutta.

Aineisto ja analyysimenetelmät

Suullinen koe olisi voitu laatia minkä tahansa opiskellun kielen kokeeksi ja näin testata sen soveltuvuutta valtakunnalliseksi lukion päättökokeeksi, mutta valitsin kohdekieleksi englannin kolmesta syystä. Oli ensinnäkin aiheellista valita kieli, jossa oppilaiden taidot olisivat pisimmälle kehittyneet, koska tällaisen kielen taitojen testaaminen edellytti monipuolisinta koetta. Näin saatujen kokemusten perusteella on mahdollista myöhemmin laatia yksinkertaistettuja versioita vähemmän opiskeltuihin kieliin. Englanti tuntui tärkeältä myös siksi, että sitä opiskelee niin suuri määrä opiskelijoita. Lisäksi englannin taidon testaamisesta on olemassa eniten kansainvälistä kirjallisuutta.

Koehenkilöt saatiin kahdesta koulusta, jotka osallistuivat opetushallituksen vuosien 1990-94 kokeiluun. Toinen oli itäsuomalainen kaupungissa toimiva normaalikoulu, toinen lounaissuomalainen maaseutukoulu. Ensin mainitussa koe suoritettiin kevätlukukaudella 1993 ja siihen osallistui 35 oppilasta, jälkimmäisessä koe oli syyslukukaudella 1993 ja osallistujia oli 25. Sekä kielistudiokoe että haastattelutesti oli esitestattu eräässä lounaissuomalaisessa kaupunkikoulussa.

Kielistudiokoe käsitti kuusi osaa: (0) helppo viritysosio, jota ei arvosteltu, (1) ääneen lukeminen, (2) arkikeskustelun tulkitseminen englanniksi, (3) sanomalehtijutun informaation välittäminen, (4) lyhyen esityksen pitäminen, (5) arkielämän tilanteissa reagoiminen ja mielipiteiden esittäminen. Kielistudiokokeen epäaitoutta pyrittiin vähentämään suunnittelemalla osakokeet mahdollisimman luonteviksi ja arkielämän viestintätilanteita vastaaviksi. Sisällöt pyrittiin laatimaan miellyttäviä assosiaatioita tuottaviksi. Kokeille laadittiin yhteinen kommunikatiivinen viitekehys. Osassa 1 oppilas lukee luokkatovereilleen kirjeen, jossa amerikkalainen nuoriso-orkesteri ilmoittaa halustaan tulla Suomeen. Toisessa osassa yksi orkesterin jäsen on majoittunut oppilaan kotiin, ja oppilas tulkitsee hänelle äitinsä suomenkielistä puhetta. Osassa 3 oppilas kertoo vieraalleen suomalaisesta sanomalehdestä lukemansa jutun sisällön. Neljännessä osassa amerikkalaisvieraat ovat oppilaan koulussa, ja oppilaan tehtävänä on selostaa heille Suomen koululaitosta, erityisesti lukiota. Myös kulttuurin tuntemusta haluttiin testata: oppilaat joutuivat vertailemaan suomalaista ja amerikkalaista koulua. Viimeisessä osassa koehenkilö on vastavierailulla USA:ssa ja joutuu päivän kuluessa erilaisiin tilanteisiin, joissa hänen on reagoitava asiallisesti ja kohteliaasti. Tässä osassa hän joutuu myös esittämään ja perustelemaan mielipiteitä.

Oppilas kuuli valmiiksi tauotetun äänimateriaalin nauhalta, jonka kesto taukoineen oli 40 minuuttia. Lisäksi oppilaalle jaettiin kirjallista materiaalia. Oppilaan tuotos nauhoitettiin, ja nauhan pituudeksi tuli 20-22 minuuttia, mitä kokeen kehittelyssä on mahdollista lyhentää. Kokeessa arvioitiin sekä interaktionaalista (vuorovaikutteista) että transaktionaalista (esittävää) puhetta. Sekä kielistudiokokeen että haastattelun arvioi testaaja ja kaksi taitavaa kieltenopettajaa. Kielistudiokokeessa kukin osakoe arvioitiin erikseen, ja kullekin kokeelle määriteltiin painokerroin. Koko kokeen tai jonkin osakokeen perusteella arvioitiin lisäksi seuraavat ominaisuudet: ääntäminen, sujuvuus, koheesio, informaation välittäminen ja sosiolingvistinen kompetenssi.

Kielistudiokokeen validiutta testattiin kansainvälisesti tunnetuimmalla haastattelutestillä nimeltään ACTFL OPI (the American Council on the Teaching of Foreign Languages Oral Proficiency Interview). Koe on alunperin laadittu arvioimaan amerikkalaisten diplomaattien ja virkamiesten kielitaitoa, mutta siitä on myöhemmin kehitetty laajasti käyttökelpoinen yleinen testi. Haastattelussa on neljä vaihetta: lämmittely, kielitaidon tason varmistukset, tunnustelut ylöspäin ja lopetus. Näiden osien lisäksi on tapana käyttää myös roolileikkiä. Tutkimusta varten kehitin uudet, erityisesti Suomen olosuhteisiin sopivat roolileikkitehtävät. Niiden avulla voitiin haastattelun lopulla tarkistaa vielä epäselväksi jääneitä seikkoja. Haastattelun arvioinnissa kielitaidon päätasoja erotettiin neljä: alkeistaso, keskitaso, edistyneiden taso ja taitajien taso, ja ne jakaantuivat ylintä lukuun ottamatta alatasoihin. Tasoa määritettäessä kiinnitettiin huomio seuraaviin piirteisiin: sujuvuus, kielioppi, pragmaattinen kompetenssi, ääntäminen, sosiolingvistinen kompetenssi ja sanasto. Haastattelut kestivät oppilaan tasosta riippuen 15-25 minuuttia. Ne nauhoitettiin äänikasetille myöhempää arvioimista varten.

Kokeilukouluissa suoritettiin ensin haastattelu, sitten kielistudiokoe ja viimeisenä asennekartoitus. Likert-tyyppisessä mittauksessa oppilaat selvittivät suhtautumistaan suulliseen harjoitteluun ja testaukseen. Lisäksi he kertoivat käsityksiään suullisen taidon opettamisesta ja vertailivat haastattelua ja kielistudiotestiä toisiinsa.

Tulokset

Ensimmäinen osakoe, ääneenlukutehtävä, oli vaikeampi kuin oli osattu odottaa: sekä ääntämisessä että sujuvuudessa oli monien kohdalla toivomisen varaa. Helpoimmaksi kokeeksi osoittautui toinen tehtävä, arkikeskustelun tulkitseminen englanniksi, ja myös kolmas tehtävä, sanomalehdessä olleen helpohkon tarinan välittäminen, onnistui hyvin. Neljäs tehtävä, esityksen pitäminen Suomen koululaitoksesta, tuotti eniten vaikeuksia. Siinä samoin kuin kolmannessa tehtävässä mahdollinen heikko tulos johtui usein sanaston puutteellisesta hallinnasta. Viidennessä osassa, jossa tehtävänä oli vieraassa maassa kohdatuista tilanteista selviäminen, sanastolliset vaikeudet pystyi usein kiertämään, mutta äänensävy ja tapa esittää asia vaikuttivat ratkaisevasti lopputulokseen. Tästä tehtävästä koehenkilöt selvisivät kohtalaisesti, mutta tuotosten arvioiminen oli vaikeaa. Hyväksi tehtävätyypiksi osoittautui mielipiteiden esittäminen.

Haastattelukokeessa saatuja tuloksia oli mahdollista varovasti verrata muualla saatuihin. Kokonaisuutena suomalaiset oppilaat selvisivät kokeesta hyvin. Suomalaisessa versiossa ylimpään kategoriaan sijoittuvia tuloksia oli paljon, mutta toisaalta oli myös heikkoja suorituksia. Kokeesta saatiin myös tärkeä tieto, joka ei sisältynyt koeasetelmaan: haastattelijan osuus, itse haastattelun suorittaminen A1-kielessä, osoittautui yllättävän vaativaksi. Jos kauimmin opiskellussa kielessä halutaan suorittaa päteviä ja luotettavia haastatteluja, haastattelijoiden kouluttaminen vaatii huomattavia resursseja.

Asennetutkimus osoitti oppilaiden suhtautuvan sekä puhumisen harjoitteluun että sen testaukseen varsin myönteisesti. Vaikka näillä oppilailla oli ollut tavallista enemmän puheharjoittelua, he toivoivat sitä vielä lisää. Heidän mielestään myös äidinkielen opetuksessa on lisättävä suullisen tuottamisen harjoittelua. Vieraan kielen

puhumista tulee testata ylioppilastutkinnossa joko kielistudiokokeessa tai mieluummin haastattelussa.

Tutkimus ei antanut selkeää kuvaa ulkomailla oleskelun vaikutuksesta suulliseen taitoon. Monet suullisessa testissä hyvin menestyneet oppilaat olivat tosin oleskelleet ulkomailla, mutta samat oppilaat menestyivät hyvin myös kirjallisissa ylioppilaskokeissa. Toisaalta oli myös oppilaita, jotka menestyivät hyvin suullisessa kokeessa, vaikka eivät olleet oleskelleet englantia puhuvassa maassa.

Tutkimuksen pääkysymykseen, voidaanko oppilaiden suullista taitoa arvioida pätevästi ja luotettavasti kielistudiokokeella, tutkimus antaa myönteisen vastauksen. Laadittu kielistudiokoe mittaa kielitaitoa monipuolisesti. Kokeen tulos, 73 % maksimisuorituksesta, oli hyvin samantapainen kuin kansainvälisesti arvostetussa haastattelukokeessa saatu tulos, 72 % maksimista. Useimmat oppilaat, 60 %, saivat asteikolla 1-5 saman tuloksen molemmissa kokeissa, ja vain yhden oppilaan tulos poikkesi kahden arvosanan verran. Se, että osa oppilaista sai eri testeissä myös erilaisen tuloksen, voi johtua tilannetekijöistä (satunnainen varianssi), mutta osoittaa myös, että testit mittaavat osittain eri alueita kielitaidosta. Kielistudiokokeen luotettavuutta haastatteluun verrattuna parantaa ratkaisevasti se, että tehtävienantovaihe - nauhalta kuultu teksti - on sama kaikille koehenkilöille. Toisaalta kielistudiokokeen luotettavuutta voi horjuttaa sen koehenkilöissä aiheuttama jännitys ja myös sen alttius teknisille häiriöille.

Johtopäätöksiä

Kokeilukoulujen oppilaiden saamat valtaosin hyvät tulokset kansainvälisessä haastattelutestissä osoittavat, että puheharjoittelun lisääminen lukion vieraan kielen opetuksessa ei mene hukkaan. Kielistudiokokeen eri alueiden tarkastelu kertoo, että lukion päättävillä oppilailla on hyvät englannin kielen valmiudet selviytyä jokapäiväisestä keskustelusta ja annetun tarinan tulkitsemisesta. Lyhyen itselaaditun suullisen esityksen pitäminen tuotti sen sijaan vaikeuksia ja kaipaa lisäharjoittelua. Vaikka aihe - suomalainen koulu - on tutuista tutuin, monen kokelaan puhe kangerteli puutteellisen sanavaraston vuoksi. Myös ääntämisen ja sujuvuuden harjoitteluun tulisi kiinnittää entistä suurempaa huomiota. Kahden koulun tuloksia ei ole syytä yleistää, mutta omat kokemukseni opettajankouluttajana tukevat edellä esitettyjä käsityksiä. On mahdollista, että pitkään muodissa ollut kommunikatiivinen menetelmä on yksipuolistanut suullista harjoittelua. Ymmärretyksi tuleminen ja luonteva vuorovaikutustilanne ovat toki välttämättömiä, mutta ne eivät riitä kymmenen vuoden kielenopiskelun tuloksiksi.

Aikaisempaan tutkimukseen perehtyminen osoitti, että suullista kielitaitoa olisi testattava erillisellä kokeella. Puhuttu ja kirjoitettu kieli eroavat toisistaan merkittävästi, ja sujuvan puheen tuottamiseen vaadittavat valmiudet, esimerkiksi automaattistuminen ja keskustelustrategiat, ovat pääosin sellaisia, joita kirjallisessa kokeessa ei voi testata. Joidenkin oppilaiden menestys suullisissa kokeissa poikkesi huomattavasti heidän menestyksestään kirjallisissa kokeissa.

Ehkä tärkein syy, miksi puhetta pitäisi testata, on päättökokeen takaistusvaikutus opetukseen. Jos puhumista testataan päättökokeessa tänään, sitä opetetaan koulussa huomenna. Asennetutkimus osoitti, että myös oppilaat, jotka ovat jo saaneet

tavallista enemmän suullisen taidon opetusta, haluavat sitä vielä lisää opetusohjelmaan. Puhetaidon testausta ei tarvitse karttaa siitäkään syystä, että tutkimus olisi osoittanut ulkomailla oleskelun lisäävän sosiaalista eriarvoisuutta. Testin säilyminen vapaaehtoisena voi sen sijaan lisätä alueellista epätasa-arvoa. Opettajista innostuneimmat valmentavat oppilaitaan myös puhetestiin, mutta välinpitämättömät ja väsyneet eivät jaksa vaivautua. Etelän taajamissa asuvat oppilaat ovat tällöinkin etuoikeutetussa asemassa, sillä edistyneitä opettajia on paljon juuri etelässä ja/tai suosituissa kaupungeissa.

Kielitaidon olemukseen paneutuminen osoitti, että kielitaito on niin monimuotoinen ilmiö, ettei sitä voida pätevästi testata vain yhdellä testityypillä. Kielistudiokokeeseen on mahdollista sisällyttää monia erilaisia testejä ja siten parantaa sen testausominaisuuksia. Päättökokeessa mahdollisesti käytettävän testin valinta riippuu myös testattavan kielen hallinnasta. Jos kieltä on opiskeltu pitkään, kuten A1-kieltä, taitoa luotettavasti mittaava haastattelutesti vaatisi erittäin ammattitaitoisia haastattelijoita ja olisi näin ollen toteutettavissa vain harvojen oppilaiden kohdalla. Sen sijaan kieliä, joita on opiskeltu lyhyemmän aikaa, voisi ilmeisesti testata joko kielistudiotestillä tai haastattelulla.

Jos testi halutaan toteuttaa valtakunnallisesti, on otettava huomioon paitsi sen pätevyys ja luotettavuus myös kustannukset. Suullisen harjoittelun mahdollisesti lisääntyessä kielistudioiden tarvetta joudutaan testaamisesta riippumatta harkitsemaan uudelleen. Oikein käytettynä studio on tehokas väline esimerkiksi sujuvuuden kehittämiseen. Testitilanteessa kielistudiosta aiheutuneita kustannuksia kompensoi ajassa saavutettava säästö. Suoritettu kokeilu osoitti sitä paitsi, että luotettavaan tulokseen voisi päästä tässä käytettyä lyhyemmällä ja yksinkertaisemmalla testillä. Silti uudesta kokeesta saattaisi aiheutua sellaista kieltenopettajien työmäärän lisääntymistä, jota Suomen kieltenopettajien liitto pelkää. Sen välttämiseksi olisi harkittava, millä tavoin nykyisessä ylioppilastutkinnossa käytössä olevia testejä voisi integroida ja yksinkertaistaa. Olisi ehkä mahdollista testata samalla kokeella sekä kuullun ymmärtämistä että puhumista. Kirjoittamista testataan nykyisin samassa tutkinnossa sekä avovastauksilla että kirjoitelmilla, mitä voisi yksinkertaistaa. Jos testaamisesta aiheutuva työmäärän lisääntyminen näin voitaisiin välttää, kieltenopettajien liitto ei varmaankaan halua vastustaa itse opetuksen monipuolistamista.

Onnistuessaan tutkimus tuo tullessaan uusia kysymyksiä ja/tai tarpeen syventää saatua tietoa. Tämä tutkimus rajoittui kahteen kouluun eikä siis antanut mitään yleistettävää tietoa. Avoimeksi jäi esimerkiksi kysymys, olisiko viime-aikaisissa peruskoulututkimuksissa ilmennyt kielitaidon epätasainen jakaantuminen maan eri osiin ollut nähtävissä myös lukiossa. Miten opettajat suhtautuisivat eri tapoihin testata suullista taitoa, jos heillä olisi niistä omakohtaista kokemusta? Mitä muita testityyppejä voitaisiin käyttää jokavuotisessa valtakunnallisessa kokeessa? Minkälaiset koemuodot sopisivat vähemmän osattuihin kieliin? Kielitaidon testaaminen on laaja ja tärkeä alue, johon yksittäiset tutkimukset voivat antaa vain kapeaa valoa. Aiheen taloudellisen ja kulttuurisen merkittävyyden huomioon ottaen tuntuisikin luonnolliselta, että sen jatkuva kehittäminen annettaisiin tehtäväksi jollekin pysyvälle elimelle, jolle osoitettaisiin riittävät resurssit.