Thematic Network Project in the area of Languages

Sub-project 10: Testing

Assessing the Language Proficiency of Modern Language (Under)Graduates: report on a survey and feasibility study

Sub-project 10: Testing

Assessing the Language Proficiency of Modern Language (Under)Graduates: report on a survey and feasibility study

Appendix to the Final Report for Year Three

List of Contents

Prepared by the Scientific Committee of the Sub-project on Testing
September 1999

Assessing the Language Proficiency of Modern Language (Under)Graduates: report on a survey and feasibility study

Preface

Angela Hasselgren
Editor, Coordinator for TNP subproject no. 10

The TNP in the Area of Languages subproject no. 10: Testing was formed early in 1997 and the Scientific Committee has consisted of:

- Helmut Bonheim, Universität zu Köln, DE
- Raymond Capré, Université de Lausanne, CH
- Gerda Cook-Bodegom, Háskóli Íslands, IS
- Angela Hasselgren, Universitetet i Bergen, NO
- Arthur Hughes, University of Reading, UK
- Heidrun Klemm, Universität Potsdam, DE
- Paz de la Serna, Universidad Autónoma de Madrid, ES
- Sauli Takala, Jyväskylän yliopisto, FI

Fellyanka Kaftandjieva (Jyväskylän yliopisto, FI) has worked with the Committee since the beginning of 1999.

The Committee members have a wide range of experience within the area of language assessment in Higher Education (HE) and have shared, from the outset, a common awareness that HE is, in many ways, the "poor relation" when it comes to assessment of language proficiency. Much investment has gone into research and development work on the testing of the foreign language ability of working adults, for example by the Council of Europe, implemented in projects such as DIALANG. The school system has also benefited from this European cooperation, as well as being generally well-endowed with national examination boards with considerable expertise.

In universities and colleges however, the assessment is frequently in the hands of individual teachers with no training or access to innovative assessment methods, and with no criteria for evaluating students' performance. Not only can this have adverse consequences for the students being assessed, but it means that no basis exists for comparison between the level of language proficiency of students from different institutions/countries. In this era of student exchange programmes and a general tendency for graduates to seek work or to continue their studies in another country, this lack of any yardstick is clearly a handicap. In the case of language majoring students, this situation has further repercussions, viz. for the quality of language teaching in Europe, and hence for the language ability of its future generations of citizens.

Against this background, the Committee decided to make it its first aim to find out what is needed in order to enhance the assessment of the language proficiency of language majoring (under) graduates and language teacher trainees. Its next aim was to investigate how these needs may be catered for through a concrete set of measures, carrying out a pilot study of these measures on a limited group of students/institutions and languages, and focusing on a particular aspect of proficiency.

This document describes the processes undergone by the Committee in carrying through its aims, and the rationale behind these. It presents the survey material used to track the

language assessment needs of HE institutions and its major findings. It also presents modified versions of the testing/self assessment instruments developed by the Committee to investigate the level of reading ability of language majoring students, together with an analysis of the trialling data. It draws conclusions on the feasibility of the future development of a fully-fledged instrument for the assessment of (under)graduate language proficiency, which would implement an adaptation of the Council of Europe Framework (CEFr) for Language Learning, Teaching and Assessment (Council of Europe 1996). Some future directions for the continuation of the work initiated by the group are indicated, and an invitation is extended, through the ELC website, to those involved in the HE assessment of language proficiency, to try out the 'products' and give feedback on the work of the Committee.

Chapter one: Overview of Activities: 1997 - 1999

Raymond Capré

This section of the report will cover the activities of subgoup Nr 10 from the time the group was established, at the beginning of 1997, up to the final Conference in Jyväskylä, in July 1999. The details of each meeting will not be gone into, but the achievements of the group will rather be concentrated on, stressing either what the group has produced or what the group has found out.

Before proceeding to a year-by-year account of the activity, an overview of the various meetings held by the Scientific Committee of subgroup Nr 10 and the main papers that have been produced are shown below.

Meetings of the Scientific Committee

22/02/1997	Berlin, Freie Universität Berlin	First meeting
31/05/1997	Reading, University of Reading	Meeting
05/07/1997	Lille, Université Charles-de-Gaulle Lille III	Evaluation Conference Meeting
25/10/1997	Potsdam, Universität Potsdam	Meeting
24/01/1998	Berlin, Freie Universität Berlin	Workshop
28-29/03/1998	Bergen, Universitetet i Bergen	Meetings
23-24/10/1998	Madrid, Universidad Autónoma de Madrid	Meetings
12-13/02/1999	Cologne, Universität zu Köln	Meetings
01-03/07/1999	Jyväskylä, Jyväskylän yliopisto	Second Conference Workshops, Meetings

Documents produced

- Policy paper, February 1997
- Lille Conference, Workshop Reports, Université Charles de Gaulle – Lille III, 1997, pp 52-54.
- Questionnaire Nr 1 on testing, Results, July 1997 (Appendix A)
- Questionnaire Nr 2 on testing, Results, February 1998 (Appendix B)
- Minutes of every meeting, with appendixes, 1997-1999
- Yearly reports, Year 1, Year 2, Year 3
- Feasibility study on exit tests in English and in French and results (see present report)

For a presentation or summary of these documents, see the ELC website:
http://www.fu-berlin.de/elc/

1.1 Year 1 – Defining objectives

The Scientific Committee had to define its objectives and plan its activities in relation to the DIALANG project launched in December 1996. Two important issues to be investigated were identified:

- Training in assessment
- Exit testing at University level

It was decided to do a survey of existing assessment practices and needs in several European countries. Two subgroups conducted the investigation. One group sought

information, on the basis of a questionnaire (Appendix A), on language procedures at the end of tertiary education, whereas the other group conducted a survey in Finland and in Iceland on how "non-language students" were evaluated as to their language abilities.

Answers to the first survey questionnaire showed several significant things (see Tables 1 to 3). Firstly, fewer than half of respondents had any kind of specified criteria for assessing written or spoken language ability. 78% of respondents said that they believed that members of their department needed guidance in developing assessment methods, and the same proportion of respondents said that it would be very useful to them to have at their disposal standardised tests which would allow them to compare their students' performance with that of other students studying the same language in Europe.

There was a generally high level of satisfaction with the way assessment was carried out (around half opting for 4 or over on a five-point scale) and two thirds believed their institutions were competent (using the same measure) in this assessment. However, these figures are undermined by the revelation that the this competence is based overwhelmingly on practical experience alone (80%), with a combination of practice and training in only 15% of cases. Moreover, the low status of assessment as a subject of study is reflected in the fact that only 39% of institutions concerned offer assessment as course components and as few as 7% carry out research on language assessment.

Table 1. Practices, level of satisfaction and perceived needs at informants' institutions

	yes	no	total of asked
have criteria for assessing written language ability	49%	44%	93%
have criteria for assessing spoken language ability	46%	46%	92%
satisfied with practice *	51%	47%	98%
consider competent*	66%	32%	98%
need for guidance	78%	22%	100%
would like standardised test	78%	10%	88%

* 'satisfied' is defined here by the electing of 4 or over on a five-point scale

The studies conducted in Iceland and in Finland came partly to the same conclusion. Although there was a relative degree of satisfaction about the way testing was carried out with language degree students in Iceland, it appeared that in Finland very few teachers were thoroughly satisfied with the guidelines for oral assessment. See tables below for more details.

Table 2. Competence in testing (as perceived by informants)

	practical experience alone	training in testing alone	both	total of asked
competence based on	80%	0%	15%	95%

Table 3. Status of assessment at informants' institutions

	coursework	research	both	neither	total of asked
assessment studied as	39%	7%	0%	51%	97%

After discussing these matters at the workshop in Lille, the group came to the conclusion that two main themes had been identified for further study:

- Development of exit tests for language students in tertiary education which would allow comparisons to be made across Europe.

- Developing a set of guidelines for the assessment of graduate/teacher trainee language proficiency.

The second year would be devoted to go deeper into those themes, whereas year three would be devoted to make proposals for a pilot study in at least one of these areas. At the same time, great attention will be paid to the developments of DIALANG in which our group should be able to find inspiring ways of tackling with the areas pointed out above and also to the Framework set up by the Council of Europe in which scales of proficiency could be used as a starting point for our project.

These points were made clear at our meeting in Potsdam. In addition to that it was decided to issue another questionnaire to be sent to university departments in order to find out more information about the ways language tests were being carried out (Appendix B).

1.2 Beginning of Year 2 - Deeper into the matter

At the beginning of year 2, the workshop held at the Freie Universität Berlin was meant to give members of the Scientific Committee information about what was being done elsewhere in our field and how it was done. Two experts attended the workshop, Digna Samson from the Dutch National Institute for Educational Measurement (CITO), and Steven Fligelstone from the Dept of Linguistics & Modern English Language, also Project Administrator in the DIALANG project.

The workshop included both points for discussion and points for information. Here are the main points brought to us for information and for feeding our reflection, the basic idea being to progress towards some proposal for exit tests in languages at university level.

- Digna Samson showed us how reading comprehension in a foreign language is tested at the end of Dutch secondary education. She discussed objectives, question formats, levels of difficulty, and the link between criteria and scales relating to examination texts. She illustrated all these aspects with examples from the Dutch examinations and she stressed the need for objectivity, particularly in large settings.

- Steve Fligelstone gave a comprehensive account of the possibilities and the potential problems for the use of computers and the internet in language assessment. He supplied us with lists of possible types of items, new possibilities offered by computers in testing as well as lists of basic questions dealing with implementation problems, security, administration, limits of computers, advantages and disadvantages, etc

- Council of Europe Framework (CEFr) Scales (Council of Europe 1996a). They had been studied by all members of the committee and were discussed then. It was generally felt that the scales of proficiency developed for the Council of Europe and the descriptors attached to each point of the scales were very well made, but some members of the group were not sure whether the scales would also work at higher University levels. It was agreed though that descriptors at higher levels of the scales could be expanded and made more explicit.

In the absence of Brian North, who had been a major contributor to the CEFr scales, an article of his "The Development of a Common Framework Scale of Descriptors of Language Proficiency Based on a Theory of Measurement" (North 1997) was discussed. Members were impressed by the thoroughness of the procedures and in the view of a possible expansion of scales by the group it was anticipated that similar procedures might be used.

- ACTFL Scales were also examined by the Committee. Two key issues were identified in the use of the scale which would be of significance in the future development of scales: the way in which descriptions of a level bundled together a number of features which might not all describe a particular learner's ability; and, related, the non-compensatory nature of level assignment (eg in speaking, the level assigned to a student would be the lowest level achieved between accent, grammar, vocabulary, and fluency).

- Questionnaire Nr 2. The results of this questionnaire were presented to us, commented and discussed. The questionnaire was meant to investigate amongst university language departments the degree of importance given to language testing at the time of a final examination, and the way it was done. Many departments admitted that it was done according to "a general impression of language use"; at the same time, it was apparent that, here and there, a wide variety of techniques were used, but also that there was probably room for improvement. The importance given to language proficiency when it comes to giving a grade seems to vary greatly, from 10% to 80%, with a peak around 30%. The great variety of techniques used, together with a wide difference in the importance given to the evaluation of language proficiency at the end of a "language" program gave confirmation to the Committee that we should work towards defining exit test specifications that would make it possible to compare levels of achievements between various languages and across borders. (Complete results of Questionnaire 2 are included in Appendix B)

- Components and format were discussed at great length. As far as format is concerned, members of the Committee tended to prefer multiple choice and fill-the-gap questions to open questions. Concerning components, there seemed general agreement that the traditional four skills should be included in the testing procedures to be proposed. Some voices stood in favour of testing also metalinguistic knowledge, cultural knowledge, grammar on specific points, vocabulary, etc, but after hearing about all the problems encountered by the teams dealing with DIALANG, it was agreed that we should limit ourselves to language testing, thus excluding linguistics, literature and culture.

1.3 End of year 2 - Defining a feasibility study

The end of year two was used to evaluate all the issues raised in Berlin, and to restrict ourselves to what would be reasonable to achieve in the three year period planned for the project. In Bergen, in March 1998, we agreed upon concentrating on scales for one skill. The decision was made to deal with **reading skills**.

We agreed to investigate amongst our colleagues and students to try to establish what they felt that students should be able to do in terms of reading abilities when successfully completing their studies in a language department. On the basis of this research we should be able to select a list of *can do*s for reading. At the same time we should select texts, and questions related to those texts, that could be connected to the *can do*s. It was also agreed to use some non text dependant questions. All this should contribute to a feasibility study: trying to define in terms of *can do*s what should be the reading ability of a language majoring student and giving examples of ways of testing these *can do*s. It

was agreed to test material both in English and in French. Material had to be collected and sent to each other prior to the beginning of year three.

1.4 Year 3 - Feasibility Study

Can-dos and sample tests were collected at the first meeting of year 3, in Madrid, and decisions were made as how to proceed with our feasibility study. The main aim would be to produce a pilot instrument, in English and in French, for measuring and describing the reading ability of students around the point of graduation from first degree ML courses, or teacher training language courses. The instrument should consist of:

- A scale or set of descriptors of graduate reading ability which would be an adapted version of the CEFr scale of general reading ability. In particular, descriptors should be considerably extended and made more specific to the skill to be assessed.

- Test and self-assessment items targeting specific aspects of abilities mentioned on the scale.

From the list of *can do*s collected by the group and from texts and questions brought to the Madrid meeting, five core areas were identified, concerning reading ability around C1 level in the CEFr.

It was then decided to test which texts would match these various areas, which items would target these, which texts students at this level would be able to cope with, how they would assess these tests (self-assessment), and how teachers would assess these students. All this had to be put in a suitable form, both in English and in French, so that students and teachers would understand how to answer. A more detailed account of the procedures followed in designing and carrying out the feasibility study is given in Chapter Two.

Producing these tests, passing them to groups of students, collecting the results, sending those results for analysis, making the analysis, giving the results of the analysis was done in three months only, between our Madrid meeting and our meeting in Cologne. This shortage of time explains the relatively low number of informants, 92 in English, 93 in French. However, that was a feasibility study, and as such it yielded some very interesting pieces of information.

The Cologne meeting was devoted to the findings from the analyses of the data from the trialling of the tests. Fellyanka Kaftandjieva, who had performed the analyses at the University of Jyväskylä, attended the meeting and gave an orientation of the outcomes of her findings. The discussion at the meeting involved the degree to which the testing instruments had elicited the information hoped for, and how this may have been improved. This discussion was of fundamental importance to the work of the group, since the current study was a feasibility study, carried out as a potential forerunner to a fully fledged project on testing the language proficiency of students, to be proposed at a later date. The actual findings from the analyses and a detailed discussion of their implications are presented in Chapter Three.

To be complete, let us just mention that year 3 ended with a series of workshops on Assessment in Higher Education during the ELC Second Conference in Jyväskylä and with our last meeting in which we agreed on proposals and projects for the future. The outcomes of both of these events are discussed in Chapter Four.

Chapter two: Towards an instrument for assessing the reading ability of exiting language graduates in HE

Gerda Cook-Bodegom, Angela Hasselgren, Heidrun Klemm

In this chapter, an outline will be presented of the development of an instrument for assessing the reading ability of exiting language graduates in HE, which formed the major part of the TNP activity from the latter part of year two, until the middle of year three. The presentation will begin, in Section 2.1, by considering the purpose the instrument was to serve, against the background of what had emerged from the work, including surveys, of the first two TNP years, with a discussion on the essential format of the instrument, and the rationale behind it. Section 2.2 will give an in-depth look at what makes up reading ability, and Section 2.3 will present an overview of the specification of the 'test' part of the instrument. Finally, in Section 2.4, an account will be given of the actual procedures followed in developing the instrument.

2.1 Development of an instrument for assessing the reading ability of exiting language graduates in HE

2.1.1 What purpose should the instrument serve?

Several factors were influential in deciding what kind of instrument we were to develop, in terms of what it would tell us about our students' ability. The results of the two surveys led us to believe that some kind of standardised testing was needed. However, if the test was to tell anything universally understandable about our learners' ability, then some concrete criteria were needed for the interpretation of scores, and as a basis for any other type of assessment which may be carried on within institutions. The lack of such criteria had already been revealed by the survey, in the case of testing the production of written and spoken language.

Since the CEFr (Council of Europe 1996a) has already been widely accepted as a cornerstone in European assessment systems, such as DIALANG (see Kaftandjieva 1999) and the European Language Portfolio (Council of Europe 1996b), it was felt that the criteria in the CEFr for assigning levels constituted a natural starting point for the criteria to be adopted in this instrument. What was felt to be needed therefore, was an instrument which provided a way of assigning HE (under)graduates to levels of performance corresponding to levels (or sub-levels) on the CEFr, but which gave information relevant to academic language ability. As the CEFr is more fully developed for the lower levels (see DIALANG's findings on the power of *can dos* to distinguish between learners at the two upper CEFr levels, in Kaftandjieva 1999), it was clear that the descriptors of ability would have to be supplemented in quantity as well as being adjusted in quality, for the learner group concerned.

2.1.2 Format and general characteristics of instrument

This section will deal with the format and general characteristics of the instrument which was developed. The discussion will consider such matters as which languages should be offered, what levels should be targeted, how long the procedure should take, what kind of what kind of information the instrument should provide us with, and what kind of judgements should be made.

It was decided at the outset that two languages should be included in this feasibility study. This would enable us to experience some of the issues involved in creating parallel

instruments while at the same time would impose reasonable limits on the piloting. French and English were chosen as these were the languages taught by the committee members, who therefore had more direct access to the students concerned. It was felt that language-majoring exiting and near-exiting students could be expected to have a level of language proficiency at least no lower than B2 on the CEFr scale. It was decided that, for practical/timetabling reasons, the entire procedure should not take longer than about one and a half hours.

In order that the instrument should serve the purposes outlined above, it had to be designed to elicit different kinds of information. Firstly it had to give some indication of what CEFr level the individual students should be considered to be 'at'. Secondly it had to give a test score that could then be roughly linked to this level. Thirdly it had to give information on individual *can dos*, in order to see how these might be incorporated in adapted descriptors of ability. And as these *can dos* are inextricably linked to different texts, some indication should be given as to which texts seem to suit students at different levels, so that ultimately these could be analysed and described in the adapted CEFR levels.

In order to elicit all of this information, it was decided that the instrument should involve both self-assessment (SA) and testing. SA has made considerable gains in status recently, not only due to its potential in enhancing assessment, but because of its benefit to the learning process (Oscarson 1999). Moreover, counter to popular belief, Oscarson demonstrates that there is empirical evidence that SA is a dependable way of assessing ability. Self placing on the CEFr, either as it stands or in adapted forms is becoming accepted practice, eg in DIALANG, and in testing projects in individual countries, eg in a project involving the assessment of Norwegian as a foreign language (see Norsk Språktest 1998). It is particularly the case for the receptive skills, such as reading, that no one knows as well as the learners what they have understood or can do. Moreover, it is difficult to accrue evidence of what learners can do with a text solely on the basis of test answers which may be influenced by irrelevant factors.

For these reasons, the instrument was given the following format. Students were asked to place themselves on the CEFr scale of reading ability. A number of texts were given (three or four) at different levels of difficulty. Each of these texts was accompanied by a set of test questions, each of which was intended to target a specific ' *can do*' from a set of core *can dos* identified as being salient to language majoring students. Moreover, students were required, after a first reading of the text and before doing the test questions, to indicate on a five-point scale how confidently they were able to 'do' the specified things for the particular text. They were also asked to rate texts in terms of difficulty, familiarity and interest. Slightly modified versions of the instrument in English and French are shown in Appendix C and D.

Teachers were also given the option of placing individual students on the scale. Scores were given on the basis of selecting the right answer from a multiple choice set. A database was created for compiling all test scores, self-assessment answers and self/teacher placings. It was arranged that the Centre for Applied Language Studies, University of Jyväskylä would analyse this data, in order to see how successful the instrument appeared to be in eliciting the kind of information intended.

2.2 Reading Comprehension

In recent years the field of reading has been studied extensively, not just by linguists, but also by experts in other disciplines, such as psychology, sociology, anthropology, philosophy and artificial intelligence. Many different approaches have emerged along with perhaps as many interpretations. It is beyond the scope of this report (and beyond the capacity of the TNP Committee to even try) to provide a comprehensive survey of the work that has been done in this field. In this section the complexity of this field will be

illustrated with a few examples; the nature of reading ability will be briefly discussed and then related to the required level(s) in Higher Education as outlined by the TNP Committee in its feasibility study.

There is general agreement among scholars that reading is a complex cognitive activity, but so far no one has come up yet with a conclusive definition of what reading is exactly. The definitions range from general views of reading to highly complex models. Thorndike for example, in 1917, saw reading as a process similar to mathematical problem solving. Widdowson, in 1979, defined reading as 'the process of getting linguistic information via print'. Carver (1977) came up with a more complex description of reading; he described it as a linear process from graphic symbols to meaning responses where the reader checks words individually and sounds them out phonetically. Gray (1960) and Robinson (1966) presented a model in which reading consists of four activities: word perception, comprehension, reaction to what is read, and assimilation of what is read through the fusion of old and new ideas. The model put forward by Smith (1988) describes reading in the following way:

> Features of sequences of words may be analysed but the letters themselves do not need to be identified when the reader's objective is the identification of words. And features of words may be analysed without the words themselves being identified when the purpose of reading is to find specific kinds of sense in the text. Readers can go straight to meaning in the text by means of prediction. Reading is not a matter of identifying word after word. (Smith 1988:285)

Carver's definition is one of the 'bottom-up' models which describe direction of processing is from 'bottom-level' features of text to 'higher levels', that is, from the identification of letters to sounds, to words, to sentences, and finally to meaning and thinking.' (Davies 1995:169) Smith's view is one of the 'top-down' models of the reading process 'which predict that the processing sequence proceeds from predictions about meaning to attention to progressively smaller units, for example letters, visual features.' (Davies 1995:175).

In 1977 Rumelhart proposed an 'interactive' model as an alternative, as a model that attempts to account for both bottom-up and top-down processing. In his view the process of reading 'begins with a flutter of patterns on the retina and ends (when successful) with a definite idea about the author's intended message. Thus reading is at once a 'perceptual' and a 'cognitive' process. Moreover, a skilled reader must be able to make use of sensory, semantic and pragmatic information to accomplish his task. These various sources of information appear to interact in many complex ways during the process of reading.' (Rumelhart 1977:573-4).

The theories mentioned above apply to reading in a first language. Whether or not the same models can be applied to reading as a second or foreign language has been the subject of much study and debate. Bilingual studies have shown that at higher levels of language competence there is a fair degree of relationship between a reader's ability in the first and in the second language. For low levels of foreign language competence foreign language reading appears to be more a language problem than a reading problem, according to Alderson (1984). For a more detailed discussion on the development of the understanding of reading ability, see Clapham 1996.

The TNP Committee was concerned with Higher Education only, and therefore confined itself to the level of reading comprehension that goes beyond decoding written symbols: to the level of understanding whole texts, where the reader's reading skills in the first language appear to transfer largely to the foreign language.

Several institutions (as well as individuals) have tried to describe an ascending series of levels for foreign language reading comprehension. The TNP Committee studied in particular the scales developed by ACTFL, where ten levels of reading competence are described, and the scales developed by the Council of Europe, which, in its CEFr, outlines six levels of proficiency in reading (as well as in the other 'skill's).

The levels outlined by **ACTFL** are: Novice-Low, Novice-Mid, Novice-High, Intermediate-Low, Intermediate-Mid, Novice-High, Advanced, Advanced-Plus, Superior and Distinguished. The three highest levels seem appropriate for Higher Education. They have been defined as follows:

Advanced Plus:
Able to follow essential points of written discourse at the Superior level in areas of special interest or knowledge. Able to understand parts of texts which are conceptually abstract and linguistically complex, and/or texts which treat unfamiliar topics and situations, as well as some texts which involve aspects of target-language culture. Able to comprehend the facts to make appropriate inferences. An emerging awareness of the aesthetic properties of language and of its literary styles permits comprehension of a wider variety of texts, including literary. Misunderstandings may occur.

Superior:
Able to read with almost complete comprehension and at a normal speed expository prose on unfamiliar subjects and a variety of literary texts. Reading ability is not dependent on subject matter knowledge, although the reader is not expected to comprehend thoroughly texts which are highly dependent on knowledge of the target culture. Reads easily for pleasure.

Superior-level texts feature hypotheses, argumentation, and supported opinions, and include grammatical patterns and vocabulary ordinarily encountered in academic/professional reading. At this level, due to the control of general vocabulary and structure, the reader is almost always able to match the meanings derived from extralinguistic knowledge with meanings derived from knowledge of the language, allowing for smooth and efficient reading of diverse texts. Occasional misunderstandings may still occur; for example, the reader may experience some difficulty with unusually complex structures and low-frequency idioms.

At the Superior level the reader can match strategies, top-down or bottom-up, which are most appropriate to the text. (Top-down strategies rely on real-world knowledge and prediction based on genre and organizational scheme of the text. Bottom-up strategies rely on actual linguistic knowledge.) Material at this level will include a variety of literary texts, editorials, correspondence, general reports, and technical material in professional fields. Rereading is rarely necessary, and misreading is rare.

Distinguished:
Able to read fluently and accurately most styles and forms of the language pertinent to academic and professional needs. Able to relate inferences in the text to real-world knowledge and understand almost all sociolinguistic and cultural references by processing language from within the cultural framework. Able to understand a writer's use of nuance and subtlety. Can readily follow unpredictable turns of thought and author intent in such materials as sophisticated editorials, specialized journals, articles, and literary texts such as novels, plays, poems, as well as in any subject matter area directed to the general reader.

The Council of Europe Framework proposal (CEFr) distinguishes several kinds of reading (Overall Reading Comprehension, Reading Correspondence, Reading for Orientation, Reading for Information, Reading Instructions) and outlines six main levels: A1 (Breakthrough), A2 (Waystage), B1 (Threshold), B2 (Vantage), C1 (Effective

Operational Proficiency)) and C2 (Mastery). The TNP Committee agreed that levels B2, C1 and C2 apply to Higher Education.

For 'Overall Reading Comprehension' the descriptors are as follows:

The student

B2
Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low-frequency idioms.

C1
Can understand in detail lengthy, complex texts, whether or not they relate to his/her area of speciality, provided he/she can reread difficult sections.

C2
Can understand and interpret critically virtually all forms of the written language including abstract, structurally complex, or highly colloquial literary and non-literary writings. Can understand a wide range of long and complex texts, appreciating subtle distinctions of style and implicit as well as explicit meaning.

When drawing up a framework for describing the reading ability of HE foreign language students for its feasibility study, the TNP Committee built on descriptors and scales developed by others, in particular by ACTFL and the Council of Europe. The results are described in section 2.4.

2.3. Test specifications

2.3.1. What are test specifications?

Test specifications explain what is tested and how it is done. They are essential guidelines for the test developer in particular. Therefore the more detailed these guidelines are the more adequate the test. Whereas test writers need to be well aware of the whole range of test specifications, others involved in the testing process like teachers, textbook writers providing the teaching material to prepare students for the tests and, of course, the test takers, require information about certain aspects of these specifications.

As we had decided to construct a reading test for exiting language students, the purpose of the test could only be a proficiency test since we intended to involve learners from different countries with different language learning backgrounds. It is assumed that these learners generally have a high level of proficiency; therefore an expanded version of the highest brackets of the CEFR scales was used for the assessment. The texts and the test items had to be selected with this high level of proficiency in mind.

2.3.2. Specifications for the Feasibility Study Reading Test

As previously stated, the test aimed at establishing the reading proficiency level of advanced learners in the foreign language. The time allowed for both the self-assessment section and the test itself was 60 - 90 minutes. The text types selected were:

- colloquial
- literary
- serious newspaper text
- academic text (review).

These types were all to be represented by authentic texts displaying complex language in different forms The language skills to be tested were:

- understanding the main idea(s) or main purpose of the text or parts of the text
- finding specific information
- making inferences
- distinguishing tone and style of texts
- understanding development and linking of ideas
- understanding the language (e.g. vocabulary).

The number of items were to be roughly the same (8-10) with equal weight given to each of them in the evaluation. The items had to be objectively markable. The test method to be used throughout was multiple choice.

2.4 Development and trialling procedures

The development and trialling of the instrument extended over a period from late spring 1998 early spring 1999. Some of the decisions taken in the course of the process, such as which languages to test and which levels of the CEFr were most appropriate, were arrived at by consensus during the TNP meetings in this period. However, others were taken as the result of, in many cases, a protracted period of work conducted by members working individually and as a group. This work can be considered to have taken place in four phases:

- preliminary work on compiling *can do* statements, suitable reading texts and accompanying tasks (April 1998 - October 1998)
- final selection of *can do*s and tasks and the laying down of test specifications and general specifications, format and characteristics of the instrument (October 1998)
- creation of tasks and final compilation of material to be included in the instruments in two languages (November 1998 - January 1999)
- trialling (February 1999).

The preliminary work began in Spring 1998 with local surveying of what students and teachers of HE language-majoring courses considered salient to reading ability at this level. This involved first giving open questions of the type, *what distinguishes between the reading ability of successful students at an early stage in their course, and at the end?* An example of the question posed to students is shown in Appendix E. At the same time, group members did a close study of a variety of higher level reading ability tests. The answers to the open survey questions, together with what group members themselves believed salient, e.g. on the basis of subskills which test questions typically targeted, were compiled by e-mail correspondence, and lists and charts were drawn up at various stages, classifying subskills of reading ability and possible modes of testing them. Examples of these are shown in Appendix F and G. The members then collected texts and provided sample questions which seemed to target the individual subskills identified.

At a meeting in Madrid in October, the final selection of *can do*s, texts and specifications were made. It was decided to limit the number of *can do*s to be individually targeted to a core group of five or six, and, by consensus, the following emerged as most salient.

1 understanding meaning (explicit and implicit, overall and detailed)
2 the structure of the discourse
3 the intention of the writer (especially if veiled)
4 the tone used
5 the style.

It was agreed that vocabulary would not specifically be targeted but the level would be implied by the texts associated with items. Texts should thus cover a range of difficulty, and a task of the group would be to try to identify texts which seemed suitable for different levels. A large number of texts were considered and three of four in each language were decided on. A number of these texts had accompanying questions, but it was agreed that these needed to be supplemented in order to target all the *can do*s. The format of the instrument, which the inclusion of a self placement on the CEFr and a assessment 'grid' to be used with each text (see Section 2.1.2) was also decided on in Madrid.

In the period following the Madrid meeting, an major tasks was the writing of test items for the chosen texts. Group members shared this task and sent round items to be considered in terms of suitability regarding their quality as reliable test items and their potential to target individual *can do*s. The self-assessment grid and self/teacher placing schema were finalised and the instruments were completed. These instruments, in both English and French, contained some tasks which were taken from existing tests, since the point here was not to develop material to be used in a finalised reading test, but to find out how the instrument worked in principle, using texts and tasks with certain given characteristics. For copyright reasons, the instruments have since been slightly adapted, and the versions shown in Appendix C and D contain only tasks developed by the group members.

The trialling of the instruments took place in February 1999 at English and/or French language departments the following six universities/colleges:

- Universitetet i Bergen (NO)
- Université de Lausanne (CH)
- Universidad Autónoma de Madrid (ES)
- Universität Potsdam (DE)
- Høgskolen i Hedmark (NO)
- Høgskolen i Sør-Trøndelag (NO)

Data was collected from 92 students (English) and 93 students (French) and sent to the University of Jyväskylä, for detailed analysis.

Chapter three: Data analysis results (pilot study)

Fellyanka Kaftandjieva & Sauli Takala

3.1. English test

3.1.1. Sample

The total sample size of English test piloting study was 92. The participants in the study were language students at universities (72%) or teacher training courses (28%) in three countries (Norway: 42%, Spain: 37% and Germany: 21%).

The level of language proficiency - reading comprehension according to CEFr scale - of the subjects taking part in the study was (according their self-estimation) as follows:

CEFr scale level	Frequency	Percent
B1	2	2.2
B2	38	41.3
C1	41	44.6
C2	4	4.3
Sub-total	85	92.4
Missing	7	7.6
Total	92	100.0

3.1.2 Pilot study design and instruments

The English Reading Comprehension Test used in the pilot study consisted of six separate parts, each of them including a text, a few reading comprehension items and/or 11 self-assessment (SA) statements as follows:

	Sub-test ST0 (stand-alone)	Sub-test A	Sub-test B	Sub-test C	Sub-test D	Total
Text for reading	0	1	1	1	1	4
Number of items	9	10	8	10	0	37
SA statements	0	0	11	11	11	3x11

The responses to all reading comprehension items are dichotomously scored and the responses to SA statements are in a 5-point rating scale (1 = not at all confident, 5 = absolutely confident)

3.1.3. Test Characteristics

TEST RELIABILITY
The average difficulty of the items included in the total test (37) is 59%, with the difficulty varying between 12 and 96 %. This wide range of item difficulty is one of the possible explanations for the comparatively low test reliability (Cronbach's a = 0.73).

The discrimination index has an average of 0.21 and varies between -0.05 and +0.48, while 16 of the items have a discrimination index below +0.20. Discarding the seven items with the lowest discrimination can increase test reliability up to 0.75 but, because of the small test length, higher reliability cannot be reached in this way. Taking into account the fact that the standard deviation of the raw test score is 4.72, the standard error of measurement is 2.47. This means that the interval estimation of reading

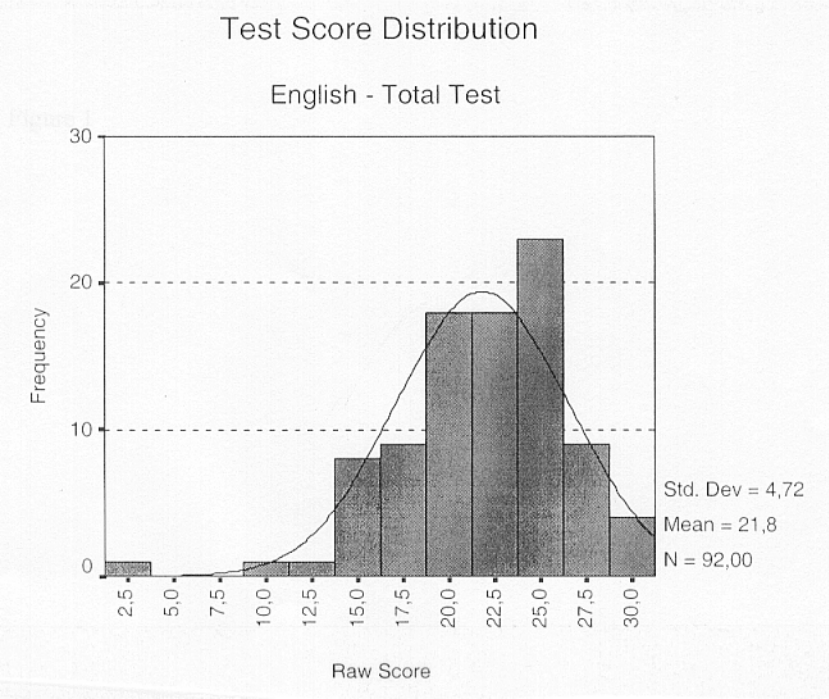
comprehension ability (English) with this test, at 95% level of confidence, is approximately X0 ± 4.84.

Due to the low reliability, the separation index for the total test is also low (1.63), which means that, based on the test results, test-takers can be reliably classified in no more than two classification groups (strata). This result is very important because it shows that the instrument would need further development, should it be used for real classification purposes.

More detailed results of reliability analysis are presented in Appendix H.

TEST SCORE DISTRIBUTION

Figure 1

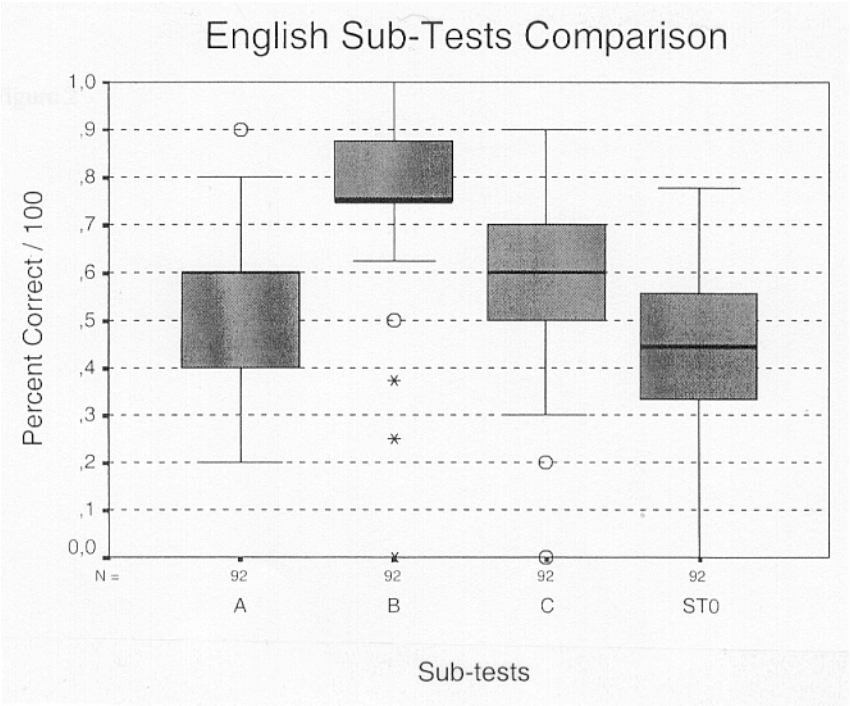


The test score distribution (Fig. 1) is normal (K-S *Z* = 0.984, r = 0.288) with a mean = 21.82 and SD = 4.72, and the total raw test scores vary between 3 and 31, which means that nobody gets a perfect score. On the other hand, the frequency distributions for the different sub-tests (A, B, C and ST = 9) are not normal and the difference between them is statistically significant for all possible pairs of sub-tests[1].

A comparative analysis of different sub-tests also shows that the four sub-tests differ considerably in their difficulty but due to the abnormality of their test score distributions it cannot be tested whether the differences between their average difficulty are significant, or not. As can be seen (Fig. 2), however, the easiest sub-test is Sub-test B and the most difficult one is Stand Alone Sub-test (ST0). Although the medians of sub-tests A and C are the same, sub-test C is comparatively easier than sub-test A.

[1]The more detailed results of statistical analysis can be found in Appendix H

Figure 2



The correlation coefficients between the sub-tests scores are statistically significant but comparatively low (Spearman's rho [+0.23; +0.42]), which may indicate that the different sub-tests measure different aspects of reading comprehension ability. However, all conclusions concerning interpretations of sub-test scores should take into account their low reliability, mainly due to the short length of the sub-tests (8-10 items).

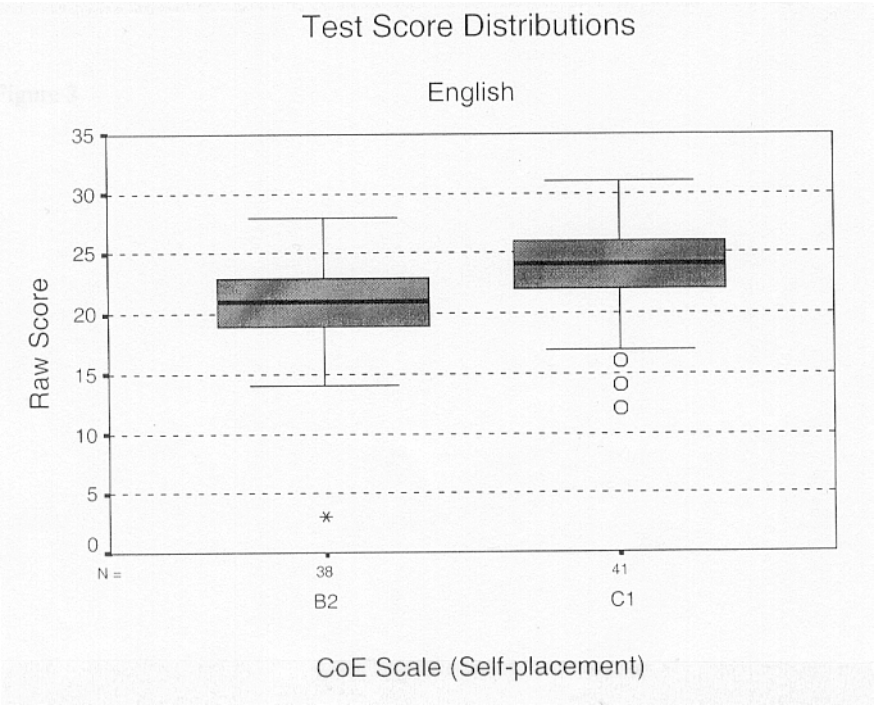
TEST VALIDITY

In this stage of test development only concurrent validity was investigated. Two external criteria were planned to be used: teacher assessment of language proficiency and student self-estimation.

Unfortunately the first criterion could not be used because although the language teachers of the participants in the pilot study were asked to assess the level of language proficiency of their students, this was done in the case of only 19 students. These students were all assessed to be at high levels of reading comprehension (B2, C1 or C2). Despite the observed trend of increasing test scores with increasing level of estimated language proficiency a more detailed analysis cannot be conducted due to the small size of the sub-samples ($n_{b2} = 6$, $n_{c1} = 9$, $n_{c2} = 4$). On the other hand, as has already been mentioned, there was another source of information about the level of students' reading comprehension ability: their self-estimation. This made it possible to compare the test results of the subjects, who estimated their ability as being at level B2 ($n_{b2} = 38$) or level C1 ($n_{c1} = 41$).

The two frequency distributions of test scores are normal ($Z_{B2} = 0.944$, r = 0.335 and $Z_{C1} = 0.944$, r = 0.335) and the difference between means for these two levels is statistically significant (t = -3.147, r = 0.002) in favour of higher level students. The two frequency distributions differ significantly (K-S Z = 1.747, r = 0.004) as can be seen in Fig. 3.

Figure 3

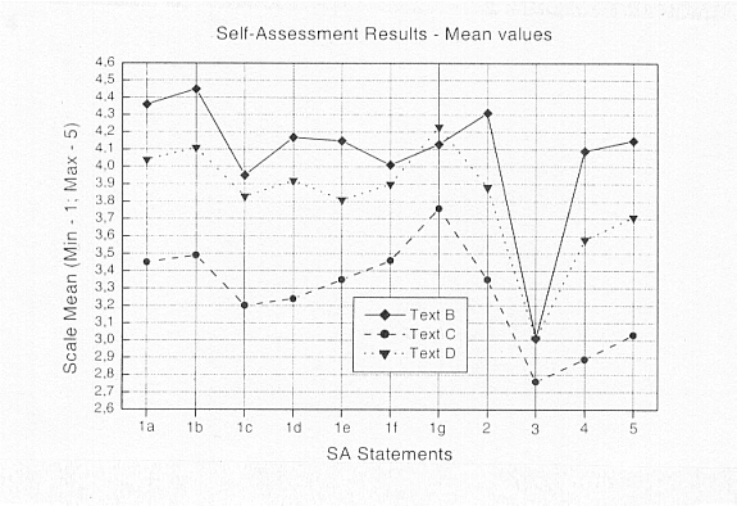


If we assume that the participants estimated their level of language proficiency relatively correctly, we can conclude from these results that the English Reading Comprehension Test might become a good instrument for measuring reading comprehension ability at higher levels if it is improved in terms of item homogeneity and test reliability.

3.1.4. Self - Assessment

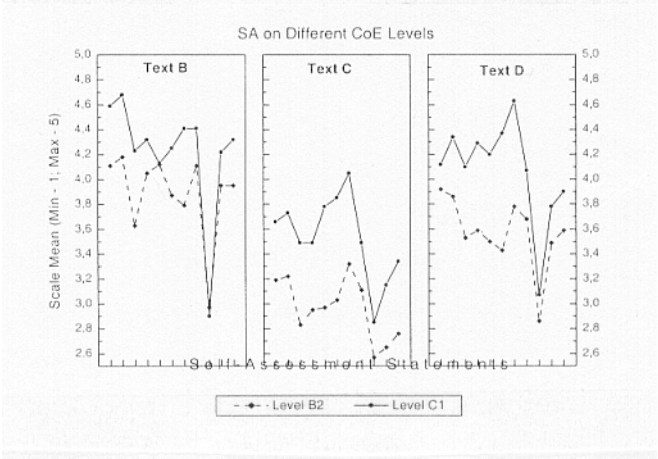
The results for self-assessment (SA) statements are not normally distributed (Appendix H) for any of the texts and therefore no parametric statistical tests for the analysis of the differences can be applied. The summary of the results (Fig. 4), however, shows quite a clear trend - Text B is regarded as the easiest and the most comprehensible and interesting text, despite the fact that the subject matter is not particularly familiar, in comparison with the other two texts. It confirms the previous conclusion that sub-test B is the easiest one. Obviously in this case, the difficulty of the test is closely related to the difficulty of the text itself.

Figure 4



Text D, for which there were no any language test items, is placed somewhere between texts B and C with one exception only (SA 1g) that its language style is the most recognisable. In general, its characteristics seem to be closer to those of text B than to those of text C.

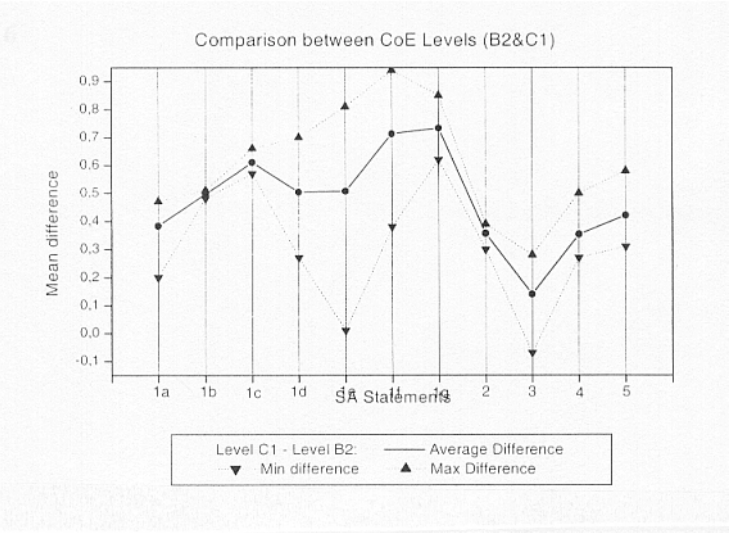
Figure 5



As far as the comparative analysis of self-assessment results at different levels of language proficiency (self-estimation) is concerned, almost one and the same pattern can be observed for all three texts (Fig. 5). All texts seems to be more understandable, more interesting, easier and in some degree even more familiar (in terms of subject matter) for participants who estimated being at a higher level of language proficiency (reading comprehension).

Although it cannot be checked whether the mean differences are statistically significant or not, the trend is obvious and it seems (Fig. 6) that the ability to recognise the style (SA 1g) and interpret correctly the writer's tone (SA 1f) distinguish best students at level C1 from students at level B2. On the other hand, familiarity (SA 3), interest in the subject matter (SA 4), and smoothness of reading (SA 2) do not appear to greatly affect reading comprehension at higher ability levels. This is, however, only a hypothesis which needs further investigation in order to be confirmed or rejected.

Figure 6



In general, the results of SA analysis confirmed that the participants in the pilot study demonstrated a high level of consistency in the self-assessment. This lends support to the claim of reliability of their self-estimation as an external criterion and in this way to give some additional support to the test validation.

3.2. French Test

3.2.1. Sample

The number of participants in the French test piloting study was 93, all of whom were language students at the Universities of Lausanne (74%), Madrid (23%) or Bergen (3%). According to the self-estimation of the participants in the study, their level of language proficiency - French reading comprehension - was as follows:

CEFr scale level	Frequency	Valid percent
A2	1	1.1
B1	25	27.2
B2	40	43.5
C1	24	26.1
C2	2	2.2
Total	92	100

3.2.2 Pilot study design

The French reading comprehension test consists of 3 sub-tests (A, B, and C), which include a text, 10 language items and 4 self-assessment statements (the same for each of the sub-tests) concerning the text comprehension.

The language items were scored dichotomously (0 = wrong, 1 = correct). For the self-assessment statements a three-point (SA1, SA2, SA3) or five-point (SA4) rating scale was used with the lowest category indicating that texts were easy, familiar, interesting and the highest category indicating difficult, less familiar, less interesting texts. (This was, in fact, in opposition to the rating scheme used in the English self-assessment system, where 'high' was equated with 'positiveness'!)

3.2.3. Test Characteristics

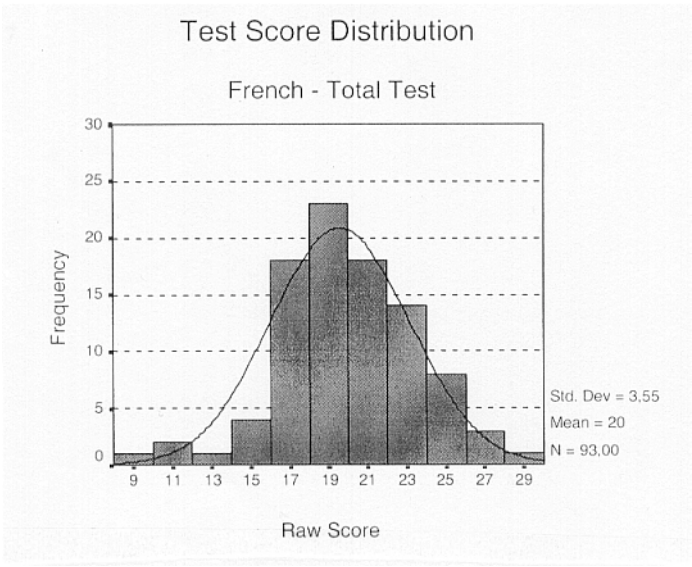
TEST RELIABILITY
The average difficulty of the items is 65%, but this varies over a very large range - between 14% and 99% - offering a possible explanation for the comparatively low test reliability (Cronbach's a = 0.58). The second possible explanation is the considerable heterogeneity of the item pool (the average inter-item correlation is 0.05) which leads to lower item discrimination and consequently lower reliability. The item discrimination averages 0.17, and varies between -0.17 and +0.37.

The maximum reliability, which can be reached discarding the 17 less discriminating items is 0.672 but even this is not satisfactory for classification purposes. The standard error of measurement with the original 30-items test is equal to 2.29 and therefore the interval estimation of reading comprehension ability at 95 % level of confidence will be approximately X0 ± 4.49, in other words, an interval of about 9 units when the range of the scale is 31 units, which is not very precise measurement.

The separation index, on the other hand, is equal to 1.18. Since this number is smaller than 2, the test as it stands cannot be used for classification purposes because it cannot distinguish more than one (1.18) ability strata. The more detailed results of reliability analysis are given in Appendix I.

TEST SCORE DISTRIBUTION
The test score distribution (Fig. 7) is normal (K-S *Z* = 0.874, r = 0.429) with a mean equal to 19.57 and standard deviation equal to 3.55. The raw test scores vary between 9 and 28 points, which means that - as in the case of English tests - nobody received a perfect score (30).

Figure 7

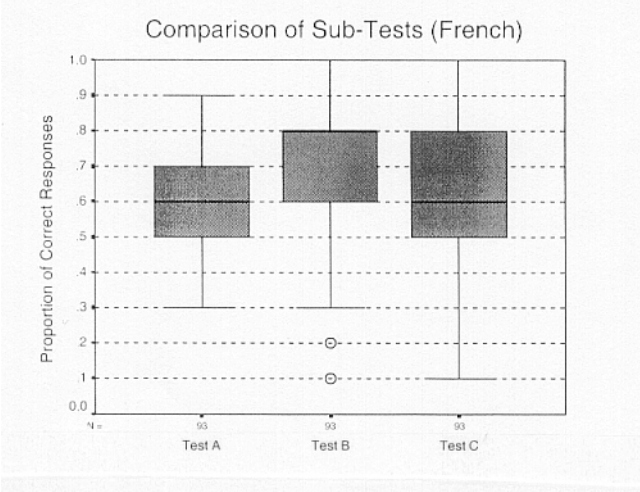


The score distributions (proportion of correct answers) for the sub-tests on the other hand are not all normally distributed and sub-test B differs significantly from the other two in terms of frequency distribution and difficulty.[2] As can be seen in Figure 8, Sub-test B is comparatively easier than the other two sub-tests and its score distribution is more clustered to the central point.

[2] See Appendix I for more statistical details

Although only one of the sub-tests A and C has a normal frequency distribution (sub-test C), the two tests do not differ significantly either in their difficulty or in their test score distributions. Sub-test A, however, is characterised by the smallest range and variance. The correlation between the three sub-tests is comparatively low, although sub-test A correlates significantly with the other two (r_{AB} = +0.22, r_{AC} = +0.37). This is an indication that the three sub-tests measure different aspects of reading comprehension ability and the total test as whole probably is multidimensional.

Figure 8

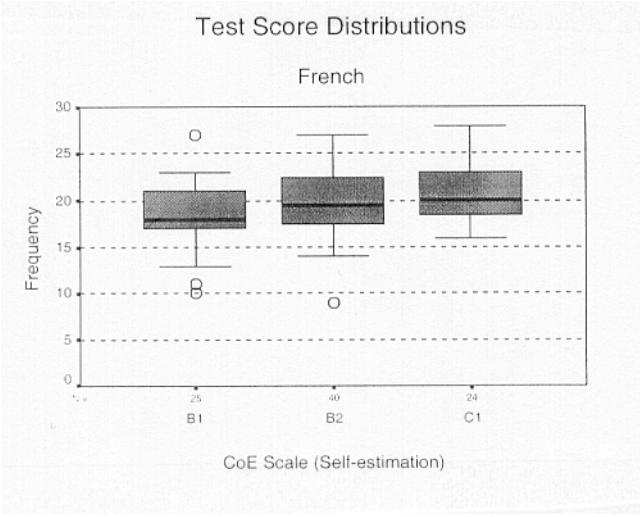


TEST VALIDITY

The self-estimation of language proficiency was the only external criterion used at this stage of test validation. According to this self-estimation, 25 of the participants were at level B1, 40 at level C2, and 26 at level C2. The sample sizes for the other levels (one for level A2 and two for level C2) were not large enough for these levels to be included in the analysis.

The test score frequency distributions for each one of the analysed levels (B1, B2, and C1) are normal (see Appendix I) and, as can be seen in Figure 9, there is a trend towards higher test scores among students at the higher levels of self-estimated language proficiency.

Figure 9

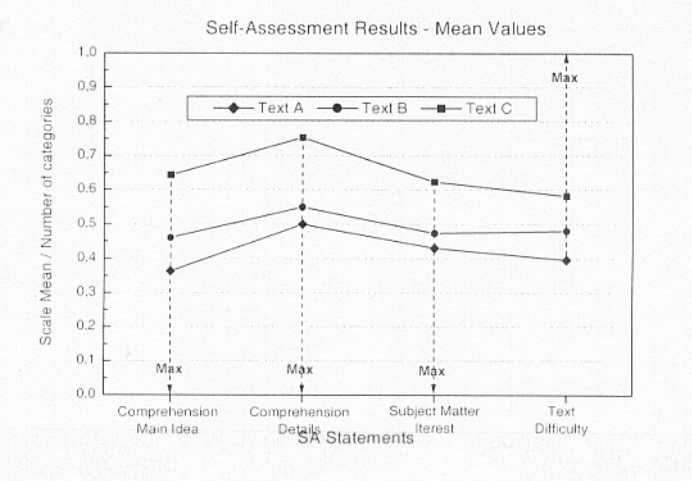


This trend, however, is not so clear as in the case of the English test, because the only significant difference is between the means of raw test scores of levels B1 and C1 (t = -2. 427, r = 0.019). The possible reasons for this might be the smaller sample size, which leads to less power of the statistical testing and/or the lower test reliability and consequently less precise measurement of language proficiency. Of course, the validity of self-estimation also needs some empirical support, which in this case is missing.

3.2.4. Self-Assessment

All frequency distributions for the self-assessment part of the test are not normally distributed (see Appendix I), which limits the options for statistical analysis. There are, however, some trends which are obvious and can be easily observed, although they cannot be tested statistically. As can be seen in Fig. 10, text A is the most comprehensive and interesting text, while text C seems to be the most difficult one. Text B, on the other hand, is somewhere between texts A and C, but closer to A than to C in terms of the measured characteristics.[3] (NB! Note that the graph must be read differently to that in Figure 4 (English). In the case of French, 'high' = 'negative')

Figure 10

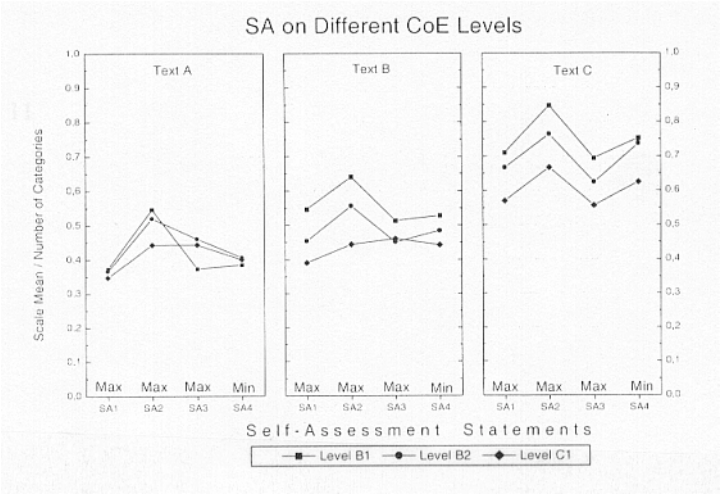


The findings shown in Figure 10 apply not only to the total sample as whole, but also to the sub-samples for different CoE levels (Fig. 11). The contradiction between these results and the previous comparison of the three sub-tests might be due to the fact that test difficulty depends only partly on the text difficulty, and the difficulty of the items is the main factor determining test difficulty. The other possible explanations might be the low test reliability, the possible low reliability of self-assessment scales and the abnormality of the frequency distributions.

On the other hand, all texts are on the average more comprehensive (in terms of main ideas and details - SA1 and SA2) for the higher level students. Consequently, as can be expected, the students who estimated themselves as being at lower levels of language proficiency rated at least two of the texts (B and C) as more difficult (SA4). There is, however, one exception of this pattern - text A, which participants with higher language proficiency (B2 and C1) considered more difficult than the participants at the lower (B1) level.

[3] For the purposes of comparative analysis, the original average scale values were divided according to the maximum number of scale points (three for SA1, SA2 and SA3 and five for SA4).

Figure 11



The interest in the subject matter, again as in the case of the English test, does not seem to greatly affect reading comprehension at higher ability levels. In general, the results based on self-assessment for English and French are similar at least for levels B2 and C1 and support the use of self-assessment as one of the external criteria in validation studies.

CONCLUSION

This chapter presents the statistical analysis of the tests of English and French reading comprehension. For each language, the chapter first describes the sample, the self-estimated level of reading comprehension in terms of the Council of Europe CEFr proficiency scale, and the number of language test and self-assessment items. This is followed by a fairly detailed analysis of the test characteristics: test reliability, score distributions, and test validity. The results concerning self-assessment are also reported.

The analyses show that the tests need further development if they are to be used for classifying students at the exit point. The results suggest areas where concrete development work is required. The results also suggest that self-assessment - as it was used in the study - may play a useful role in the validation of the tests.

Chapter four: a final word

Angela Hasselgren

This document has described in some detail the aims, approaches and activity of the TNP no. 10 from its rather rushed beginning in 1997 to its end, in its present form in 1999. Although the feasibility study revealed much about the weaknesses of the piloted instrument produced by the group, I believe it also showed that we were on the right course. Working as a group of professional people already heavily committed to our own employment situations, we were forced to economise on such things as the time spent on making test items and the size of the pilot population. However, the study did what it was intended to, in that it gave an indication of what was possible to do, given more roomy circumstances, in the way of

- identifying core subskills specific to the proficiency of language (under)graduates
- identifying the levels on the CEFr which seem to approximate this proficiency
- identifying texts appropriate to levels on this scale, and hopefully characterising these in terms of key features
- self-placing of learners on an scale thus adapted from the CEFr
- placing of learners on the scale through test scores
- self-testing of students on the core subskills
- ultimately validating the scale.

The conclusions on the above points create a basis for a blueprint for a full-scale instrument for the assessment of the proficiency of language (under)graduates. The creation of such an instrument, together with the compilation of a set of general guidelines, and increased training opportunities for those involved in language assessment in HEs would go a long way towards meeting the needs of HE institutions in their assessment processes, as highlighted in the early stage of the TNP work.

Mindful of these aims, and with the experience gained during the feasibility study fresh in mind, the group finalised its activity at two events in Jyväskylä, July 1999: the workshop on testing/assessment at the 2nd ELC Conference, and the TNP no. 10's last meeting.

4.1 The workshop on testing/assessment at the 2nd ELC Conference

The workshop consisted of the following four sessions:

- DIALANG
- The Council of Europe's European Language Portfolio
- New Technologies in Language Testing
- A Pilot project for Exit Tests

DIALANG
Presenters: Sauli Takala, Ari Huhta, Fellyanka Kaftandjieva (University of Jyväskylä), Neus Figueras (University of Barcelona)

An overview was given of the aims and intentions of DIALANG, its practical implementation and the findings of the data analysis on its recent trialling for the testing of Finnish. The discussion which followed reflected the genuine interest of the participants in DIALANG, which is of major importance as a potential model for the incorporation of new methodologies in language testing in higher education.

The Council of Europe's European Language Portfolio
Presenter: Rolf Schärer (formerly of Eurocentres)

An general introduction was given to the Portfolio, and ways of implementing it in HE were exemplified. The lively discussion reflected the awareness of the group of the need for radical rethinking of the way assessment is carried out in HE, and an openness to the idea of the portfolio in this context. It was also evident that a number of questions need to be addressed regarding the practical issues involved in portfolio use and the suitability of the current CEFr to the language proficiency of language majoring HE students.

New Technologies in Language Testing
Presenter: Glenn Fulcher (University of Surrey)

The participants were introduced to the three main types of computer-based language testing: CBTs (Computer-Based Tests), CATS (Computer Adaptive Tests) and WEBCBTs (Web-based CBTs). These types of testing were first illustrated and practical advice given on the procedures and software available for computerising tests. Participants were then given the opportunity to try out computerised tests in the lab. The questioning reflected the interest among the group in making use of computerised testing, although, as might have been expected, a certain scepticism was expressed, eg concerning the security aspects of examining by computer and the limitations of the computer, particularly in the testing of productive language skills.

A Pilot project for Exit Tests
Presenters: Raymond Capré (Université de Lausanne), Angela Hasselgren (Universitetet i Bergen)

In this session the two major outcomes of work of the TNP no. 10 (surveys and testing instruments) were presented to the group. The participants were given the opportunity to reflect on their own situation regarding testing and were also confronted with some of the main issues that the group had addressed in the course of their work, such as that of expanding the CEFR to capture what (under)graduates ' *can do*' in their language reading ability.

Outcomes of the workshop
The outcomes of the workshop consisted of two types. The first concerned what the individual group members gained from the workshop, in order to be able to influence the assessment processes in their own institutions. The second concerned suggestions for joint European activity to be carried out in the area of assessment in HE.

4.2 The last meeting of sub-project 10

The suggested activities emerging from the workshop formed the basis for the discussion at the final meeting of the TNP group. The suggestions which were incorporated in TNP no. 10's proposals for future action can be summed up as follows:

1 A full-scale project for the testing of the written language skills of (under)graduates majoring in language at European HEs. This project would draw on the findings of TNP feasibility study, centred on CEFr.

2 An in-serve training course in language assessment for HE teachers, e.g. as an activity to be carried out in the programme of the European Centre for Modern Languages (ECML) in Graz.

3 A website on which to continue the surveying and pilot testing work initiated by sub-project 10. Those involved in language assessment in HE are encouraged to use this website, both in the interests of their own assessment

practices, and in order to further consolidate the findings of this important preliminary research.

A plan of action was laid down at the meeting for concrete ways of realising the above suggestions. A proposal for an in-service training course has now been submitted to the ECML (August 1999), and work on the website will commence during the autumn of 1999. (In the meantime, progress can be followed via the ELC-website). A proposal for the large scale project will be submitted to a major European funding body in year 2000 or 2001. It is hoped that the website will not only elicit feedback on the work of the group, but will also attract new potential partners in the work initiated by the group.

Finally .....

This work has, so far, been difficult, frustrating and time-consuming. It has also been immensely challenging, interesting, revealing and, in the end, rewarding. Although the future of TNP no. 10, as it stands, has not been decided, there is a clearly expressed wish among the group for a continuation of their activity. This activity has so far succeed in opening up an area too-long closed, and has shown some clear routes to take, which will enhance the assessment of languages in HE, providing yardsticks for comparing student groups, and drawing on and linking up with the invaluable work already undertaken within leading European language testing projects.

REFERENCES

- ACTFL: see ACTFL Guidelines: http://carla.acad.umn.edu/ACTFLguideln.html
- Alderson, J. C. and Urquhart, A. H. (eds.)1984. *Reading in a Foreign Language*. London: Longman.
- Carver, A. P. 1977. 'A Theory of Reading Comprehension and Rauding'. *Reading Research Quarterly 13: 8-64.*
- Clapham, C. 1996. *The Development of IELTS*. Cambridge University Press.
- Council of Europe. 1996a. *Common European Framework of Reference for Language Learning, Teaching and Assessment.* Strasbourg: Council for Cultural Cooperation.
- Council of Europe. 1996b. *European Language Portfolio Feasibility Study.* Strasbourg. Council for Cultural Cooperation.
- Davies, F. 1995. *Introducing Reading*. London: Penguin.
- Gray, W. S. 1960. *The Major Aspects of Reading*. In H. M. Robinson (ed.) 1960.
- Huhta, A, V. Kohonen, L. Kurki-Suonio and S. Luoma (eds) 1997. *Current Developments and Alternatives in Language Assessment.* University of Jyväskylä.
- Kaftandjieva, F. 1999. DIALANG: *Some Results from Data Analysis*. Paper given at workshop on testing: language assessment in higher education. European Language Council 2nd Conference. July 1-3, 1999, Jyväskylä.
- Norsk Språktest 1998. *Vurdering av funksjonell norskferdighet hos voksne fremmedspråklige på AMO-kurs, tiltak eller arbeidsliv, som har avlagt Språkprøven i norsk/Mellomnivåtesten - et samarbeidsprosjekt mellom Arbeidsdirektorat, IFF-Oslo og Norsk språktest*. University of Bergen: unpublished report.
- North, B. 1997. "The Development of a Common Framework Scale of Descriptors of Language Proficiency Based on a Theory of Measurement" in Huhta, A, V. Kohonen, L. Kurki-Suonio and S. Luoma (eds.) 1997.
- Oscarson, M. (1999). "Estimating Language Ability by Self-Assessment – A review of some of the issues". In *Papers on LANGUAGE - Learning, Teaching, Assessment. Festskrift till Torsten Lindblad*. Rapport från Institutionen för pedagogik och didaktik, Göteborgs universitet.
- Robinson H. M. (ed.) 1966. *Sequential Developments of Reading Abilities*. Supplementary Educational Monographs 90, Chicago IL: University of Chicago Press.
- Rumelhart, D. E. 1977. *Introduction to Human Information Processing,*. New York: Wiley.
- Smith, F. 1988. *Understanding Reading*. New York.
- Thorndike, E. L. 1917. 'Cognitive Structures in Comprehension and Memory of Narrative discourse.' *Cognitive Psychology*, 9, 77-110.
- Widdowson, H. G. 1979. *Explorations in Applied Linguistics*. Oxford University Press.

ELC website: http://www.fu-berlin.de/elc/