ARI HUHTA

INNOVATIONS IN DIAGNOSTIC ASSESSMENT AND FEEDBACK:
AN ANALYSIS OF THE USEFULNESS OF THE DIALANG LANGUAGE ASSESSMENT SYSTEM
Ari Huhta

Innovations in diagnostic assessment and feedback:
An analysis of the usefulness of the DIALANG language assessment system


Academic dissertation to be publicly discussed, by permission of the Faculty of Humanities of the University of Jyväskylä, in the Buiding Villa Rana, Auditorium Paulaharju, on February 27th at 12 o’clock noon.
ABSTRACT

The thesis analyses the quality of DIALANG, which is a multilingual language assessment system that provides learners with immediate, on-line feedback about the strengths and weaknesses in their proficiency. Some of the evidence for its quality is provided by the present author in the articles appended to this thesis; other evidence comes from a range of studies summarized in the synthesis part of the thesis. The quality evidence divides into theoretical arguments that were presented in the design documents of DIALANG and into empirical evidence that was collected in the piloting and operational stages of the system. All this evidence is presented and analysed with reference to the Bachman and Palmer (1996) framework on test usefulness that uses usefulness as the super-ordinate term that characterizes the quality of a test or assessment system. According to Bachman and Palmer several factors contribute to the usefulness of a test for its intended purpose(s): reliability, construct validity, authenticity, interactiveness, impact, and practicality. In principle, the better the quality of each of these aspects is, the more useful the test is.

Available evidence does not allow us to obtain a balanced picture of all the aspects of test usefulness for DIALANG, neither are all its languages equally covered. The most comprehensive evidence concerns impact and construct validity. Overall, the system appears reasonably useful for the purposes that concern individual learners and their teachers – such as finding out the learners’ level of proficiency and their profile of strengths and weaknesses. A survey of over 550 DIALANG users by the author showed that a significant majority of the informants reacted positively to the system and found the feedback they got useful. User statistics show that hundreds of thousands of learners have taken DIALANG tests or visited its website – and even more must have been informed about the system by their teachers and institutions. The impact of DIALANG is greatest in the Netherlands, Germany, France, and Finland. Many institutions use the system for placing their students on language courses even though it was not developed for that purpose and thus lacks many features that would be useful for a proper placement test. Evidence also indicates certain problems with the system, most notably with the placement test, as well as with some technical aspects of the system, at least with some learners and institutions.

In the longer term the most important impact of DIALANG is probably its scientific impact on applied linguistics: it brought to researchers’ attention the fact how little we in fact know about diagnosing foreign language learning, which has spurred a considerable amount of new, interesting research into the diagnosis of language learning.

Keywords: language assessment, diagnosis of foreign language proficiency, diagnosis of foreign language learning, feedback, computerised testing, self-assessment
Author’s address
Ari Huhta
Centre for Applied Language Studies
University of Jyväskylä
Box 35
40014 University of Jyväskylä
Finland
ari.huhta@jyu.fi

Supervisors
Prof. Hannele Dufva
Department of Languages
University of Jyväskylä

Reviewers
Prof. emeritus Mats Oscarson
Institutionen för pedagogik och didaktik
Göteborgs Universitet
Sverige

Prof. Angela Hasselgreen
Senter för utdanningsforskning
Høgskolen i Bergen / Universitetet i Bergen
Norge

Opponents
Prof. emeritus Mats Oscarson
Institutionen för pedagogik och didaktik
Göteborgs Universitet
Sweden
This work is based on the DIALANG project (1996-2004), in which my own department, the Centre for Applied Language Studies (CALS) was involved, first as the coordinator and then as one of the four core project partners. I had the fortune to work in the project in many different roles, as is described in Appendix 2 of the synthesis part of this thesis. None of us who were involved in DIALANG at its early stages could foresee how long the project would last, nor how laborious and, sometimes, how frustrating the whole venture could be. However, designing this multilingual assessment system also turned out to be highly stimulating and rewarding for those participating in its development, as we had to face and solve a number of truly interesting and challenging problems along the way. Although we were able to tackle many questions about diagnosing strengths and weaknesses in language learners’ proficiency, ironically, the most important lesson we learned from the work on DIALANG was how little we in fact know about diagnosing second / foreign language proficiency and learning. Hopefully this realization will lead to more research into diagnosing proficiency so that if anything like DIALANG gets designed in the future, the result will be something that is significantly more useful to the individual learner than what DIALANG can be.

Considerable amount of research was carried out during the development of DIALANG to address the challenges facing the project, and studies on the system have continued after the project came to an end in 2004. Much of the earlier research is reported in Alderson’s 2005 book that brought our lack of knowledge on diagnosing second / foreign language learning to the attention of the language testing world. There is, however, no comprehensive account of the different kinds of studies that have analysed different features of DIALANG or that have used DIALANG as a tool to study other things and have, as a byproduct, yielded interesting information about its quality. The synthesis part of my thesis is an attempt to bring together as much information as possible about how DIALANG has succeeded (or in some cases, failed) to meet different expectations and to fulfill the purposes it was created for.

No study is possible without the professional and other support of many people. In addition to being lucky enough to participate in the design of DIALANG, I have been fortunate in having a chance to work in a very stimulating and supporting environment. I would like to warmly thank my former and present colleagues in the CALS and its predecessor, the Language Centre for Finnish Universities, and also elsewhere in the University of Jyväskylä as well as in several other universities and organizations in Finland and abroad, for countless meetings, discussions, and other exchange of ideas that have directly and indirectly contributed to my work on DIALANG and everything else that I have done during my career.

I would like to single out three persons who have profoundly influenced not only DIALANG but also my study on that system, as well as my own professional development in general: The late Professor and Director of CALS Kari Sajavaara, Professor emeritus Sauli Takala, and Professor J. Charles Alderson. I was first a student of Kari Sajavaara and Sauli Takala in the 1980s and then worked with them for about 15 years in numerous projects from the late 1980s until their retirement in the early 2000s. Our joint projects included, for example, the development of a national test for business and administration purposes (the TKD or Työelämän kieli-diplomi), the National Certificates examination system, the IEA Language Education Study, and, finally, DIALANG. Our co-operation continued even after their retirement, although, naturally, on a much smaller scale. My co-operation with Charles
Alderson also goes back a long time: it started with an international survey of standards in language testing in the mid 1990s, and was followed by other international projects, the most important of which was obviously DIALANG. Our collaboration continues in a research project aiming to shed more light on the diagnosis of second / foreign language learning. I owe Kari, Sauli and Charles much of what I know about applied linguistics, language test and research design, international networking, and project management.

I am indebted to a large number of people who helped me in the collection of the data for my own empirical studies on DIALANG. I am particularly grateful to the 557 language learners in Finland and Germany who took the trouble to fill in the extensive questionnaire that I used in my data collection. Some of them also consented to be interviewed by me. Equally important were the language teachers in numerous institutions who were already, or became during the study, interested in using DIALANG as part of their teaching and who encouraged their students to fill in my questionnaire. Some of the teachers, too, were kind enough to let me interview them about their experiences in using the system. Another vital group in the data gathering was the roughly 30 teachers and colleagues around Finland and at the CALS who volunteered to assist in the concept mapping study of self-assessment that is partly reported in one of the articles that form my thesis. Several students who worked at the CALS as trainees at different times helped me in getting my questionnaire and interview data ready for analyses – their assistance was simply invaluable and is gratefully acknowledged here.

Without the love, support and patience of my wife Anne and my sons Otto and Riku I could not have carried out my work. They have also shown me how there is life outside work by providing opportunities to relax at home in the evenings and on our trips to Lapland and even further away. I have been very fortunate in having a wife who is an academician and understands what research is like and the burden it often places on the family. Perhaps the amount of paper and books lying around in our house diminishes a bit now that I am finally completing the thesis – if I can only fit them in my office!

Palokka, February 15th 2010
Ari Huhta
LIST OF ORIGINAL PUBLICATIONS


VI Huhta, Ari (submitted & being revised) Concept-mapping – an approach to understanding self-assessment

Copies of these publications are appended to the end of this thesis.
Table of contents

ABSTRACT
PREFACE
LIST OF ORIGINAL PUBLICATIONS

1. INTRODUCTION ................................................................. 4
   1.1. Organisation of arguments and evidence for the usefulness
         of DIALANG in this paper ............................................. 5
   1.2. Validation before and after: the *a priori* and *a posteriori*
         approaches to validation ............................................ 9

2. THE BACHMAN AND PALMER FRAMEWORK OF TEST USEFULNESS
   2.1. Purpose of DIALANG .................................................. 11
   2.2. Summary of the Bachman and Palmer framework of test usefulness 15

3. ANALYSIS OF THE ARGUMENTS AND EVIDENCE FOR THE
   USEFULNESS OF DIALANG ................................................. 17
   3.1 Qualifications of the DIALANG development team ...................... 20
   3.2 Reliability of DIALANG ................................................ 24
       3.2.1 Theoretical / logical attempts to achieve reliability ......... 24
       3.2.2 Empirical evidence for the reliability of DIALANG .......... 26
       3.2.3 Reliability of self-assessment statements .................... 33
   3.3 Construct validity, authenticity, and interactiveness .................. 35
       3.3.1 Theoretical basis of the construct validity, authenticity,
            and interactiveness of DIALANG ............................... 38
            3.3.1.1 Diagnosis of L2 proficiency – what is it? .......... 39
            3.3.1.2 Language tests ........................................... 42
            3.3.1.3 Feedback and self-assessment ......................... 47
            3.3.1.4 Computer as a testing platform ....................... 53
       3.3.2 Empirical evidence about the construct validity, authenticity
            and interactiveness of DIALANG ............................... 55
            3.3.2.1 Evidence about reliability is also evidence about
                   construct validity ........................................ 55
            3.3.2.2 Concurrent, criterion-related validity evidence ........ 56
                   A. Internal structure of DIALANG test batteries ...... 57
                      a. The DIALANG pilot test study ..................... 61
                      b. Present author’s survey study of DIALANG users 68
                      c. Satakunta Polytechnic study ....................... 71
                      d. HISBUS English proficiency testing project ...... 75
                   B. Internal structure of DIALANG self-assessment
                      instruments .................................................. 76
                   C. Relationship of DIALANG tests with external measures
                      of language proficiency ..................................... 79
                         a. Self-assessments as concurrent validity evidence
                            for DIALANG language tests ........................... 79
b. HISBUS English proficiency testing project ..... 81

c. Satakunta Polytechnic study ......................... 85

d. Present author’s survey study of DIALANG
users ......................................................... 88

D. Relationship of DIALANG self-assessments with
other measures of language proficiency ............. 91

3.3.2.3 Learners’ reactions to DIALANG, its tests, self-assessment
instruments and feedback ............................... 99

A. Present author’s survey study of DIALANG users ..... 100

a. Users’ overall reactions to DIALANG .......... 100

b. Learners’ reactions to the VSPT ................. 102

c. Learners’ reactions to self-assessment and
self-assessment feedback ............................. 103

d. Learners’ reactions to other DIALANG
feedback ..................................................... 107

B. Other studies on DIALANG users’ reactions .......... 115

3.3.2.4 Other empirical evidence about DIALANG feedback ..... 117

3.4 Impact of DIALANG .................................................. 121

3.4.1 Impact on individual learners ........................... 122

3.4.2 Impact on language teachers, schools and other educational
institutions ................................................. 126

3.4.3 How many people are affected by DIALANG? ........ 129

3.4.4 Scientific impact ............................................ 137

3.4.5 Use of DIALANG as a research instrument .......... 142

3.4.6 Reviews of DIALANG for other purposes ............. 145

3.5 Practicality of DIALANG ............................. 148

3.5.1 Practicality of the computerized system underlying DIALANG .... 149

3.5.2 Practicality of the content (assessment instruments and feedback) .. 152

4. SUMMARY AND CONCLUSION .......................... 158

4.1. A priori theoretical arguments for the usefulness of DIALANG ..... 159

4.2. A posteriori empirical evidence for the usefulness of
DIALANG .................................................. 162

4.3. Overall usefulness of DIALANG ......................... 172

APPENDIX 1: Learners’ overall views on the best and worst aspects of
DIALANG .................................................. 177

APPENDIX 2: Own contribution in the two joint articles ..................... 192

APPENDIX 3: Concept map of the advantages of DIALANG
self-assessment based on learners’ sorting ................ 197

APPENDIX 4: Summary of the content of the six articles that form the thesis .... 199

References ......................................................... 201
THE ORIGINAL PUBLICATIONS APPENDED TO THE SYNTHESIS PART OF THE THESIS:


IV  Huhta, A. 2007b.


VI  Huhta, A. (submitted)
1. INTRODUCTION

This paper serves two intertwined purposes. First, it summarizes and synthesizes the content of the six articles on the DIALANG assessment system that comprise (together with this paper) the doctoral thesis of the present writer, namely Huhta & Figueras (2004), Huhta & Alderson (2005), Huhta (2007b), Huhta (2007a), Huhta (2008), and Huhta (submitted). The content of these articles is summarized in Appendix 4. However, the present paper also gives a more comprehensive overview of what we know about the usefulness of DIALANG on the basis of research by both the present author and by several other researchers who have either specifically studied DIALANG or who have used it as a data gathering instrument in their studies and who, as a byproduct of their research, provide us with information that sheds light on the usefulness of DIALANG. As a clarification, it ought to be mentioned at this point that the current author has carried out more research on DIALANG than what is reported in the six articles listed above, and that some of that research is reported in this summary paper whenever it contributes to empirical evidence about DIALANG. The thesis as a whole intends to provide answers to the following, rather broad questions:

- What do we know about the usefulness of DIALANG on the basis of the research and development work carried out and reported by the present author (both in the six articles and in this paper)?
- What do we know about the usefulness of DIALANG on the basis of research results, statistics on the users of the system, and other relevant information gathered by others working in the field of language education?
- What is the overall picture of the usefulness of DIALANG when we combine the two sources of information above?

Obviously, the empirically oriented studies by both the present author and the other researchers reviewed in this paper try to answer a number of considerably more specific research questions, but the three questions above capture the essence of what this thesis aims to accomplish.
1.1. Organization of arguments and evidence for the usefulness of DIALANG in this paper

The term *usefulness* is used in this paper because the present writer mainly relies on the Bachman and Palmer (1996) framework on test usefulness in order to organize the presentation: Bachman and Palmer use *usefulness* as the super-ordinate term that characterizes the quality of a test or assessment system. A detailed description of the Bachman and Palmer framework is presented in Chapter 2 below. Instead of ‘usefulness’, it would be possible to use terms such as ‘quality’ or ‘validity’ without radically changing the meaning of what the Bachman and Palmer framework is about.

Besides the Bachman and Palmer framework, literature on educational assessment and language testing provides other approaches that could be used in an overall analysis of the quality of an assessment procedure. Basically, most, if not all studies on language tests can be thought of as being some sort of validation studies of the tests in question: they aim at providing evidence and justification for different aspects of the validity of the tests or the interpretations made about the test takers on the basis of test scores.

Thus, it would be possible to build this review and synthesis of research on DIALANG on the traditional aspects of validity, and use such categories as *criterion-related* (or predictive) validity, *content* validity, and *construct* validity (see e.g. Cronbach, 1971). In fact, the number of aspects of validity that has been presented in the literature is very high: for example, Cummins and Berwick (1996) list about 20 types of validity. Although some aspects of validity are rather clear and easy to define, such a long list of validities would not be an ideal way to organize an analysis of validity evidence for a particular test. The great number of categories would in itself be a problem for a coherent analysis, and it may also be difficult to decide, in a principled way, which types of validity to use in and which to exclude from the analysis. In addition, validity theory has made progress since the days when validity was simply conceived of as a set of different but partially overlapping categories.

Probably the most widely used framework for conceptualizing validity is the one proposed by Samuel Messick in his extensive treatment of validity in 1989 (Messick 1989), which he later
presented in a more concise and slightly modified form (Messick 1995, 1996). According to Messick,

*Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment.* (Messick 1989, 13; emphasis in the original)

... validity is broadly defined as nothing less than an evaluative summary of both the evidence for and the actual – as well as potential – consequences of score interpretation and use … This comprehensive view of validity integrates considerations of content, criteria, and consequences into a construct framework for empirically testing rational hypotheses about score meaning and utility. Therefore, it is fundamental that score validation is an empirical evaluation of the meaning and consequences of measurement. As such, validation combines scientific inquiry with rational arguments to justify (or nullify) score interpretations and use. (Messick 1995, 742)

Messick’s validity framework has two facets (Messick 1989, 20). One facet is the basis of justification of the testing and consists of (1) evidence or (2) consequences. The other facet is the function or outcome of testing, which can concern either (3) test interpretation or (4) test use.

The **evidential basis of test interpretation** (i.e. the intersection of points 1 and 3 above) is construct validity (Messick 1989, 20). Messick (1989, 1995, 1996) further specifies that construct validity involves at least the following aspects. The **content** aspect of construct validity relates to the relevance, representativeness and technical quality of test content. The **substantive** aspect concerns evidence on test takers’ responses and the processes underlying those responses – the extent to which they match the processes predicted by the theory. The **structural** aspect concerns the match between the scoring criteria and rubrics used in the test and the construct domain to be tested. The **generalizability** aspect relates to the extent to which the interpretations drawn from the test results generalize across different groups of test takers and across different settings and tasks. As the generalizability of the interpretations across different occasions and raters is included in this aspect, the **reliability** of the test falls into this category. Finally, the **external** aspect of construct validity includes convergent and divergent evidence about relationships between the test and different external measures.
The **evidential basis of test use** (1 & 4 above) is also construct validity but with the added consideration of the relevance and utility of the test for some specific, applied purpose or setting. Thus, the focus of attention is the degree of match between the test results and some relevant criteria or decisions (e.g. whether test takers are correctly selected or placed on the basis of the results).

The **consequential basis of test interpretation** (2 & 3 above) concerns values implied by how the constructs are labeled, which theories underlie test score interpretations, and which ideologies are involved in those theories. Messick (op cit. 20) argues that the main issue in this is whether the theoretical implications and value implications are commensurate because values are always involved when test results are interpreted.

The fourth intersection of the two facets in Messick’s framework concerns the **consequential basis of test use** (2 & 4 above) and is about the planned and unplanned social consequences of the test. However, it could be argued that this fourth combination of the two facets of the validity framework also involves consideration of all the other intersections of the two facets, i.e., construct validity (1 & 3), the relevance and utility of the test (1 & 4), and value implications (2 & 3) (Messick 1995, 748).

The Messick framework could fairly easily be used to organize the presentation of validity arguments for DIALANG in the present paper since it has been used elsewhere for categorizing validation studies and for organizing the presentation and analysis of validity evidence for a particular test, either according to the main facets of validity, as has been done, for example, by Cumming (1996) and Kunnan (1997) or according the more specific aspects of validity (i.e., content, substantive, structural, etc.) as has been done, for example, by Hasselgreen (2004). In fact, Messick’s framework shares many characteristics with the Bachman and Palmer framework, as will become evident when the latter is described below. Thus, the choice between the two frameworks is largely a matter of preference. The decision to use the Bachman and Palmer approach was based on the fact that it is slightly more practically oriented – it in fact includes ‘practicality’ as one of the aspects of test usefulness – and that it appears to be less often used to organize the presentation of validity arguments. In fact, the author is not aware of it being used outside Bachman and Palmer’s (1996) book, in which they apply their framework to analyze several language tests. On the whole, it is unlikely that the present synthesis of research on DIALANG would have been substantially
different had it been based on the categorizations used by Messick. It is hard to imagine that, for example, any of the empirical studies reported in this paper would have been left out from this review or that the conclusions drawn from them about the quality and characteristics of DIALANG would have been significantly different.

However, there is one apparent difference between the two frameworks for analyzing validity arguments: the Bachman and Palmer framework does not directly address the consequential basis of test interpretation in Messick’s system. Thus, the consideration of the values underlying DIALANG might not be addressed if the present author strictly adhered to Bachman and Palmer’s categories. There are no studies on the value basis of DIALANG that the author is aware of but there is some critical discussion in the assessment literature of the values implied by the Common European Framework of Reference and of the potential dangers of its institutionalization and politicalization (e.g. Fulcher 2004).

One more validity framework that has recently gained popularity among language testing researchers should be mentioned before we move on to describe the Bachman and Palmer framework. That is the approach based on argumentation theory and concerns the development of interpretative and validity arguments for tests and other assessment procedures. It is based on Toulmin’s (1958, 2003) work on reasoning in non-mathematical sciences such as law and sociology, and has been applied in educational and psychological testing by Kane (1992, 2001) and Mislevy et al. (2003). Probably the best-known recent example from the field of language testing is the application of the argumentation theory in presenting validity evidence for the new TOEFL test (Chapelle et al. 2008). In this approach, validity arguments for any interpretation of the test scores are structured in a specific way into a series of claims, warrants (reasons) with appropriate backing, and possible rebuttals (examples against the claim).

Since the overall orientation of the above framework is substantially different from Messick’s, and Bachman and Palmer’s approaches to validation, it is not made use of in the present discussion of the validity arguments and evidence on DIALANG.
1.2. Validation before and after: the *a priori* and *a posteriori* approaches to validation

Finally, before describing the Bachman and Palmer framework in detail, it should be stated that the presentation of the evidence on the usefulness of DIALANG in this paper also makes some use of Weir’s (1991, 1993, 2005) work on validity theory. Weir makes the distinction between *a priori* and *a posteriori* validation of assessment procedures. *A priori* validation includes all non-empirical work that the test designers do during the development of the test, before it is piloted or used for real, which is why this type of validation can also be called theoretical validation (this, in fact, is similar to the point made by Messick (1989) above about theoretical rationales being one aspect of test validation). At the *a priori* validation stage, the designers attempt to ensure, with careful, knowledgeable planning that the test is as good and as valid as it possibly can.

*A posteriori* validation uses empirical data that are gathered when the test is actually administered to real language learners, either when the test is piloted or when it is has reached the operational stage – or, ideally, both. The empirical data used in *a posteriori* validation typically consists of test scores and ratings, which are analyzed statistically, but often also qualitative data are used to find out whether the test actually works as it should, for example, interviews or questionnaire surveys of test takers or other stakeholders, think-aloud data from test takers or analyses of responses to open-ended test items.

A well-grounded test design project relies on both *a priori* and *a posteriori* type of evidence about the quality of the test. In the case of DIALANG, the *a priori* evidence about the quality of the assessment system is mainly provided in the documents that guided the construction of the testing system: the DIALANG Assessment Framework document, test specifications, and a range of other documents that guided the development and review of the software, translations, feedback, and self-assessment instruments. These documents specify both the content of the various aspects of the assessment system and the procedures by which they were to be designed. It should be clarified that not all the documents mentioned above were created at the beginning of the project; the design and development of DIALANG took several years, from 1997 to about 2002. Some of the documents were needed and were thus created only after the initial phase of, for example, item writing was over, and these were intended to guide the reviewing and fine-tuning of the various parts of the system.
In addition to the documents, another key aspect of the *a priori* validation could be argued to be the personnel available to the test development project, although this is not specifically mentioned by Weir, for example. An experienced test design team is obviously more likely to be able to develop the documentation and procedures that ensure a high-quality test. Thus, in the case of DIALANG, it is also relevant to evaluate the level of **experience and expertise of the personnel** involved in developing the system.

As far as the *a posteriori* validation of DIALANG is concerned, it comprises all the empirical evidence gathered:

- by the project members themselves, during the development of DIALANG,
  - on the development of certain aspects of the system such as the scales used for self-assessment and for reporting test results (see Kaftandjieva & Takala, 2002);
  - on the piloting of the Finnish tests (which were the first ones to be piloted: the procedures developed for them were used in the data analysis and item selection for the other language tests (see Kaftandjieva, Verhelst, & Takala, 1999; Luoma 2004);
  - on the piloting of English (Luoma 2004; Alderson 2005; Alderson and Huhta 2005);
  - on the standard setting exercises (not reported except in project-internal memos and, to a limited extent, in Alderson 2005 and Alderson & Huhta 2005);
- by the project members, based on the operational version of DIALANG (e.g. Huhta 2007a, 2007b; Huhta, submitted; and this paper);
- by outsiders to the DIALANG Project, based on the operational version of DIALANG (some of publications listed in the column ‘Empirical evidence’ in Table 2)
2. THE BACHMAN AND PALMER FRAMEWORK OF TEST USEFULNESS

The framework that is mainly used in this paper to organize arguments and evidence about DIALANG is based on the notion of test usefulness developed by Lyle Bachman and Adrian Palmer (Bachman & Palmer, 1996), which consists of a number of elements or aspects, most of which also appear in the Messick framework of validity.

2.1 Purpose of DIALANG

Bachman and Palmer tie usefulness very closely with the purpose of the test: a test and its uses may be useful, to some degree, in terms of its intended purpose(s) (Bachman & Palmer, 1996). Although DIALANG has been used for several purposes, the number of its intended purposes is limited, and, furthermore, some of them are very closely related to each other. Thus, it appears feasible to use this purpose-dependent framework as a conceptual basis for organizing the presentation of research on DIALANG. Obviously, the more different purposes a test has the more complex the analysis and evaluation of its usefulness for each different purpose becomes. However, if the reality is complex, we should not expect the evaluation of such multi-faceted reality to be simple either.

Let us examine more closely the purpose of DIALANG and its feedback, as they have been described by the developers of the system. The DIALANG website (www.dialang.org) provides this information:

*Welcome to DIALANG, where you can learn about your strengths and weaknesses in a foreign language, and find out what level you are at. …*

*DIALANG does not issue certificates. It allows you to find out what level you are at and where your strengths and weaknesses are, so that you can decide how best to develop your mastery of a language.*

*DIALANG will help you to take control of your language learning and it will increase your awareness both of what you can do at the moment, and of what it means to know a language.*
The above points are elaborated on in the publications on DIALANG such as Alderson and Huhta (2005) who write that

**DIALANG is the first large-scale language assessment system that aims at diagnosing rather than certifying language proficiency.** (p. 302)

**DIALANG aims to be a tool that supports independent, life-long language learning by providing the users with a wide range of feedback that helps them to diagnose strengths and weaknesses in their proficiency, to plan further language learning and to become more aware of their language skills.** (p. 319-320)

Alderson (2005, 30) further elaborates what diagnosis means in the context of DIALANG. According to him, DIALANG can be argued to be diagnostic in a number of ways and at different levels. At the macro level, the diagnosis is about finding out the learners’ CEFR level, which “can be used, for example, to help learners of their teachers decide whether they are ready to take a public or commercial examination at that level, or to take an exam at a lower level.” Alderson adds that such general information can also be used for placement purposes.

At a more micro level the diagnosis can be about subskills (Alderson 2005, 30) and individual items. The learners may find out that they are good at a certain subskills (e.g. understanding the main ideas) but not at others (e.g. making inferences), which information may be used by them and their teachers to guide further learning and teaching.

A further dimension of diagnosis according to Alderson (2005, 31) relates to raising learners’ self-awareness. DIALANG, and especially its self-assessment related feedback and information is “… intended to raise learners’ awareness about the nature of language and language learning, and thus to help them diagnose for themselves what their strengths and weaknesses might be in the light of such feedback”. The learners are encouraged, through the provision of verbal feedback, “to reflect upon their performance, how it compares with the CEFR and their own beliefs and expectations about their ability” (op cit. p. 31). Feedback also aims “to encourage learners to think about what might be an appropriate action if they wish to improve their language proficiency. In short, the DIALANG system aims to encourage self-diagnosis and self-help.” (op cit. p. 30).
To sum up, the main purpose of DIALANG is to diagnose language learners’ proficiency. It aims to do that in several different but complementary ways and at different levels of specificity. Key points in this include:

- Finding out about one’s **level** in the skill and language tested;
- Finding out about one’s **profile**, i.e., strengths and weaknesses (at the level of skills and sub-skills, at least);
- Increasing one’s **awareness** of
  - one’s own proficiency
  - language proficiency in general
  - the CEFR and its levels
  - the match or mismatch between one’s tested proficiency and one’s expectations and beliefs about it (i.e., self-assessment);
- Taking more **control** of one’s own learning: with the help of the test results and awareness-raising information, it is argued, learners are better able to plan their language studies and to decide on whether they are ready for a formal language examination;
- Also the **teachers** will be able to use the above information to plan their teaching of particular learners and/or to make judgments about learners’ readiness for a particular level of course or an examination;
- An extension of the last purpose is that **institutions** where the teachers work use the information about learners’ proficiency to decide on learners’ readiness to participate in different types and levels of language courses. The use of DIALANG as a **placement** test in schools and other institutions was not one of the explicit goals of the system but it is in line with the envisaged use of the test results by language teachers for decisions concerning teaching. This purpose has in fact become very important as most users of the system probably take DIALANG tests in institutions that want to place them on language courses.

According to Millman and Greene’s (1993, 336) classification of test purposes, the **domain of inference** that pertains to most of the above-mentioned purposes is the **criterion-based** one (see Table 1, the last column on the right), as these purposes relate to knowledge, skills or behavior that are required for success in specific settings – all these are specified in the CEFR and the DIALANG Assessment Framework and Specifications. Some purposes of DIALANG – those that are related to learners’ awareness and control of learning – appear to belong to Millman and Greene’s second, **cognitive**, domain of inference. The aims of DIALANG do not appear to be directly relevant to their third domain, **curriculum**, because DIALANG is not related to any particular course or curriculum. However, we might argue that perhaps placing students on courses by teachers and institutions is a purpose that is at least indirectly related to the curricula of the institutions that made the decision to use DIALANG for placement purposes.
Millman and Greene (1993) also define three **types of inference** that different purposes may focus on: describing individuals, describing a system or a group, and making mastery decisions (see Table 1). The majority of the above mentioned purposes of DIALANG concern **describing individuals** (e.g., their level of proficiency in different skills, or performance on individual items). Also placing students on courses probably relates to that same type of inference. Thus, the envisaged uses of DIALANG do not concern the other two types of inferences in the Millman and Greene framework describing groups or systems, or making mastery decisions about individuals. However, as we will see later, some researchers have used DIALANG to study the proficiency of groups of language learners and have thus extended the scope of envisaged purposes of the system.

**Table 1.** Functional description of achievement and ability test purposes by Millman & Greene (1993, 336)

<table>
<thead>
<tr>
<th>Type of inference desired</th>
<th>Domain to which inferences will be made</th>
<th>Curricular Domain</th>
<th>Cognitive Domain</th>
<th>Future Criterion Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Before instruction</td>
<td>During instruction</td>
<td>After instruction</td>
</tr>
<tr>
<td>Description of individual examinees’ attainments</td>
<td>Placement</td>
<td>Diagnosis</td>
<td>Grading</td>
<td>Reporting (e.g., to parents)</td>
</tr>
<tr>
<td>Mastery decision (above/below a cutoff for individual examinees)</td>
<td>Selection</td>
<td>Instructional guidance</td>
<td>Promotion</td>
<td>Certification Licensing</td>
</tr>
<tr>
<td>Description of performance for a group of system</td>
<td>Pre-instruction status for research / evaluation</td>
<td>Process and curriculum evaluation</td>
<td>Post-instruction status for research / evaluation Reporting (e.g., on effectiveness of instructional programme)</td>
<td>Construct measurement for research / evaluation Reporting (e.g., on educational achievement at state or national level)</td>
</tr>
</tbody>
</table>

Wolfe and Smith (2007) further elaborate on the types of purposes that a test can have by referring to the kinds of **decisions** that are made on the basis of the test results. They state
that in most cases tests depict status for the sake of describing individuals or groups, which is true also for most of the purposes of DIALANG listed above. Other types of decision concern selecting individuals who are at a particular level of performance (cf. placement of students on courses) or evaluating claims based on a theory. This last type of decision is not covered by any of the purposes envisaged for DIALANG.

2.2 Summary of the Bachman and Palmer framework of test usefulness

According to Bachman and Palmer (1996), several factors contribute to the usefulness of a test for its intended purpose(s): reliability, construct validity, authenticity, interactivity, impact, and practicality. In principle, the better the quality of each of these aspects is, the more useful the test is. However, there is usually a tension between the different factors since maximising the quality of one is often detrimental to the quality of another feature. Thus, it is essential to consider the purpose of the assessment to be able to decide if all the aspects of quality are equally important for any given purpose. In fact, the overall usefulness of the assessment is what really counts, and that cannot be evaluated without taking into account, and weighing all the factors above that contribute to usefulness (op cit. p. 18).

The first aspect of test usefulness in the Bachman and Palmer framework is reliability, which has been one of the key qualities of measurement as long as testing and assessment have existed as a profession and an area of study. Reliability of a test concerns the extent to which it measures whatever it is designed to measure in a systematic and dependable fashion.

The next three aspects of test usefulness in the Bachman and Palmer framework – construct validity, authenticity, and interactivity – are quite closely connected (this will be discussed in more detail in Section 3.3; see also Figure 1 in that section). Construct validity concerns the interpretations made on the basis of a test result: Are the interpretations justifiable and appropriate? A construct is a specific term for the ability that a test intends to measure, for example, foreign language ability.

Authenticity relates to the relationship between test takers’ performance on the test tasks and tasks in the ‘real world’: What is the domain of generalization to which one wants to generalize the interpretations made from the test scores? The key consideration in
determining the authenticity of a test is the correspondence between the test tasks and the tasks in the target language use domain (domain of generalization).

**Interactiveness** can be defined with reference to both construct validity and authenticity. Interactiveness is about the extent to which the test tasks engage the test taker’s areas of language ability and other relevant characteristics. The key point is thus the interaction between the test taker and the task.

The **impact** of test use operates at two levels: micro (individuals) and macro (institutions, society). The impact of assessment is often discussed in terms of how positive or negative it is, and the designers and administrators of assessments generally hope that the impact will be positive or that there will be no unintended consequences, especially negative ones.

**Practicality** differs from the other test qualities as it relates to availability of resources and to decisions to use or develop a particular test – or not. A test is practical if the resources that it requires (time, money, work, personnel) do not exceed the available resources; a test that is too long or expensive for its purpose is impractical and is not likely to be used. It is probably the case that the most common tension that occurs between the different aspects of usefulness takes place between practicality and the other aspects.

It is obvious from the brief description above that the factors contributing to test usefulness in the Bachman and Palmer model are in many ways similar to the aspects of validity in the other validity frameworks discussed earlier. In the presentation of the validity arguments and evidence for DIALANG in the chapters that will follow, occasional reference will be made to how different aspects of the Bachman and Palmer framework relate to the Messick framework, in particular.
3. ANALYSIS OF THE ARGUMENTS AND EVIDENCE FOR THE USEFULNESS OF DIALANG

In this main chapter of the paper, a range of arguments and evidence for the validity of the DIALANG assessment system are presented. As described above, the chapter is organized along the aspects of test usefulness proposed in the Bachman and Palmer (1996) framework. Additionally, each aspect is viewed from two complementary viewpoints: theoretical / logical (a priori) arguments, on the one hand, and empirical (a posteriori) evidence, on the other hand, which is an approach to validation advocated by Weir (1991, 1993) in particular, but which is not incompatible with how both Messick and Bachman and Palmer view validation.

Table 2 presents a summary of this chapter by describing, for each aspect of test usefulness, the theoretical-logical arguments and empirical evidence for the quality of that aspect, as well as the documents where the arguments and evidence are presented. The table is organised according to the aspects of usefulness (rows) and the type of evidence or arguments that are provided for each aspect (columns). The logical arguments comprise the a priori stage of validation and the empirical evidence the a posteriori stage. The articles and topics that relate to the present author’s doctoral thesis are in bold typeface.
Table 2. Summary of the theoretical / logical arguments and empirical evidence for the usefulness of DIALANG

<table>
<thead>
<tr>
<th>Aspect of usefulness</th>
<th>Theoretical / logical arguments (a priori validation)</th>
<th>Empirical evidence (a posteriori validation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td>All documents and procedures that aimed to ensure consistency of the assessment system: (1) The detailed item specifications that guided the operationalization of the assessment framework and aimed at ensuring high quality of items. (2) Guidelines for ensuring uniform review of the test items, the accuracy and comparability of the translations used in the system and the comparability of standard setting sessions. (3) Specifications of the computerized system and programming that aimed to guarantee reliable service. (4) Scoring keys to ensure consistent and adequate automatic scoring of short-answer items.</td>
<td>Evidence about - reliability and validity of the DIALANG scales (Kaftandjieva &amp; Takala 2002) - reliability of the English pilot tests (Alderson &amp; Huhta 2005; Alderson 2005) - reliability of standard setting (Alderson &amp; Huhta 2005; Alderson 2005) - reliability and validity of the Vocabulary Size Placement Test (Huhta 2007a) - reliability of scoring open-ended items (the present paper; Appendix 1) - reliability of service (technical reliability of the computerised system) (the present paper; Appendix 1)</td>
</tr>
<tr>
<td>Construct validity</td>
<td>(1) The definitions of language and use of language in the CEFR. (2) The overall assessment plan (framework) of DIALANG that defined how DIALANG draws on the CEFR. (3) The operationalisation of the assessment framework as defined in the test specifications and guidelines for item review and translations. (4) The planning, development and implementation of self-assessment and feedback into the system. (5) The development, trialling and calibration of test and self-assessment items. (6) The development of linking procedures to the CEFR (standard setting). (7) The definition and implementation of diagnosis of language proficiency in the system.</td>
<td>All the evidence about the reliability of various aspects of the system (see above). Concurrent validity evidence (1) against self-assessments and some external tests (Finnish Matriculation examination, school grades) (the present paper) and (2) against Matriculation exam and school grades (Jaatinen 2005) Concurrent validity evidence of DIALANG self-assessments against language tests (Desroches et al. 2005; Demeuse et al. 2006). Evidence about the internal structure and relationships between DIALANG tests (the present paper; Peschel et al. 2006). Evidence about users’ reactions to the system in general and to its</td>
</tr>
<tr>
<td>Authenticity</td>
<td>The way in which DIALANG drew on the CEFR content categories and definitions of what it means to know and use a language.</td>
<td>Users’ reactions to and views on different parts and functions of the system, including test content, feedback and self-assessment (Huhta 2007b; the present paper)</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Documents:</td>
<td>see above (mainly the CEFR, DAF, DAS, and the item review guidelines)</td>
<td>Some of users’ comments on their test taking experience, particularly on the VSPT. Other comments on the system, tasks, self-assessment or feedback may throw light on the interactiveness of DIALANG (Huhta 2007a; Huhta 2007b; the present paper)</td>
</tr>
<tr>
<td>Interactiveness</td>
<td>The way in which DIALANG assessment system in general and the assessment specifications in particular aimed at designing of tasks and activities that engage the users’ language knowledge, strategic competence, metacognitive strategies, affective schemata, and topical knowledge.</td>
<td>Impact on individual learners studied so far (Huhta 2007a; 2007b; the present paper) Impact on the institutions using DIALANG for placement and</td>
</tr>
<tr>
<td>Impact</td>
<td>The plans to create an assessment system that would achieve the different stated (mainly individual) purposes of DIALANG.</td>
<td>Impact on individual learners studied so far (Huhta 2007a; 2007b; the present paper) Impact on the institutions using DIALANG for placement and</td>
</tr>
</tbody>
</table>
### Practicability

The different design features of DIALANG that aim it to be as easy, accessible, practical, quick, and inexpensive to use as possible: multilingual tests, instructions and feedback, benefits of computerisation (e.g. immediate results) and the use of the Internet (informative website, downloadable from Internet, upgrading and test-taking via Internet), free of charge.

Documents:

- Figueras et al. 2000
- Fligelstone & Treweek 2000
- Fligelstone 2001
- Huhta et al. 2002
- Alderson 2005
- Alderson & Huhta 2005
- Huhta & Figueras 2004

### Users’ reactions to the practicality of DIALANG, such as ease of use, clarity of navigation and instructions, technical matters (Huhta 2007b; the present paper; Floropoulou 2002a)

Practicality of DIALANG for other purposes than originally envisaged (West 2003; Haahr et. al 2004)

### 3.1 Qualifications of the DIALANG development team

The level of experience and expertise of the test development team is not discussed in the literature on test validation because the publications on validation focus on the areas of work and procedures that are typically required to produce valid and useful assessments, not the people who do the job (except raters and item writers, whose qualification are sometimes referred to in the literature). However, it goes without saying that designing a multilingual, large-scale assessment system is an activity that requires a certain amount of expertise in
order to provide its users with meaningful results and feedback. Certainly, DIALANG advertises itself on its website in the following way:

*DIALANG offers carefully designed and validated tests of different language skills, together with a range of feedback and expert advice on how to improve your skills.*

*DIALANG also offers scientifically validated self-assessment activities, because it is now widely recognized that being able to judge your own language proficiency is an important part of learning a language.* …

*DIALANG has been developed by more than 20 major European institutions, with the backing of the European Commission.*

The main reason for including the above statement was to make it clear to potential users of the system that DIALANG was qualitatively different from most other language tests that one can encounter on the Internet – that it was better and more dependable than most of the other tests. That was the ‘sales pitch’ targeted at the lay audience, but the inclusion of a list of institutions involved in the creation of the systems, which is also found on the DIALANG website, was intended to convince language education and testing professionals that DIALANG was a serious, professionally developed language assessment system. (The conference presentations and academic journal articles by Project members also served the same purpose.)

A possible way to link the level of expertise of the test developers to validity theory is via Weir’s (1991, 1993, 2005) discussion of the *a priori* and *a posteriori* validation. Developers’ expertise – or lack of it – could perhaps be seen as part of *a priori* validation of a test. The selection of the persons who are responsible for the testing process takes place before the actual use of the tests in the same way as the design of test blueprints and other documents listed in the column ‘Theoretical / logical arguments’ in Table 2, and these selections are as essential to the quality of the test as those documents.

As far as the expertise of the DIALANG development team is concerned, it can be argued to be quite considerable, which gives credibility to the *a priori* claims that DIALANG is a useful system of high quality. The key arguments for the expertise of the development team include at least the following:
The partnership included institutions with a proven track record in doing high-quality research on language testing (mainly Lancaster University and Cito; University of Jyväskylä was beginning to get attention in this respect at that time), in developing and maintaining large-scale testing systems (Cito and U. of Jyväskylä), in participating in large-scale international comparative studies in education (U. of Jyväskylä and Cito) and in psychometrics (Cito).

The partnership included a number of individuals, in different phases of the project, both from the above-mentioned institutions or from other partner institutions, who were world-class experts in their areas and/or who had extensive experience in coordinating European networks and large-scale projects and in working with the Council of Europe and European Commission such as J. Charles Alderson (Alderson & Urquhart 1984; Alderson & Windeatt, 1991; Alderson, Clapham, & Wall, 1995; Alderson, 2000), Neus Figueras (2001; Huhta & Figueras 2004; Alderson et al. 2004), Felianka Kaftandjieva (Takala & Kaftandjieva 2000); (Kaftandjieva, 2004), Sari Luoma (2001, 2004), Wolfgang Mackiewicz (Mackiewicz, 1995), Mats Oscarson (Oscarson, 1978, 1984); (Oscarson, 1997)), Kari Sajavaara (1981, 1992); (Takala & Sajavaara, 2003), Sauli Takala (Purves & Takala, 1982; Takala & Kaftandjieva, 2000), Mirja Tarnanen (2002; Luoma & Tarnanen 2003), Alex Teasdale (Teasdale & Leung 2000), and Norman Verhelst (Mislevy & Verhelst, 1990; Verhelst & Glas, 1993). Brian North, the key person behind the scales found in the CEFR participated in the work of the self-assessment and feedback working group of the project and reviewed the self-assessment instruments developed by the project. A number of other competent individuals contributed to the project (see Alderson 2005, vi – xii for an exhaustive list of people involved in DIALANG), some of whom were already established figures in language testing and language education more generally.

The partnership’s expertise in item writing varied across languages. In some cases, the item writers were experienced professionals who had worked for established language examinations for years but in other cases they were less experienced as item writers although they were professionals in language teaching, for example. Thus, the expectations of the quality of the items varied
somewhat depending on the language. In any case, however, the item development and review followed specified procedures that aimed at ensuring that all the languages would achieve the same minimum quality in terms of items produced. (Piloting was intended to be the final check on how well the work had succeeded but it could unfortunately be completed only for some of the languages.)

- The partnership’s expertise in creating the feedback system for DIALANG was perhaps the least developed of the content areas. This was largely due to the general lack of development of diagnosis of foreign language learning and proficiency (see Alderson 2005 for detailed discussion). If you cannot properly diagnose FL proficiency, you cannot give really good and useful feedback about it – or you can try, as we did in DIALANG, and devise a range of different types of feedback in the hope that different learners might benefit from some of the feedback. Thus, research into the usefulness of the feedback from DIALANG is particularly interesting as the developers faced a much tougher challenge there because of the lack of established predecessors and procedures – unlike in most other areas of work in DIALANG, with perhaps the exception of linking the test results with the CEFR, which was another novelty of the system.

- The partnership did not have expertise in programming but it had to hire outsiders to do this work. It is difficult to evaluate the expertise of the programmers who worked for the project, except by referring to the fact that they managed to create a system that has been running successfully, with minimal maintenance, since the end of the project in 2004 (see Alderson 2005, 38, on the difficulties hiring programmers for the project). The partnership also lacked expertise in commercialising DIALANG, until about 2005. This did not directly impact on the development of DIALANG but it made it extremely difficult to find a solution for the permanent maintenance and further development of the system. At the time of writing this paper, the future of DIALANG is still quite open, and it is not clear how long the core partners can maintain the system in the future.

On the whole, it can be argued that the DIALANG partnership possessed considerable expertise that should have enabled it to develop an assessment system which serves the
purposes it was designed for and which is thus useful for its users. However, as with the *a priori* validation in general, an expert design team does not in itself guarantee that the test to be developed is valid and useful – no more than good test specifications guarantee the quality of the tests. We need empirical evidence in order to evaluate to what extent we can argue DIALANG to be a useful system for the purposes it is used.

### 3.2 Reliability of DIALANG

Reliability of a test concerns the extent to which it measures whatever it is designed to measure in a systematic, consistent and dependable fashion. A person taking the test on two occasions should receive the same score or grade, assuming he/she has not made significant progress in the skills and knowledge measured by the test between the two occasions.

As with any other aspect of usefulness in the Bachman and Palmer framework, reliability can be analysed from two complementary perspectives: theoretical / logical and empirical. The theoretical / logical perspective covers all the actions that take place before the test is actually used (*a priori*): it is about the design of the test (e.g. Weir 2005). All activities that potentially lead to the system measuring the intended skills is a systematic, unbiased way can be thought to contribute to the reliability of the tests. At this stage, it is often difficult to distinguish between reliability and validity: most of design features of a testing system contribute to both its validity and reliability. Indeed, Messick (1995) regards reliability as one aspect of generalizability of meaning of test scores, which is one aspect of construct validity (i.e., evidential basis for test interpretations).

#### 3.2.1 Theoretical / logical attempts to achieve reliability

At the heart of *a priori* attempts to ensure the reliability of a test are test specifications that define how the areas of knowledge and skills are to be operationalized as concrete test items and that determine how different item types are to be designed in such a way that they function consistently and in the way they were planned to function. Failure of the specifications to do this in a clear and detailed way is likely to result in unreliable tests. The complete lack of test specifications – which is unfortunately the case with many real-life tests
– is an extremely serious threat to the reliability or, indeed, to any other aspect of quality of a test.

The development of DIALANG was guided by test specifications. In addition, the process of item review and construction of tests was supported by other documents that can be considered more detailed extensions of the test specifications. These included the item review guidelines that defined in a more precise way how the items selected for piloting were to be checked in the specific technological context of DIALANG (Huhta & Noijons, 2002). They also specified how translations of the test instructions and the programme interface were to be done, reviewed, and implemented in the system (Fligelstone & Huhta 2001; Huhta, 2002). Thus, they contributed to the creation of as uniform an experience to the users of the system as possible – which is essential from the point of view of reliability of the tests and of the system as a whole. Furthermore, guidelines existed for the systematic conduct of the standard setting exercises that provided data for determining the cut-off points on the test score scales for the six CEFR levels (Kaftandjieva et al. 1999; Noijons et al., 2002).

A small point in the context of the whole system but something that had a significant impact on the reliability of the test results was the creation of good scoring keys for free-response items. A ‘dumb’ scoring system such as the one used in DIALANG is dependent on the test designers’ ability to create an exhaustive list of acceptable answers for the items in which the test takers have to write, rather than select, their responses. While some of the scoring decisions have direct bearing on the validity of the test (e.g., Do you penalise for spelling errors in open-ended reading comprehension items?), the main challenge relates to reliability: Can you anticipate all valid answers so that you can create a comprehensive, reliable scoring system?

As a computer system operating via the Internet, DIALANG faces other kind of challenges in its attempts to provide a reliable testing system to the learners compared with a paper-and-pencil test. There are issues with the dependability of the system across different hardware and operating systems, and networks. The reliability of the Internet service in general poses potential problems for the systematic functioning of the system that are not directly related to the reliability of the test system as such – problems that the system designers cannot always control for. The careful planning of the overall system (e.g. use of mirror servers that can handle break-ups in service) and the different software components,
and the choice of programming language, etc. are aspects of design that have direct effect on the reliability of IT side of the testing system (cf. Figueras, Taalas, & Treweek, 2000).

It should be added that, in the case of DIALANG, the procedures for ensuring high reliability in the test design phase were not limited to providing the relevant actors with documents such as item writing and review guidelines. In addition, the whole process of training item writers and face-to-face and on-line consultation and advice were important elements of attempts to ensure, for example, uniform and systematic item writing and review (Alderson 2005).

3.2.2 Empirical evidence for the reliability of DIALANG

No comprehensive study has been carried out on the reliability – nor the validity – of DIALANG. It would in fact be impossible to provide only one picture of the quality of DIALANG because of the different state of readiness of the languages in the system: While the tests in some languages have been properly piloted and the results analysed other languages remain non-piloted, which means that the reliability and validity of the 14 different language test systems of DIALANG are bound to differ. However, we have some evidence about the reliability of certain parts of the system either from its developmental stages or from the actual operational system.

There is evidence of a fair level of internal consistency of the English pilot tests (see Alderson & Huhta 2005; Alderson 2005), which suggests that the operational English tests in DIALANG Version 1 are quite reliable. However, there is no published research into the reliability of the other piloted languages (Dutch, Finnish, French, German, Italian, and Spanish) although the relevant pilot data are exist. The study of the reliability of the tests in the non-piloted languages (e.g. Danish, Irish, Norwegian, Swedish) would be more difficult because the number of pilot test takers in those languages was so small in the lifetime of the Project that the analysis and calibration of the test items could not be carried out. Although standardised item design and review procedures across all DIALANG languages suggest that the results obtained for English might tell us something about the reliability of the other tests, we cannot obviously be sure.

Another aspect of the development of the system for which we have reliability evidence is standard setting, which is a procedure by which test developers attempt to set meaningful
cut-off points on the test score scale. For example, the cut-off point for level B1 in the intermediate test of reading in English might be 45 (on 0 – 100) scale but the same cut-off score for reading in Spanish might be 56. The reason for the differences in the cut-offs is that each test is an individual with a somewhat different combination of items of different difficulties, and thus, they cannot have the same cut-off points. The cut-offs were determined by a combination of information about the empirical item difficulty (if it was a piloted language) and judgements about the level of the items made by experts in the standard setting sessions, guided by standardised procedures developed in the Project (Kaftandjieva et al., 1999; Noijons et al., 2002). For details on standard setting, see Alderson (2005) and Alderson and Huhta (2005). As in any other judgemental task, the judges of item difficulty can do their job more or less systematically, both with themselves and with the other judges. Alderson and Huhta (2005, 317) provide evidence for a very high intra and inter-judge agreement for the German standard setters; most of the correlations were in the range of .8 or .9. Obviously, the standard setting teams in different languages varied to some extent as far as their consistency is concerned, and the degree of their consistency has had an effect on the accuracy of the cut-off points and, thus, on the reliability, and validity, of the test results in different tests and languages.

One of the articles comprising the present thesis (Huhta 2007a) in an exception to the lack of reliability studies on the operational DIALANG system as it focuses on the reliability and validity of the Vocabulary Size Placement Test (VSPT). Any test score has some degree of imprecision and measurement error but it appeared that the current version of the VSPT had some serious problems. The study into the VSPT was prompted by feedback from many users of DIALANG who claimed that the VSPT worked quite strangely for them. Some users appeared to receive scores that were totally out of line with what they could expect to get. The problem seemed to be that these test takers received too low scores from the VSPT – even zero points – which was incompatible with what the test takers themselves and their teachers knew about their level of proficiency, or even with the results of the main DIALANG skills tests. Furthermore, other users appeared to receive very different VSPT scores when they re-took the test and applied a somewhat different test-taking strategy.

Huhta (2007a) argues that there are at least two reliability issues with the current VSPT. The first one relates to inaccurate VSPT scores themselves. A wrong VSPT score means that the description of the meaning of the score that the learner can read after the test does not match
what he/she can do in the tested language. The way in which the VSPT scoring algorithm works almost certainly guarantees that the wrong score is practically always too low rather than too high. Any test score that clearly underestimates the test taker’s proficiency will almost certainly cause a negative reaction of some sort, either towards the test (‘what a bad test’ type of reaction) or towards the learner him/herself (‘am I really this bad?’ type of reaction). Huhta (2007a) provides evidence that low VSPT scores indeed provoked this type of reactions from the users. Such reactions to zero scores on the VSPT were also reported by Ylönen (2002) who used DIALANG with her learners of Finnish as L2 in Germany and who regarded such results as totally incorrect for the particular students on the basis of what she knew of their proficiency. The fact that it is possible to a zero score from the VSPT clearly reinforces such negative reactions – it is after all somewhat unusual to get a zero as a result of any test.

It was thus clear on the basis of the author’s questionnaire survey of DIALANG users, which is reported in part in Huhta (2007a), that the VPST scores caused quite negative reactions in some users. One can of course ask if it really matters if test takers do not like a particular test or the results they get from it. Learners’ perceptions of the test may not, however, be related to the way they perform on it. If it is important to do well in the test, most test takers will probably try their best no matter how stupid, boring, uninteresting, or unfair the test appears to them. However, most testing experts would nowadays agree that in normal circumstances it is not fair and ethical to ignore users’ perceptions of the validity and fairness of the tests that they are required to take. For DIALANG, which is often taken voluntarily by its users, it is obviously even more important to avoid unnecessarily antagonising its users.

In the case of DIALANG, the consequences of wrong VSPT scores are not limited to angering or depressing some of the users but inaccurate VSPT scores can make their main test results for reading, listening, writing, vocabulary, and structures unreliable. Because the VSPT is used to pre-estimate the learner’s general level of language proficiency, it plays a key part in the decision by the system to allocate to the learner the most appropriate level of test. Since the wrong VSPT result is practically always an underestimation of the learner, he/she will be assigned to a test that is too easy. It is difficult to say if too easy a test affects actual test performances but the main danger in terms of reliability of DIALANG test results lies in the fact that its test versions are optimised to measure a certain range of proficiency. Thus, an easy test version mainly consists of easy items although it also includes a fair
number of intermediate (especially B1) items, but only a few if any really hard items. An advanced learner whose level in the skill tested is B2, C1, or C2, but who was assigned to an easy test because of a poor VSPT score, is the most likely candidate to receive an inaccurate result in reading, listening, or whatever the skill tested. The easy test versions are simply not suitable for the testing of very advanced learners, even if it may be possible for the test taker to get a very high level from an easy test version if he/she obtains a perfect, or near perfect, score in the skills test. The possibility of getting a very high result on an easy DIALANG test varies from language and skill to the next depending on the availability of items for each CEFR level in the particular test and on where exactly the cut-off points were placed in different tests. In some cases, full points on an easy test of, say, reading, result in the level C1 assigned to the user while in other languages and tests a C2 is given.

Feedback received from users and the testing of the system by the project members towards the end of the DIALANG Project all pointed to the same direction: If the test taker made a lot of guesses, the VSPT score could easily go down to zero. Consequently, institutions using DIALANG were instructed by the project members to advise their students to avoid guessing when taking the VSPT but to claim a word to be a real word only if they were sure about it. Most of the informants in the questionnaire survey, the results of which are reported in Huhta 2007a, 2007b and in this paper, were also given this advice.

The detailed study on the VSPT (Huhta 2007a) examined in more detail the mechanism that resulted in the underestimation of some learners’ vocabulary knowledge by the current version of the VSPT, and it also tried to estimate the proportion of test takers thus affected. Systematic trials with different patterns of responding to the VSPT confirmed that the main source of unreliable VSPT scores is the current scoring algorithm which is over-sensitive to guessing: Only a few false claims about words being real words when, in fact, they are non-words, result in a zero score (see Huhta 2007, p. 52). Thus, guessing, which is a very sensible test-taking strategy for any other selected response test that the learners are familiar with, does not work at all with the VSPT. It may be that the VSPT instructions contribute to the problem, as the wording deviates from the instructions in previous yes/no vocabulary tests by asking the learners to identify which words are real and which are not, rather than asking them to say which words they know vs. don’t know. The number of DIALANG test takers affected by unreliable VSPT scores is difficult to estimate precisely but a reanalysis of the
English pilot test data suggested that it may be close to 10% of the test takers (op cit. page 51).

The other reliability issues of DIALANG about which we have at least some empirical evidence include the technical reliability of the IT system that underlies DIALANG assessment procedures and the reliability of the scoring of open-ended questions found in its language tests.

The technical reliability of service must be a top priority for any IT system that operates over the Internet. As was briefly described above, and elaborated in Alderson (2005), the DIALANG Project spent considerable time and resources throughout the project to ensure that the system would work reliably across different hardware and operating systems. Again, it is impossible to give precise figures, but based on the statistics on DIALANG tests taken and on the present author’s experience in advising users of the system (see Appendix 2 for details on the present author’s role in the DIALANG Project), it seems safe to say that the great majority of all DIALANG tests have been successfully completed, at least from the technical point of view. There are institutions that hardly ever face technical problems when hundreds or even thousands of their students take DIALANG tests.

However, not all learners are free from technical problems when using DIALANG. There are individual users and institutions that probably never managed to make DIALANG work due to some technical problem. Often the most insurmountable problems seemed to relate to firewalls that control and safeguard communication between a local system and the Internet (e.g., it turned out extremely difficult to make DIALANG work at the European Commission, the main funding body of the Project, because of the high security of their IT system). Certain firewall configurations appeared make the use of a system such as DIALANG impossible. Then there are obviously users and sites that encounter sporadic breakdowns of service or some lesser technical problems. In addition, it should be mentioned that the whole DIALANG system was down because of hardware failures for 1-3 month periods on at least two occasions since the end of the Project in 2004; the length of the two periods of closure was due to the minimal maintenance resources currently available for the system.

That technical issues are probably the most noticeable and serious of all types of problems that a computer-based assessment system can encounter is also indicated by the fact that the
great majority of all problems reported to the Project and to the present author by the test sites and individuals during the Project concerned technical issues. Feedback from sites around Europe indicated that technical problems did occur in certain places and at certain times but it is impossible to say how DIALANG compares with other complex Internet-based systems in terms of technical reliability. When technical problems occur in DIALANG they are obviously very annoying as they may totally disrupt a learner’s testing session. The analysis of the users’ responses to the questionnaire survey by the present author showed that different types of technical problems were the most often mentioned problem together with the problems with the VSPT (see Table 2 in Appendix 1 for details, and Section 3.3.2.3 below). It is thus clear that DIALANG users sometimes face technical problems that range from serious breakdowns of service to smaller hick-ups. It is impossible to estimate the proportion of test sessions affected in this way but they seem to be frequent enough so that they feature on the top of the list of problems mentioned by the learners surveyed.

Besides the present author, also Floropoulou (2002a) studied the technical aspects of DIALANG (navigation, interface, and a range of other matters, some of which were clearly technical while others were more related to content). She observed and interviewed six informants who took DIALANG tests at Lancaster University and found quite a few technical problems with the system that ranged from serious (two of the test takers experienced a crash of the system) to minor problems. What is clear from her study, although she did not specifically mention it, is that the average user is likely to take some time to learn to use DIALANG; the first-time user is very likely to encounter points in the test taking and navigation process where he or she is not quite sure what to do and may have to resort to a trial-and-error approach in order to figure out how to proceed. However, this is probably true for almost any computer programme, as very few are so intuitive that the user immediately understands what to do, especially when the typical session with the programme consists of quite a few rather different stages and steps such as in DIALANG. Neither Floropoulou’s nor the present author’s study informs us about how soon a typical user gets acquainted with navigating through the different stages of DIALANG but it is probably safe to assume that after taking a couple of tests the user will have a fairly good grasp of the system.

Floropoulou’s detailed study revealed for example the following problems or issues:

- Some users experienced crashes or slow functioning;
In the beginning, the users were somewhat confused about what to do; in particular, they took time figuring out how to move forward in the programme;
- The fact that the forward and skip-to-next-section buttons were next to each other caused problems (there is a warning after you click on the skip button so that you can return to the previous point, however, if you accidentally pressed on the skip);
- The immediate feedback pop-up screen caused problems for navigation;
- The existence of the VSPT was puzzling at first: some users wondered why they were given a vocabulary test when they had just selected to take a listening test, for example;
- Some users did not realize that they had to mark also the non-words in the VSPT by answering ‘no’ for them;
- The popping up of the additional character box in open-ended items confused some users;
- Adjusting of the volume in listening items was not clear to everybody;

Some points were also mentioned by Floropoulou’s informants that are not technical problems but rather aspects of the system or choices made by the designers that some users find surprising or objectionable. These included:

- The user cannot go back in the test and review and change their answers;
- The user can only listen to the audio recording once, rather than twice, in most listening items;
- The user has to answer yes / no to the self-assessment statements (there are no intermediate categories such as ‘sometimes’);
- The user has to answer yes / no to the VSPT items (there is no ‘I don’t know’ option);
- The test does not have a time limit;
- Items are not categorized in terms of rubrics or content but follow each other in apparently a random fashion;
- The exact number of remaining items is not indicated (although there is a blue bar that shows roughly how far you are in the test at any given moment; see below).

The user also liked many technical and content aspects of DIALANG, including:

- The font and the size of the letters;
- The visually clear icons;
- Colours were considered warm;
- The chance for many users to read the rubrics in their native language was considered very user-friendly.
- Feedback on the match between self-assessment and test performance was considered useful
- Useful instructions;
The pull-down menus, the radio buttons for selection and the ‘combo box’ were considered highly innovative;

- The blue bar at the top of the screen during the language test showing progress in the test was liked by the informants.

Floropoulou’s (2002a) study illustrates very well how even a very small-scale but detailed observation and interview-based investigation can yield considerable amount of useful information about the technical aspects, in particular, of a computer-based language test. While some of the things considered problems by the informants may not be so regarded by the designers of the system, it is clear that Floropoulou’s study found out about many of the technical problems that other users of DIALANG sent us between 2001 and 2004. Her study thus confirms the argument by Nielsen (2004), quoted in Huhta, submitted) that to find out about the most serious technical navigation and interface problems in an Internet-based system it is typically enough to study only about five users.

Finally, the questionnaire survey by the present author suggests that the scoring of the open-ended questions was sometimes unreliable for at least some users (see Appendix 1). The study suggests that this may not be a widespread problem, however, as only about three percent of the respondents to the survey of users mentioned this as a problem with DIALANG (see Table 2 in Appendix 1). However, also the informants in Floropoulou’s (2002a) small-scale study experienced this problem, so we cannot be sure what proportion of users actually suffer from it.

### 3.2.3 Reliability of self-assessment statements

The test – retest type of reliability is often quoted as a prototypical example when reliability is defined in measurement literature. Interestingly, retesting with the same test is hardly ever done in practice because it is highly problematic to administer the same test to the same test takers on two occasions in order just to see if the results of the two rounds of testing agree. First, it is impractical as it takes time and it is not easy to convince the test takers and their teachers (if it is a school context) of the need to be subjected to so much testing for apparently no direct benefit to the participants. Second, it is possible that the test takers remember some of the test content and they may even have learnt something from the first administration of the test, which complicates the interpretation of the results. It may be that the most typical occasions where a test is actually administered twice to the same test takers...
are studies that try to find out if learners’ have changed over time and where it has not been possible to develop truly parallel versions of the same test – in such cases the best approach may be to test the learners twice with the same instruments before and after the learning / teaching period the effects of which one is interested in measuring (e.g. Huhta & Suontausta, 1996)

Interestingly, some of the problems inherent in the test – retest reliability of formal tests may not apply to self-assessment. For example, you do not ‘learn’ from the first administration of a typical self-assessment instrument in the same way as you might from a test measuring skills or knowledge. The practicalities and motivation involved in re-administering a self-assessment instrument may also be less problematic than with skills tests. Despite these considerations, the estimation of test – retest reliability of SA instruments appears to be very rare, if not non-existent, in measurement literature. Thus, it does not come as a surprise that there have been no studies into this type of reliability of self-assessment instruments used in DIALANG.

It is easier to study the internal consistency of any measurement instrument because it can be done without administering the instrument more than once, so the calculation of indices such as Cronbach’s alphas is the most typical approach to estimating the reliability of tests. It should obviously be born in mind that this is conceptually different from comparing the results of two administrations of the same test. Internal consistency of a test may in fact be much closer to construct validity than reliability (when the latter is defined in terms of repeatability and consistency of measurement): An analysis of the dimensionality implied in internal consistency measures resembles correlational, concurrent study of the relationships between different tests of the same test battery or between the test and other tests measuring similar or different abilities than the test that is being validated.

There is one study into the internal consistency type of reliability of DIALANG self-assessment statements, and it is based on the DIALANG pilot testing data (Alderson 2005, 100-101). Alderson’s analyses show that the internal consistency of the three sets of 18 SA statements chosen for the operational version of DIALANG was quite high: the Cronbach alpha for reading was .814, for writing .874, and for listening .824 (estimated alphas for hypothetical 45-statement versions of the SA instruments were .916 – .945).
3.3 Construct validity, authenticity, and interactivity

Three aspects of test usefulness in the Bachman and Palmer framework – construct validity, authenticity, and interactivity – are quite closely connected (see Figure 1). This is why the discussion of them below is intertwined, especially when the a priori theoretical validation of DIALANG is concerned. Largely the same documents and design procedures were used to ensure that the system, when it was ready, would have the characteristics that are captured in the concepts of construct validity, authenticity and interactivity in the Bachman and Palmer framework. When it comes to analysing empirical evidence, it is somewhat easier to distinguish which type of evidence sheds light on each of the three aspects of usefulness, and the presentation of such evidence will be organised accordingly in the passages below.

An analysis of the construct validity, authenticity and interactivity of DIALANG below is not limited to the language tests found in the system. Because DIALANG is not just a set of language tests but rather a more comprehensive assessment system that makes use of modern technology, it is necessary to consider how these aspects of test usefulness relate to such parts of the system as the self-assessment tasks, the whole range of feedback that includes some that are very unusual in any testing system, and the computational and technological operations that underlie, for example, the delivery of the test tasks, capturing of users’ responses, scoring of the responses, computation of the results, and presentation of the different types of feedback.

According to Bachman and Palmer (1996, 21), construct validity is about the interpretations made on the basis of test results: To what extent are the interpretations justifiable and appropriate? (see also Figure 1 below). A construct is a specific term for the ability that a test intends to measure, for example, foreign language ability or an area of that ability. Construct validity thus means “the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure” (op cit, p. 21). Several different kinds of evidence can be used to build a case for the construct validity of a particular interpretation of test scores, including content relevance and coverage, concurrent criterion relatedness, or predictive utility (op cit., p. 22).
Authenticity in the Bachman and Palmer framework relates to the relationship between test takers’ performance on the test tasks and tasks outside the test situation: What is the domain of generalization to which one wants to generalize the interpretations made from the test scores? The key consideration in determining the authenticity of a test is the correspondence between the test tasks and the tasks in the target language use domain (domain of generalization). The degree of a test’s authenticity can be evaluated by detailed analyses of the characteristics of the tasks in the test and in the target domain. Such characteristics include the setting, input, expected response, and the relationship between input and response. One of the reasons why it is important to study authenticity of a test is that it affects the test takers’ perceptions of the test and thus potentially their performance on it (op cit., p. 24).

Figure 1. Relationships between construct validity, authenticity, and interactiveness in the Bachman & Palmer framework (1996, p. 22)

The third part of this group – interactiveness – can be defined with reference to both construct validity and authenticity. Interactiveness is about the extent to which the test tasks engage the test taker’s areas of language ability, personal characteristics, topical knowledge,
strategic competence, and affective schemata (Bachman & Palmer 1996, 22; 152). The key point is thus the interaction between the test taker and the task.

A few words are in order about the relationship between the three aspects of usefulness described above and certain aspects of construct validity in Messick’s validity framework. Interestingly, Messick’s content aspect appears to be divided between Bachman and Palmer’ construct validity, authenticity, and interactivene ss. Content is part of the construct definition in their framework and one way of providing evidence for the construct validity of a test involves examining the relevance and coverage of test content (Bachman & Palmer 1996, 22). Content is also part of authenticity because content is obviously a key element of the input and output (response) of a task – be it a test task or a ‘real life’ task. And topical knowledge (i.e. content) is also an aspect of interactivene ss, as one of the validity questions concerning interactivene ss is to what degree the task engages the test taker’s topical knowledge in the same way as the target language task would (op cit., p. 152).

Messick’s substantive aspect of construct validity comes close in meaning to Bachman and Palmer’s interactivene ss, as the former is about test takers’ responses and the processes underlying those responses, and the latter concerns the interaction between test taker’s (cognitive and affective) characteristics and the task. Messick’s structural aspect of construct validity concerns the match between the scoring criteria and rubrics used in the test and the construct domain to be tested, which is clearly a construct validity question in the Bachman and Palmer framework (op cit., p. 151).

Finally, the generalizability aspect in Messick’s model, which relates to the extent to which the interpretations drawn from the test results generalize across different groups of test takers and across different settings and tasks, is close to what Bachman and Palmer include in their authenticity and domain of generalization. We should also remember that Messick includes reliability in generalizability, as was mentioned earlier, unlike Bachman and Palmer who consider it a separate element from the other aspects of usefulness.
3.3.1 Theoretical basis of the construct validity, authenticity, and interactiveness of DIALANG

DIALANG was the first major language assessment system that was systematically built on the Common European Framework of Reference (CEFR). Only the preliminary version of the CEFR (Council of Europe, 1995, 1997) was ready when the groundwork for the assessment system was done in the late 1990s, but that first version of the CEFR already contained the essential elements of the Framework whose final version was published several years later (Council of Europe, 2001). Thus, the main theoretical basis for the construct validity of DIALANG are the definitions found in the CEFR of what is involved in using and learning a language. The CEFR consists of two dimensions: The best known is the vertical one, i.e., the six-point scale of language proficiency that ranges from a beginning proficiency to a very good command of a language. In fact, the CEFR contains a many such scales describing proficiency in a number of different, more precisely defined areas and communicative activities. The fact that most of the scales found in the CEFR are not informed speculation as most proficiency scales are; they are rather based on solid empirical research by North (North, 1995), which gives extra credibility to the theoretical arguments about the construct validity of DIALANG. The second, horizontal, dimension of the CEFR includes definitions of different conceptual categories of language use: Domains of use, communicative activities, functions, topics, texts, and competences. However, as will be described below, the Project found the CEFR alone to be insufficient as the basis for defining the construct of reading, listening, writing, structures and vocabulary, and it had to resort to a variety of other sources of information in order to construct the test specifications for DIALANG.

Some of the articles that comprise the present thesis include brief accounts of the basis and development of DIALANG, and provide theoretical and logical evidence for the construct validity, authenticity, and interactiveness of DIALANG. In particular, Huhta & Figueras (2004), Alderson & Huhta (2005), and Huhta (2008), but also the introductory parts of the other, more empirical articles give brief descriptions of the development of the system. However, the most comprehensive and detailed documents that pertain to the theoretical validation (cf. Weir, 1991, 2005) of DIALANG were the documents designed during the project to guide the development of the system. Since the articles written or co-written by the
present author cannot give a full picture of this aspect of the project, it is useful to elaborate here on the documents and procedures that underlie the development of DIALANG.

### 3.3.1.1 Diagnosis of L2 proficiency – what is it?

Before considering the theoretical arguments for the construct validity, authenticity, and interactiveness of DIALANG it is important to discuss the very notion of diagnosis of foreign / second language proficiency because it was a major challenge to the Project due to lack of attention paid to it by language testing researchers (Alderson 2005). Although diagnosis had received some attention by researchers working on the construct validation of certain large-scale language tests (e.g. Buck & Tatsuoka, 1998) or in certain areas of 'alternative' assessment (see Huhta 2008 for details), empirical research that could be relied on was scarce. DIALANG can thus be considered a pragmatic approach to tackling the problems in diagnosing foreign language proficiency in the absence of really useful information about the field. DIALANG has, however, turned out to provide important impetus to new research into diagnostic testing in the language testing community, thanks to the presentations and publications by several project members, notably by Alderson (2005), but also all the other publications on DIALANG listed in the List of References at the end of this document. There is now clearly more research and development in progress that aims to understand better language proficiency and learning, and to diagnose it more appropriately; some of this interest manifests itself in a closer co-operation between second language acquisition researchers and language testers (this will be discussed in more detail in Section 3.4.4 on the scientific impact of DIALANG).

Part of the work carried out in the DIALANG Project and afterwards by project members on diagnosis involved obtaining a **better understanding of the nature of diagnostic assessment** and its relationship with other types and functions of assessment. A key characteristic of diagnostic assessment is that it aims to support learning, either by providing the learners themselves with information that can influence what they do, or by providing their teachers with information that can help them adjust teaching in order to improve their students’ language proficiency. Thus, diagnostic assessment must have a lot in common with other types of assessment that also aim at supporting learning. Huhta (2008) is an attempt by the present author to explore and clarify the relationship between diagnostic and other,
closely related types of assessment, and to try to define some of the characteristic features of diagnostic assessment.

Alderson (2005) contrasts and compares diagnostic testing with, for example, placement testing, with which it has often been associated in the language testing literature. Huhta (2008) reports on the intertwined history and development of diagnostic and formative assessment, and concludes that the two cannot clearly be distinguished. In fact, they have often been treated as almost synonymous in the general assessment literature. It seems possible, however, to distinguish formative / diagnostic assessment from proficiency / placement / summative assessment by making reference to the level of detail that they address (Figure 1 in Huhta 2008, page 473). Diagnostic / formative assessment typically focuses on the details of the content and skills to be learned and it also provides detailed information and feedback to learners and teachers. In contrast, the proficiency / placement / summative types of assessment are generally satisfied with getting an overall picture of the learners’ proficiency in order to award marks, grades, and certificates, or to place students on appropriate courses. In addition to the degree of detail involved, a second dimension that can be used in the analysis of the purposes of assessment is to see to what extent they derive their constructs and content from either (1) theories or other such general frameworks of language, or from (2) a specific curriculum or even a course or a textbook. This distinction helps us to distinguish between theory-related proficiency tests and curriculum-related summative assessment. Formative assessment is also typically curriculum-based, but, confusingly, diagnostic assessment can be either theory-based or curriculum-based, depending on which authority on (general) assessment you happen to read.

Viewed through the lens of this kind of analytical framework (Huhta 2008, 473), DIALANG turns out to be an unusual type of diagnostic test, which cannot easily be placed in the framework and compared with other types of tests. It is clear that DIALANG is based on a theory-like framework (i.e., the CEFR), rather than on any specific curriculum. It is less clear, however, which place DIALANG should occupy along the detailed vs. general axis, i.e., the vertical axis in Figure 1 (Huhta 2008, 473). In that framework, DIALANG is placed just above the middle point on that continuum, on the side of ‘less detailed content and feedback’, because much of what it does involves the measurement of, and feedback against, the 6-level CEFR scale, which can hardly be described as a very detailed approach to diagnosis. However, DIALANG also provides more detailed feedback on the sub-skills
measured by its tests, and importantly, on each individual test item. Thus, DIALANG has features that are characteristic of the ‘more detailed’ end of the continuum, which places it apart from the typically less detailed proficiency and placement tests.

### 3.3.1.2 Language tests

A key document for the construct validity and authenticity of DIALANG language tests is the **DIALANG Assessment Framework (DAF)** that defines which CEFR content categories (domains, activities, functions, etc) are covered by the DIALANG tests (CALS & DIALANG Project, 1998a); see also Huhta et al. 2002; Alderson 2005; Alderson & Huhta 2005). In a way, the DAF is a summary of the most important categories of the CEFR. It differs from the CEFR in that some categories are excluded as not directly relevant to DIALANG (e.g., the domain of education) and in that it is supplemented by the elaborations on the various activities, functions, etc. that were found in the three earlier Council of Europe documents: the Threshold, Waystage, and Vantage level definitions (Ek & Trimm, 1991, 1999, 2000). These define in greater detail the prototypical content of the levels B1, A2, and B2, respectively, and they were included in the DAF in order to improve the validity (or construct validity and authenticity in the language of Bachman and Palmer) of the DIALANG assessment instruments. Checklists were designed for the item writing teams on the basis of the DAF categories to ensure that all relevant CEFR content was adequately represented in the items that they designed; the checklists also included categories derived from other sources than the CEFR (see below).

Because the CEFR and the Threshold, Waystage, and Vantage definitions were incorporated into the DAF, it could be argued that these Council of Europe documents are as essential to DIALANG as those documents developed by the project. In fact, the DAF did not reproduce all CEFR content categories but instructed the item writers to consult specific chapters in the CEFR for details.

Much of the construct validity and authenticity of DIALANG is thus based on the DAF. The key points for defining the constructs measured by DIALANG can be found in the sections of the CEFR that deal with communicative competences. The core linguistic and pragmatic competences defined in the CEFR are very familiar to language testers and applied linguists more generally, as they can be found in almost the same form in earlier, well-known
publications in the field (e.g. Bachman, 1990; Canale & Swain, 1980). The DAF specifies, however, that certain more general competences included in the CEFR are not the specific focus of DIALANG tests, although such competences are obviously present in some way in all language use, test-taking included.

Authenticity of DIALANG is obviously highly dependent on the successful capture in the DAF of the key CEFR (and other) categories pertaining to the situations in which the target language is used. DIALANG focuses on the use of language in the private, public, and at a rather general level, also in the occupational domains. It thus excludes situations which are typical of the educational domain and which take place in schools or other educational institutions. Also, very specific situations found in the occupational domain are not included in DIALANG either: In terms of content, there is thus only one version of DIALANG that measures ‘general’ rather than job-specific language proficiency.

The main documents that guided the actual item writing were the five DIALANG Assessment Specifications (DAS), which detailed how the theoretical basis defined in the DAF and the CEFR was to be operationalized as concrete test items. Specifications were designed for all the five skills or areas of proficiency covered in DIALANG, and they defined in detail the sub-skills and sub-categories to be tapped by test items, as well as the format and other characteristics of the items (CALS & DIALANG Project, 1998b, 1998c, 1998d, 1998e, 1998f, 1998g). It should be noted that the Specifications made considerable use of other sources than the CEFR, as the CEFR is not a recipe for test development but a more general inventory of conceptual categories involved in language learning, teaching, and assessment.

In particular, the CEFR does not offer much support to the testing of vocabulary and structures, and thus, the specification of the categories for vocabulary and structural knowledge were drawn from other sources. Also, the elaboration of reading, listening, and writing skills was based on previous work on these areas not included in the CEFR (see Alderson 2005; Huhta & Figueras 2004 for details). The DAS not only complemented the obvious gaps in the CEFR, but better definitions for some categories described in the CEFR were found in other literature on applied linguistics. For example, the definition of text forms and types was based on work by Werlich (1976, 1988) because that offered a more systematic basis than the CEFR for selecting texts for the reading tasks (Alderson 2005, 123).
However, the main problem with the CERF as the basis of constructing tests was that it lacked detailed descriptions of the constructs of interest (reading, listening, etc), especially from the cognitive point of view. Although the CEFR scales provide some help by describing at a fairly general level what the levels A1 to C2 mean in different areas of proficiency, they do not define the constructs of reading, listening, etc at the level of specificity that would have allowed the DAS to define precisely how to write items, for example, for level B1. The item writers were asked to write items that covered a range of different levels, and they were also asked to estimate the CEFR level of the items they wrote, to ensure that all levels would be covered. However, it was expected that guessing the level of say reading and listening items would be difficult based on earlier research (Alderson, 1990) which was why it was absolutely essential to pilot, calibrate and standard set the items. Indeed, the comparison of the predicted and empirical item difficulties (i.e., CERF levels) showed that the items writers were not very successful in predicting the actual level of the reading and listening items (Alderson 2005, 130 & 146).

It should be noted that the issue of using the CEFR for designing comprehension tests was later examined in greater depth in the Dutch Construct Project that developed a tool for linking the content of the test with CEFR categories (see Alderson et al. 2006 for details). At the time of the development of DIALANG, such tools were unavailable, however.

For defining writing in the DAS, extensive use was made of the model developed by Anneli Vähäpassi in the IEA International Study of Writing (Gorman, Purves, & Degenhart, 1988).

The only exception among the language tests in DIALANG was the Vocabulary Size Placement Test (VSPT), which was not based on the CEFR and which was not defined in the DAF and the DASes, except in passing when the entire test-taking process was described for the item writers in the specifications. The design of the VSPT was outsourced to Paul Meara, who had carried out considerable amount of research on yes/no vocabulary tests similar to the VSPT. Meara designed the tests in cooperation with the DIALANG language teams, who checked the words and pseudo-words proposed by him. The VSPT is based on lists of verbs in the tested languages and it serves as a quick instrument to place users of DIALANG on the most appropriate level of reading, listening, writing, vocabulary or structures test that they chose (see Alderson 2005, Alderson & Huhta 2005; Huhta 2007a, for more information about the design of the VSPT).
In addition to the Assessment Framework and Specifications, it was necessary to develop other, complementary documents to support further work on the test items, as they needed to be input into the item database, and an extra review of the items selected for piloting was also required. Also, the inputting and review of translations of the texts included in the system needed specific guidelines to ensure their clarity and comparability across languages (see Fligelstone & Huhta, 2001; Huhta & Noijons 2002; Huhta 2002). Activities such as item review were necessary in any case, but the fact that DIALANG runs on computer software poses somewhat different demands for the inputting and review of test items and translations compared with paper-and-pencil based testing systems, which needed specific guidance to succeed. Furthermore, the different components of software were completed only after the initial phases of the Project, hence it was necessary to complement the DAS with these additional, software-specific guidelines once the exact characteristics of the software were known.

It is important to add that the various documents intended for item developers and reviewers and for translators were only one aspect of the test development process. The item writers could not be expected to design high-quality items just by reading the relevant documentation, no matter how clear and detailed it may be. A necessary aspect of item development were numerous training sessions and countless one-to-one consultations with item writers, translators, and reviewers via e-mail, phone, and in face-to-face meetings during the project. Such training and consultation were done by the members of the coordinating centre (Centre for Applied Language Studies) in Phase 1 of the Project and by the Test Development and Translation Coordinators in the later phases (see Appendix 2).

Besides the documents and procedures related to item writing and review, the validity of DIALANG is very much dependent on the piloting of test items (and self-assessment statements, see below), analyses of the pilot test data that led to the construction of the tests with unique scoring rules for each (piloted) test, and the procedures to link the test results to the CEFR levels via standard setting. Detailed description of these aspects of the work is beyond the scope of the current thesis. Suffice it to say that the Project had to break new ground also in these areas, as DIALANG was the first test to be linked with the CEFR levels and because most standard setting methods had been developed for pass / fail decisions rather than for placing cut-off points for a multi-level scale (see Kaftandjieva, Verhelst & Takala
1999; Kaftandjieva & Takala 2002; Alderson 2005). Some of these procedures developed in DIALANG turned out to be robust enough to contribute significantly to the Council of Europe’s Manual on linking examinations with the CEFR (Council of Europe, 2004a, 2004b, 2009); Figueras et al. 2005; Alderson et al., 2006).

The five DIALANG Assessment Specifications together with the supplementary documents and procedures described above formed an essential and a more concrete level than the DAF in ensuring the construct validity and authenticity of DIALANG. They also contributed towards the interactiveness of the system by laying out the context in which the test tasks were to engage the test takers’ language competences, topical knowledge, etc. in a way that closely resembles the way in which their competences are engaged in non-test situations. The DIALANG documents do not explicitly discuss interactiveness, as the Bachman and Palmer framework was not directly used in the test development work, but the issues at hand were obviously familiar to the Project. The DASes and the various review guidelines, however, deal with the relationships between tasks and language abilities, in particular, in several places, and discuss how the two could be aligned and how construct-irrelevance could be avoided in designing tasks and items for specific skills, subskills, and areas of proficiency. The topics covered by the CEFR and DIALANG are rather varied but general in nature, and consequently, the specifications guided the item writers to select rather general topics (as defined in the CEFR and DAF) and to avoid highly technical and specific texts and topics that an average European adult would find unfamiliar.

It should be mentioned that there is an obvious contradiction between how DIALANG measures language proficiency and how language competences are defined in the CEFR, which may be an issue as regards the construct validity, authenticity, and interactiveness of DIALANG tests. DIALANG follows the traditional division of tests into reading, listening, and writing, and furthermore, into vocabulary and structures, whereas the model of language proficiency developed by Canale and Swain (1980), further refined by Bachman (Bachman, 1990), and included in the CEFR, defines language abilities somewhat differently. It is difficult to say to what extent this discrepancy in format leads to discrepancy in terms of authenticity and interactiveness. In general, it seems that DIALANG tests incorporate, for example, the ideas of sociolinguistic appropriacy and functional knowledge, which are so central in the models of communicative competence by Canale & Swain and Bachman.
One area of mismatch seems obvious, however, namely integrated vs. skill-specific tasks, use, and processing of language. Each DIALANG test, and the tasks that comprise the test, cover one major skill or area of language only, thus the tests do not engage users in such interactiveness that involves the integration of, say, listening and reading skills or listening and writing, except to a trivial extent when the learners have to produce very short replies to open-ended comprehension questions. Neither do the DIALANG tasks resemble integrative tasks that are possible according to the CEFR and that certainly are common in everyday life – which is obviously an issue for the authenticity of the tasks and, thus, for the domain of generalization in the interpretation of DIALANG test results.

Although integration of different tasks and skills is not unknown in language tests, it has nevertheless been quite rare in high-stakes examinations, for example. In this light, the recent revision of the TOEFL into TOEFL iBT is a major event, as it involves a radical rethinking of language constructs and their operationalization in tests (Chapelle et al., 2008).

Authenticity and interactiveness appear to be somewhat difficult aspects of usefulness to apply to a diagnostic test, however. It should be clear from the above discussion that if target language use involves a lot of integration of skills, a skills-based language test may be found somewhat lacking in terms of authenticity and interactiveness. It is not surprising that a proficiency test such as the TOEFL adopts a much more integrated approach in its test design than has been the case previously – it is in fact somewhat surprising that it has taken this long for a major international testing system to break away from the traditional four skills ‘straightjacket’. However, as Alderson (2005) argues, what is important for a proficiency test may be less relevant for a diagnostic test. In order to be able to provide the kind of detailed information about proficiency, a diagnostic test may have to be, at least partly, less communicative than most modern language tests are. Thus, it may be that authenticity and interactiveness – as defined by Bachman and Palmer – are aspects of usefulness in which the domain of generalization and the relationship between the task characteristics and language abilities is much more complex in diagnostic tests than in many other types of tests.
3.3.1.3 Feedback and self-assessment

DIALANG is not only a battery of language tests, and there is, thus, more to the a priori construct validity, authenticity, and interactivity of the system than what concerns the language tests themselves. Because DIALANG was intended to be a diagnostic system that helps its users to become aware of strengths and weaknesses in their proficiency and to plan further language studies, it was evident from the beginning that the system ought to provide learners with much more than just a single test score or mark. Diagnosing foreign language proficiency with the help of a large-scale, computerized system was quite a novel thing to do. In fact, the whole area of diagnostic testing of L2 skills had not been studied much, and this meant it was not obvious what kind of feedback the system should give and how. From the very beginning, self-assessment of language proficiency was considered one of the key parts of the system because of its key role in fostering learner autonomy and self-directedness (see Huhta et al. 2002; Huhta & Figureas 2004; Luoma 2004; Alderson 2005 for more information). A specific working group was set up to plan and design the feedback and self-assessment components for the DIALANG system (see Appendix 2).

The work on feedback and self-assessment became an important part of the a priori validation of DIALANG. Similarly to the DIALANG Assessment Framework and Assessment Specifications, the CEFR constituted an important theoretical basis for the development of self-assessment and feedback. The self-assessment instrument, in particular, was derived almost entirely and directly from the CEFR scale descriptors – and the comparison that users of DIALANG can make between their test results and self-assessments is based on the CEFR scale system. The CEFR was also used as the basis for the design of certain parts of the feedback, such as the descriptions of the CEFR levels that appear in the test results part of the system and in the ‘Advisory feedback’ (extended level descriptions), as described in Huhta & Figueras (2004; see also Alderson & Huhta 2005; Alderson 2005; see also Appendix C in the CEFR). However, some of the feedback available in DIALANG is not directly based on the CEFR but is grounded on specific test items (the review of answers to the items; information about the sub-skills tested by each item) or comes from other sources (the VSPT feedback; information about self-assessment; advice on how to make progress from one CEFR level to the next). Since the test items were developed with reference to the CEFR, feedback on individual items, too, could be seen to be linked with the Framework, at least indirectly.
In the piloting of DIALANG, two types of self-assessment were used. The first was a six-level scale that described the CEFR levels A1 – C2 for a particular skill covered by DIALANG (i.e., reading, listening, writing). The second was a set of ‘can do’ statements related to the same skill as the overall self-assessment scale. Both SA instruments were derived directly from the CEFR scales but some of the can dos were slightly modified to simplify the language for ordinary learners who may not have been familiar with some of the terms or expressions used in the original statements (see Huhta & Figueras 2004; Luoma 2004; both instruments can be found in 14 languages in the DIALANG Project website, which is accessible via the DIALANG website at www.dialang.org).

During the data analyses of the pilot test results, the quality of individual self-assessment statements was analyzed and 18 best statements were selected to be included in the operational version of the assessment system. Analyses were also carried out on how the self-assessment statements could work as a pre-estimator of learners’ proficiency in the skill in question because there were to be three levels of tests in DIALANG – easy, intermediate, and hard – and self-assessment was planned not only to provide feedback to the learners on the match between their self-assessments and test results but it was also to help the system to place them to the most appropriate level of test. Thus, an algorithm had to be created on the basis of the analyses that would estimate the learners’ likely level of proficiency on the basis of the pattern of responses to the SA statements, which also involved determining two cut-off points for the SA score scale for the automatic placement decisions that the system had to make. In addition, the result from the SA task had to be combined with the learners’ result from the Vocabulary Size Placement Test, if the learner takes the VSPT, because also the VSPT was to be used for pre-estimation purposes (see Luoma 2004; Alderson 2005; Alderson & Huhta 2005, for more information on the development of the self-assessment for DIALANG).

For the operational DIALANG, only the more detailed self-assessment instruments based on 18 individual SA statements per skill were selected. The main reason for this was that two self-assessment tasks were considered unnecessary. Furthermore, some previous research suggested that the detailed SA might be easier for many learners as it describes more detailed instances of language use than the broader overall scales (see the review by Oskarsson 1984, 31). It could be mentioned, however, that the Project seriously considered changing the
content and functioning of the self-assessment system in the later stages of the project. The reason for this was that responding to the detailed SA statements does not allow learners to form a clear idea of where they stand on the CEFR scale in the skill tested – the learners cannot see a full CEFR scale and ponder which level on the scale best matches their own idea of their proficiency. In fact, a learner who takes DIALANG for the first time does not see a full six-point CEFR scale until reviewing his or her test result. The Project considered making the overall self-assessment based on the six-point scales the main SA instrument of the system – and perhaps even a compulsory part of the test-taking process – and making the detailed SA optional, for those who would be interested in more in-depth self-assessment activities. However, the need to allocate all programming resources to completing the missing vital elements of the system prevented any substantial changes to the way the self-assessment works in the current version of DIALANG.

The *a priori* construct validity, authenticity, and possibly also interactivity of feedback and self-assessment statements is to a considerable extent dependent on the definitions and scales found in the CEFR and on their operationalization in the DIALANG system – as is the case of the test items. Thus, much of the quality of the self-assessment statements and the feedback that relates directly to the language tests (the main test result on the CERF scale, comparison of SA and test result, but also the extended CEFR level descriptions and perhaps item level feedback) depends on the quality of the CEFR. However, some of the feedback is not derived from the CERF but from a range of other sources, and there the quality of the sources as well as the expertise of the persons making the selection (i.e. the Self-assessment and Feedback Working Group) is decisive.

It should be noted that *a priori* validation in terms of construct validity, authenticity and interactivity is rather more straightforward in the case of self-assessment than feedback. Self-assessment instruments are very similar to language tests: in both of them the learner completes tasks or answers questions, and the outcome of both is a score, grade or level. **Feedback is not, however, a unitary concept** (see also Mory 1996, 2004). Some of it is inseparable from the tests – for example, the test result is feedback that is directly related to the test. Thus, whatever applies to the *a priori* validation and theoretical arguments for the three aspects of the usefulness of DIALANG presented above also applies to feedback on test results, and, as was just pointed out, to the feedback on self-assessment and its match with the test result. In addition to the test result (i.e. the CEFR level and its verbal description),
other DIALANG feedback that is directly related to the language tests include the extended descriptions of the CEFR levels that learners can study separately from the regular, short description that the users of DIALANG can see when they view their test results. Also, the immediate and post-test review of responses to individual test items is such test-dependent feedback, as is the display of results categorized by sub-skills or sub-areas of reading, listening, writing, vocabulary, and structures that is part of the delayed, post-test item review.

DIALANG contains at least two types of feedback that may not lend themselves to be evaluated in exactly the same way as the tests and test-dependent feedback: advice to the test taker on how to improve his/her proficiency towards the next CEFR level in reading, listening, or writing, and the explanations of why a learner’s self-assessment may not match his/her test result.

In a certain way, these two types of feedback are related to the results of the tests and self-assessments, however. The advice on making progress towards upper CEFR levels is language, skill and level-specific. If you have taken a Danish reading comprehension test in DIALANG with the result of ‘C1’, and you wish to see what the ‘Advice’ section of the feedback menu contains, you will be taken to the screen that displays the extended description of what typical learners at your level (i.e. C1) can read, and you can compare that with a description of the two adjacent levels B2 and C2. From the screen, you will have direct access to the section of advice on language learning that is targeted at improving level C1 readers’ skills so that they could reach the next level C2 (see Figure 2). Thus, the advice is level-specific, and the relevance (usefulness) of the advice is obviously dependent on the correctness of the reading test result. Some of the advice also mentions the language tested but this does not suggest that the advice is only relevant to that language – Danish in this example. It rather illustrates one of the ways in which DIALANG feedback was individualized to users.
The following suggestions may help you make progress towards C2:

- Actively enlarge your vocabulary by, for example, deciding for a period to look up all new unknown words in a dictionary or by questioning particular meanings of known words in context or by attempting to notice and "collect" idiomatic usages and the contexts in which they appear.

- Try to learn the fine distinctions of meaning between similar expressions (see for example dictionaries and guidebooks on style and writing).

- Pay attention to how the writers put the intended message across (openly or in hidden ways), what resources are used and whether they consider themselves as outsiders or as co-participants in the actions they describe, whether they consider themselves part of the audience they write to or not.

- Access the Danish version of the instructions for new machines rather than those in your first language.

- Keep reading Danish popular novels as a leisure activity and read non-literary texts in your field or area of interest in preference to those in your first language.

Figure 2. Example of advice type of feedback in DIALANG: advice given to the users who has taken Danish reading comprehension tests, with English as the interface language, and who has received level C1 as their test result.

However, the Self-assessment and Feedback Working Group recognized the impossibility of categorizing advice in such a way that only certain advice would be given to those who had received A1 and who wanted to progress towards A2, whereas different advice would be given to those moving from A2 to B1, and so on. No existing theory on language learning could guide the linking of advice and proficiency levels, and thus experience of language educators, textbook writers, and whatever analyses existed of the ways in which different learners go about learning languages had to be used in deciding which advice to place at different points of the CEFR scale. Certainly, some of the advice makes sense to beginners only whereas only advanced learners can benefit from doing certain other things, so it is possible to rank the advice in very broad terms according to learners’ proficiency level. In the current version of DIALANG, however, the advice is displayed level by level and the user can only read one set at a time. This is not the ideal solution but it was technically and
navigationally the least problematic solution. Direct and easy access to the other sets of advice for the other levels is provided to the users via buttons on the upper part of each advice screen (see Figure 2 above), which partly tackles the difficulty that the Project had to decide, in a theoretically sound way, which advice to give to learners at different CEFR levels.

The feedback that gives DIALANG users information about possible reasons for a mismatch between self-assessments and test results is also related to test results, at least to an extent. These explanations are accessible to anybody taking DIALANG tests but the likely readers are obviously those whose self-assessment feedback indicated a mismatch between their SA and test. Thus, the relevance of the explanations is to an extent dependent on the validity of both the test results and the self-assessments. If either the test result or the result of the self-assessment is not accurate, feedback on the mismatch – or the match – between the two is not accurate either. However, some of the information in this section of the feedback goes beyond describing potential reasons for a mismatch and elaborate on, for example, what kind of information about one’s proficiency different kinds of language tests can or cannot provide. These texts, then, have a slightly different approach to raising learners’ awareness of language learning and proficiency from the other feedback in DIALANG.

As was the case with deciding on where on the CEFR scale to place different pieces of advice on language learning, the lack of a sound theoretical basis and the scarcity of relevant research results hampered the design of the explanations for mismatch between SA and test results. Again, common sense, educators’ experience and whatever little research existed had to be relied on (e.g. McFarland & Miller 1994, and the studies reviewed in Oskarsson 1984;).

To summarize the theoretical arguments for the construct validity, authenticity and interactiveness of the two types of DIALANG feedback discussed above, it could be said that they are partly based on the test and self-assessment results, which means that the quality of the tests and self-assessments partially determines their quality and usefulness. However, the fact that the content of the two sets of feedback (advice and reasons for SA / test mismatch) is not based on such developed theories or frameworks as the tests themselves obviously makes their a priori usefulness less certain. It is also less straightforward to conceptualize what construct validity, authenticity and interactiveness actually mean for these two types of feedback. Perhaps the best one can say about them is that if this feedback manages to capture,
for example, important metacognitive skills that relate to learning to learn and awareness about learning, they can argued to possess construct validity and authenticity. And if the feedback can actually make learners engage their metacognitive skills when reading such feedback, and perhaps also afterwards in different learning contexts, then the feedback may be argued to be high in terms of its interactiveness.

3.3.1.4 Computer as a testing platform

As DIALANG runs on computer software, considerable time and resources had to be devoted to software design and programming (Alderson 2005). It is essential for a computerized assessment system to function in a way that makes the test-taking experience as smooth as possible so that, for example, the way in which the learner interacts with the test tasks would be as ‘natural’ as possible. For example, it should be clear to the learner how he or she ought to reply to different kinds of items. Navigation in the system should be as easy and intuitive as possible. Thus, the design of the system as a whole and in particular its interface were matters of great importance to the project, and several reports, guidelines and specifications were drawn in this area of the project to ensure that the software underlying the system, as well as its interface, were on a par with its content (e.g., Fligelstone & Treweek 2002; Fligelstone 2001; Scott 2001; Figueras et al. 2000).

Since DIALANG is a computer-based assessment system the medium by which its users take the tests differs from the medium that is still the most familiar one to most test takers (i.e., paper and pencil). It is possible that computerized tests measure somewhat different constructs than paper-and-pencil tests (Chapelle & Douglas 2006, 12-15). The issue is particularly important with computerized tests that automatically rate speaking and writing performances, including answers to short-answer questions, because the algorithms developed to analyze test takers language do not use exactly the same criteria as a human rater would – or more precisely, computers can rate performances by using only some of the criteria that human judges can potentially use. Since DIALANG does not analyze learners’ performances but only compares their responses with the list of acceptable answers that have been input for each individual test item, the main construct issue with computerized tests is less relevant in this case. However, it is still true that the use of DIALANG requires certain basic IT skills and some minimum level of familiarity with computer programmes – simply because it is a computer programme. It is also possible that computer anxiety and attitudes to
computers may affect the way the learners can demonstrate their language proficiency via a computerized test (McDonald 2002). Because DIALANG tests have some fill-in and short-answer items where the learners have to write one or more words in order to complete the task, the construct validity of such items is potentially an issue. It is possible that some correct answers get marked as incorrect because the list of acceptable answers in the scoring key in the system is not complete.

The main issue with a computerized test like DIALANG is probably the fact that the test tasks used in it may change the typical interaction between the learner and the task, even if the task on the computer screen looks exactly the same as on paper. In the language of the Bachman and Palmer framework, the interactiveness between the test takers’ characteristics and a computerized test may be different from interactiveness with a paper-and-pencil test. The likelihood of the interaction being very different is obviously greater with computer-based tests that use very innovative items that cannot be used in paper-and-pencil tests (Chapelle & Douglas 2006, 40). As the current version of DIALANG uses only very traditional item types, such a threat (if it is considered a threat in the first place) is quite small.

Despite the worries that computerized tests measure somewhat different constructs than paper-and-pencil tests and that the interaction between test takers’ characteristics and test tasks may be different, it was clear from the beginning of the project that the computer would be the only feasible platform for a large-scale diagnostic test as DIALANG. Delivering feedback to large numbers of learners across many languages, skills, and sub-skills in a reasonable time and with reasonable costs would be unthinkable with a paper-based system. For the feedback to be effective, it needs to be delivered to the learners as soon as possible – and certainly not after several weeks or months as is the case with international paper-based examinations. Thus, although there were obvious practical questions that dictated the format of DIALANG, the purely construct-related considerations on how a diagnostic testing system delivering feedback should work made the computer the obvious choice as the delivery platform (see Alderson 2005; Mory 1996, 2004). In addition, the Project’s intention was to use more innovative item types in DIALANG than what its first operational version uses which would fully utilize the strengths of the IT (Alderson 2005, 221-242, and http://www.lancs.ac.uk/fss/projects/linguistics/experimental/start.htm), precisely because certain new item types might be better for diagnostic purposes than, for example, the
traditional multiple-choice and gap-fill. Furthermore, the computer as a medium for anything is increasingly the default platform in the modern world, at least in the developed countries, and the issue of the difference between computerized and paper-and-pencil tests must be considered in a different light – it is all right for computerized tests to be different as long as their users can be familiarized with the specific skills and knowledge that taking them might require.

3.3.2 Empirical evidence about the construct validity, authenticity and interactivity of DIALANG

This section will present empirical evidence mainly about the construct validity for DIALANG. However, the evidence discussed below is not purely about construct validity in the sense defined by Bachman and Palmer (1996), although the great majority of evidence focuses on that aspect of usefulness. The discussion starts by making the point that the distinction between reliability and construct validity is not necessarily always clear and that the evidence about the reliability of DIALANG in fact contributes at least partly to its construct validity. The second reason why the discussion below is not only about construct validity is the intertwined nature of construct validity, authenticity, and interactivity aspects of usefulness, as was elaborated in Section 2 above (see also Figure 1). Furthermore, there is rather little direct empirical evidence about the authenticity and interactivity of DIALANG, which also makes is sensible to discuss them together with the far more extensive evidence on construct validity. Thus, the implications of the various research findings to the authenticity and interactivity of DIALANG will be mentioned below only when the studies appear to shed light not only on the construct validity but also on these other closely related aspects of usefulness.

3.3.2.1 Evidence about reliability is also evidence about construct validity

Bachman and Palmer (1996) make a distinction between reliability and (construct) validity, as most measurement specialists do. The two concepts are closely related because reliability is generally considered a necessary but not sufficient prerequisite of validity. Also, the two are sometimes so intertwined that they cannot be separated easily. For example, it is not straightforward to decide if the accuracy of the placement procedures used in DIALANG is a matter of reliability or validity. The study of the Vocabulary Size Placement Test (Huhta
2007a) was presented in Section 3.1.2 as evidence of certain problems in the reliability of the placement procedure. However, if one considers the prototypical validity questions ‘Does the VSPT test what it is supposed to test?’ or ‘Are the inferences made from the VSPT results correct?’, the problems discovered in the VSPT might as well be considered validity problems. It was found that one of the reasons for unreliable, or invalid, VSPT scores was a ‘wrong’ test-taking strategy, namely guessing. Is that a reliability or validity problem? Or perhaps the problem was that the interaction between that test and some test takers’ language proficiency did not work the way the test designers had planned. If this interpretation is the most accurate one, then at least some of the empirical evidence about the reliability of the VSPT is in fact (also) evidence about the interactiveness – or problems in that – for that part of DIALANG.

The point here is not to try to establish beyond reasonable doubt differences or similarities between reliability and validity in the Bachman & Palmer framework but to make the point that since reliability is usually considered a necessary prerequisite for validity, all reliability evidence can also be considered one type of validity evidence. Thus, all the empirical evidence about reliability of DIALANG presented above and detailed in Huhta & Alderson (2005) and Huhta (2007a & 2007b) contribute also to the validity of DIALANG.

### 3.3.2.2 Concurrent, criterion-related validity evidence

A very common approach to study construct validity empirically is to examine how the test of interest relates to other tests that are known or believed to measure either very similar or very different properties (see e.g., Cumming & Berwick 1996; Chapelle 1999; Weir 2005). Tests that are supposed to measure very similar skills should produce very similar results as the test of concern, whereas tests measuring very different skills should yield results that are only remotely if at all related. The tests (i.e. external criteria or criterion measures) with which a particular test is compared may be part of the same test battery that was administered to test takers at the same time as the test of interest, or the tests may be truly external measures that the learners take either before or after the test that is being studied. A somewhat different type of use of external criteria is involved in predictive studies. In such research, investigators are interested in finding out if the language test scores obtained by different groups of learners are in line with what is known about the proficiency of the groups...
(e.g. number of years they have studied the language tested) or if the test results correctly place learners to different types or levels of courses.

There is some empirical evidence about the concurrent validity of DIALANG tests as concerns all three types of evidence above: the relationships (1) among different DIALANG tests, (2) between the DIALANG tests and external measures of proficiency, as well as (3) the ability of DIALANG tests to predict the proficiency levels of different, well-defined groups of learners.

A. Internal structure of DIALANG test batteries

The first type of concurrent validity evidence traditionally concerns the correlations between different tests in the same battery of tests, as they indicate to what extent the tests measure the same or different aspects of, for example, language proficiency. If the battery happens to include two or more tests that are intended to measure the same area of ability – for example reading – it can be expected that these reading tests correlate more strongly with each other than with, say, tests of speaking or listening abilities.

The interpretation of concurrent, correlational validity evidence requires at least two kinds of information to be meaningful: information about the constructs that the tests purport to measure and information about the test takers’ abilities. We need to have a fairly good idea of what constructs each of the tests in the test battery is likely to measure, on the basis of theory, previous research into the constructs in question, and the test specifications for the tests to be studied, to name the most obvious sources of a priori information about the relationships between the tests and constructs. In the case of DIALANG, a conscious effort was made to create tests that would tap five different areas of language proficiency: reading, listening, writing, structures, and vocabulary. Thus, we would expect the tests not to have perfect inter-correlations. If two tests were found to correlate with each other extremely strongly, one could question the need and practicality to test that same skill with two tests instead of only one. On the other hand, however, previous research indicates that language skills tend to be correlated and that it is often possible to find one, common ‘general’ language factor underlying batteries of language tests (see Skehan’s 1988 review). Thus, it could be expected that the DIALANG tests, too, would correlate with each other. Not to find
at least moderate correlations between the tests would in fact be very surprising and difficult to explain, and it would probably lead to a suspicion that something is wrong with the tests.

Secondly, we need to have a fairly good idea of the test takers’ relevant background characteristics, the most important of which is probably their level of proficiency. This is because previous research suggests that the relationship between language skills is likely to differ depending on the learner’s overall level of proficiency. Specifically, research into the structure of language proficiency in the 1970s and 1980s demonstrated that the proficiency of advanced learners tended to be more uniform than that of beginners (Skehan, 1988, 213). That is, it is common for the advanced learners to have developed all or most of their language abilities to a fairly high level – it is somewhat uncommon to find otherwise advanced learners whose reading or speaking, for example, would significantly lag behind the other skills. Their profile of language skills could thus be described as ‘even’. On the contrary, it is not at all uncommon to encounter beginners or intermediate learners who have developed one or two areas of proficiency to a clearly higher level than the other areas. Probably the most important explanatory factor of this is simply that the beginning learners have had less time and a fewer number of opportunities to learn and use the language than advanced learners, and thus their different individual needs and interests cause them to focus on certain types of language activities that involve some skills more than others. Some may develop good oral communication skills but may not need to write or read much in what they do. Others may need to read and study a lot but lack the opportunity or need to speak in the foreign language. All this results in what can be described as an ‘uneven’ or ‘peaked’ skill profile.

In order to interpret correlational evidence for the internal structure of DIALANG tests, we need, thus, know not only about the theoretical basis of the tests but it would also be useful to know as much as possible about the language learners whose test performances are used for concurrent validation purposes. It should be mentioned at this point that knowing about the tests takers and their likely level of proficiency is also important to interpret concurrent validity evidence that is based on comparing the tests with some external measures, such as other language tests, teacher grades or even self-assessments.

In the light of the above, it may thus be useful to study also the profiles of test results across the tests, in addition to the magnitude of correlations between different DIALANG tests (test
result profiles and (rank-order) correlations between tests are obviously related but do not yield exactly the same information about learners’ proficiency across skills. This is obviously somewhat tentative because the design of the studies reported here varies, as do the learners studied in them, and we may not always know as much as we would like about the learners involved. However, since quite a lot of such evidence exists, it would be unwise to ignore it because it can potentially provide evidence for or against the interpretations we can draw from DIALANG results.

Should we assume that levels in the numerous CEFR scales for different skills (e.g., reading, writing, speaking) or for different activities (e.g. writing notes and messages, listening to public announcements) are equally easy or difficult? Is writing a letter at level B1 equally demanding as understanding interaction between native speakers at level B1?

It seems unlikely that full cognitive, social or linguistic correspondence could ever be established between so different activities or types of skills. Admittedly, the CEFR does not claim that such correspondence exists. What it does argue is that B1 in one language can be thought to correspond, in communicative terms, with B1 in another language. However, the fact that the same six main levels A1 – C2 are used in the CEFR to describe proficiency in a range of very different linguistic skills and activities certainly implies that some rough equivalence exists between, say, B1 reading and B1 speaking. The CEFR ties performance across various skills together with the help of larger scale contexts such as ‘everyday communication’ or ‘basic personal needs’. For example, it is characteristic of level B1 that the skills and activities typically involved at that level concern common, everyday situations related to work, travel, and everyday living in which fairly straightforward, non-complex communication is required. It is thus this context that helps one to define and understand how well different aspects of language should be mastered in order to be called B1 proficiency.

The DIALANG Project neatly avoided the issue of equivalence of skills by designing different tests for reading, listening, writing, vocabulary, and structures, and by scaling and constructing these tests independently of each other. DIALANG aims to be diagnostic by providing feedback separately on the five areas of proficiency, and it does not provide overall results for a language across skills – unlike certification tests that may only report an overall grade. From that skill-specific point of view, it is not crucial to establish equivalent levels across the different skills tested. This was, then, the main reason why the pilot test results
were analyzed, and the test items calibrated, separately for each skill, in the languages where piloting was completed before the project came to an end.

On the other hand, however, one of the aims of DIALANG is to provide the learners with a profile of strengths and weaknesses, and one obvious profile that the learner can obtain from the system is his / her level of proficiency across the skills. The system does not, in fact, show a summary profile of results based on all the tests that the learner has taken because the results are reported only one test at a time but interested users are likely to write down their results and note if there are lower results (i.e., weaknesses) and higher results (i.e., strengths) in different main skills, and may then plan their further language studies accordingly. Thus, DIALANG cannot totally ignore the question about equivalence of levels across tests. In hindsight, it would have been interesting to run the calibration analyses twice, first skill by skill (as was done), and then re-calibrate all piloted items across all the five areas of proficiency together. The latter may have allowed the Project to link all five tests together with a help of a common scale. Obviously, such a joint analysis would have required decisions on such issues as what to do when slightly different items turn out misfitting depending on the type of analysis (joint vs. skill-specific). Overall, it is possible that a joint analyses of five different areas of proficiency would have resulted in problems because Rasch analyses assume unidimensionality, i.e., that all items test the same skill (see, however, Henning, 1992 who argues that it is possible to achieve psychometric unidimensionality even if the measured construct is multidimensional).

What can we expect to learn from the examination of the DIALANG score profiles discovered in various studies? Probably the most important finding would be if it could be established that one or more of the five tests in a particular language yields results that are systematically lower or higher than the results of the other tests, even with groups of learners who are likely to have rather even skill profiles. This finding would indicate either that the way the CEFR defines a particular skill (reading, writing, or listening in this case) is somehow different from the others or that there is something wrong in the way DIALANG converts test scores to CEFR levels. The key phase of work that is most likely to cause a ‘wrong’ conversion between a test score and the CEFR level that is reported to the learner are the standard setting activities (see Alderson 2005 for details on standard setting). Standard setting largely depends on human judgement and is thus prone to bias no matter how carefully it is done. Different standard setting methods have also been shown to produce
different cut-off points and thus change the results reported to the test takers (see e.g. Cizek 2006; Kaftandjieva 2004). A particularly vulnerable part of DIALANG when it comes to linkage with the CEFR levels are the structures and vocabulary tests because there was very little in the CEFR that could be used as a basis for constructing these tests and the descriptive structure and vocabulary scales used in the standard setting exercises for these skills (see Alderson 2005).

On the whole, all DIALANG profiles should be interpreted with caution. However, if we notice that a DIALANG test profile is rather even when the test takers have mostly been advanced learners, this may provide some validity evidence for DIALANG because this would be consistent with the findings that advanced learners have fairly even skill profiles. Also, if we notice that beginners’ or intermediate learners’ DIALANG profiles are at least partly uneven, this would provide similar validity evidence. However, the discovery of an uneven profile with advanced learners or an even profile with beginners would be more complex to interpret because it is entirely possible to find groups of advanced learners with uneven profiles and beginners with even profiles – the tendencies found in language testing literature are only tendencies and may not apply to all groups of learners.

There are at least four studies that provide us with information about the internal structure of DIALANG tests, either in terms of correlations between the tests or in terms of score profiles across the tests, or both. The first is based on the English pilot test data reported in Alderson (2005) and Alderson and Huhta (2005). The others are based on using the operational version of DIALANG for research purposes at a Finnish polytechnic university (Jaatinen 2005) and across a range of Finnish and German educational institutions (the present author’s survey of DIALANG users reported in Huhta 2007a, 2007b, and this paper). Also, a German study on university students’ English proficiency yielded some relevant information (Peschel et al. 2006).

a. The DIALANG pilot test study

The first study that yielded empirical evidence about the internal structure of DIALANG tests was based on the pilot test data and concerned only the English tests. The results of the study are reported in detail in Alderson (2005) and more selectively in Alderson and Huhta (2005). Because of the design of the pilot booklets, the informants only took certain combinations of
skills (e.g. reading and grammar items or listening and vocabulary), which means that not all skills can be directly compared by computing correlation coefficients between them. However, because the test takers were randomly assigned to different pilot test booklets, it should be possible to compare the distribution of test scores across skills.

Table 3 shows that the correlations between those pairs of skills in the DIALANG pilot test data for English are quite strong but far from perfect. They are, thus, of the magnitude that could be expected from a test battery that includes tests that aim at tapping skills or areas of proficiency that can be theoretically and logically regarded as different skills. On the other hand, as is the case with most language test batteries, the tests are quite clearly associated with each other, which is in line with the idea of an underlying ‘general language’ competence that is often found in empirical research into language proficiency. These findings are thus in line with theoretical expectations and previous research. They also suggest that it is meaningful to provide separate diagnostic feedback on each of the skills tested in DIALANG because the tests appear to measure somewhat different aspects of proficiency.

**Table 3. Correlations between different skills in the English pilot test data**

<table>
<thead>
<tr>
<th></th>
<th>Structures</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>.685</td>
<td>-</td>
</tr>
<tr>
<td>Listening</td>
<td>-</td>
<td>.650</td>
</tr>
<tr>
<td>Writing</td>
<td>769</td>
<td>.795</td>
</tr>
</tbody>
</table>

Figures 3 and 4 display the distribution of English pilot test results, as CEFR levels, across the five areas of proficiency. The previous studies (Alderson 2005; Alderson & Huhta 2005) do not report these which is why the present author calculated them from the pilot test data for this paper, so that we see how these results were actually distributed across the CEFR levels in the five skills. The conversion of the pilot test takers’ scores (or IRT-based ability estimates, to be precise) to CEFR levels was done by applying two different sets of cut-off points for the English tests. As Alderson (2005, 61-78) describes, three different types of standard setting procedures were tried out during the development of DIALANG, and English was standard set twice. In addition, the programme developed at CITO by Norman
Verhelst to analyze the standard setting data also evolved over time and with experience, so that the cut-off points calculated from the same data were different, depending on the version of the programme.

Here, two different sets of cut-off points calculated for the English tests have been applied. Figure 3 presents the results of applying an older set of cut-off points, which are in fact very close to the ones used by the current version of DIALANG, if not identical. Figure 4 displays the results of the pilot test data when a newer set of cut-off points are applied; these were calculated for the English tests during the detailed analyses of the pilot test data in 2003 but have not been implemented in the operational DIALANG.

Figure 3. Results of the DIALANG English pilot tests with an older set of cut-off points applied

N: reading = 624, writing = 645, listening = 534, structures = 1084, vocabulary = 975
Let us examine what the application of two different cut-off points mean for the internal (correlational) structure of the English tests in DIALANG. First, we can see that the older cut-off points are more lenient: when they are applied, there are many more test takers who are awarded C1 or C2 on the basis of their pilot test results (Figure 3) compared with the newer cut-offs (Figure 4). Second, we can notice that the shape of the score distribution changes more with some tests than others depending on which cut-off points are used. Listening is the test whose score distribution changes the least, which is also indicated by a very high (.928) rank order correlation between the two ‘versions’ of listening tests in Figure 3 and 4. Writing also changes only a little (.912), although the number of C1-C2 recipients decreases considerably when one moves from the old to new cut-offs. Vocabulary changes the most, and there the correlation between two different cut-off based results is only .781. Also reading results change considerably if different cut-offs are applied (.822). Grammar results change, too, but not quite as much (.865).

Figure 4. Results of the DIALANG English pilot tests with the 2003 cut-off points applied
Do different cut-off points change the correlations between different tests? That is, does the evidence for the internal structure, and thus for the construct validity of the DIALANG tests change if the cut-off points are changed? Table 4 examines this question. For the sake of comparison, the original Spearman rank order correlation coefficients reported in Table 3 (and in Alderson 2005, and in Alderson & Huhta 2005) are included in this table. Because those correlations were calculated from the precise IRT-based ability values for each informant, they are obviously somewhat higher than the respective coefficients calculated from the converted CEFR-scaled results that are based only on the more coarse 6-point scale. What is of interest here is to compare the correlations for the two different cut-off points that both refer to the CEFR scale.

Table 4. Changes in the correlations between different skills in the DIALANG English pilot test data depending on how they are calculated and which cut-off points are used (Alderson 2005, 188 & 205)

<table>
<thead>
<tr>
<th>Skill</th>
<th>Structures</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reading</strong></td>
<td>IRT-ability estimate</td>
<td>.685</td>
</tr>
<tr>
<td></td>
<td>CEFR level (old cut-offs)</td>
<td>.673</td>
</tr>
<tr>
<td></td>
<td>CEFR level (new cut-offs)</td>
<td>.562</td>
</tr>
<tr>
<td><strong>Listening</strong></td>
<td>IRT-ability estimate</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>CEFR level (old cut-offs)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>CEFR level (new cut-offs)</td>
<td>-</td>
</tr>
<tr>
<td><strong>Writing</strong></td>
<td>IRT-ability estimate</td>
<td>.769</td>
</tr>
<tr>
<td></td>
<td>CEFR level (old cut-offs)</td>
<td>.724</td>
</tr>
<tr>
<td></td>
<td>CEFR level (new cut-offs)</td>
<td>.669</td>
</tr>
</tbody>
</table>

Table 4 displays a clear trend in the correlations: the intercorrelations of the tests are higher when the older cut-off points are used in all the pairs of tests that can be calculated from the pilot test data. The change in the magnitude of correlations ranges from about .05 to .12, which is not a terribly big change but it can nevertheless decrease the amount of shared variance between a pair of tests by as much as 15%. Overall, however, the correlations between different DIALANG tests remain highly significant (at .000 level) and **moderately high irrespective of the cut-offs used**, so that the **overall conclusion** about the construct validity of DIALANG presented above **does not change**.
To further examine potential evidence that score distributions of the DIALANG pilot tests can offer, two extreme groups of learners were compared. The first group are beginning English learners, who are defined as those who reported having studied the language either less than one year or between one and two years (there were 44 – 107 such learners, depending on the skill). The second group are experienced or advanced learners who are here defined as those who had studied English for at least nine years (254 – 427 learners per skill). As mentioned above, previous research indicates that beginners often have an uneven skill profile in the language they study whereas the profile of more advanced learners can be expected to be more even.

Figure 5 below presents the score profile of the beginners and Figure 6 shows the profile of the advanced learners in the DIALANG pilot test data (the old cut-off points are used here). Overall, the test results seem to be in line with the prediction above: the profile of the advanced learners is indeed quite even, with the slight exception that there are somewhat fewer low (A1-B1) results for structures and vocabulary. In contrast, the picture for the beginners is clearly more varied, and it appears that listening, in particularly, differs from the other tests. Also, structures and vocabulary differ from the three other tests in their general distribution of results, and, like the advanced learners, the beginners, too, found these two areas of proficiency somewhat easier than the three skills of reading, writing, and listening.

Interestingly, the results of the advanced learners suggest that perhaps the English reading test is somewhat more difficult than the other tests, as very few of even the most experienced learners in the pilot tests achieved level C2. The result can also suggest that the cut-off point for C2 is higher for reading than for the other tests. Counterintuitive as it may seem, if the cut-off points for the English reading tests are higher than for the other skills, this can in fact be caused by the relative easiness of the reading tests due to the lack of really difficult items. Alderson (2005, 130) notes that there is only one C1 and no C2 reading items in the English tests; some of the items intended to be really demanding reading items had be discarded on the basis of the results of the item analyses and detailed expert reviews. It is in fact the case that in order to reach C2 in reading in English in DIALANG, the learner needs to get every item correct even in the hardest test version. No wonder so few pilot test takers (or anybody else; see the other studies reported on below) reached the highest level in reading in English.
Figure 5. Score distribution of the beginning English learners in the DIALANG pilot tests (learners had studied the language for two years or less)

Figure 6. Score distribution of the advanced English learners in the DIALANG pilot tests (learners had studied the language for at least nine years)
b. Present author’s survey study of DIALANG users

The second study yielding information about the intercorrelations of DIALANG tests is the survey of over 550 DIALANG users by the present author (reported partly in Huhta 2007a, 2007b, and more fully in this paper). In this study, the learners took operational versions of the language tests. The informants were asked to report the results they had obtained in the DIALANG tests they had taken by the time they replied to the questionnaire. For the correlational analyses, only those informants who had taken tests in only one language were selected, in order to guarantee the meaningfulness of the findings. About 100 – 150 learners per skill who had reported their test results and who had taken tests only in one language qualified for this analysis; the most common language taken by them was English, but also Finnish and French were fairly common as test languages.

Table 5 shows the correlations between the operational DIALANG tests taken by the learners. The intercorrelations of the five main skills tests ranged from .449 to .690, and they are in the same range as those reported in Table 4 above for the CEFR-scaled results based on the old cut-off points (which are practically the same as those used in the operational version of DIALANG). The two studies are based on different groups of learners, which may explain the slight differences in the correlations found: in the Alderson study, they came from all over Europe, and even around the world, whereas the study by the present author was based mostly on Finnish and to a lesser extent German and international learners. The range of learners studied also varied (there were fewer beginners in the author’s study), and it included a somewhat different mixture of languages: not all informants in the survey had taken only English tests as in Alderson’s study. Table 5 shows that the lowest correlation (under .5) was found between listening and writing; on the whole, the correlations for listening were slightly lower than for the other skills. The highest correlation was found between the vocabulary and structures tests (.690).
Table 5. Correlations between DIALANG tests in the author’s survey of users of the operational system

<table>
<thead>
<tr>
<th></th>
<th>Reading</th>
<th>Listening</th>
<th>Writing</th>
<th>Vocabulary</th>
<th>Structures</th>
<th>VSPT (Banded)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>1.000</td>
<td>.523</td>
<td>.653</td>
<td>.639</td>
<td>.568</td>
<td>.520</td>
</tr>
<tr>
<td>N</td>
<td>274</td>
<td>112</td>
<td>108</td>
<td>152</td>
<td>133</td>
<td>172</td>
</tr>
<tr>
<td>Listening</td>
<td>.523</td>
<td>1.000</td>
<td>.449</td>
<td>.550</td>
<td>.545</td>
<td>.321</td>
</tr>
<tr>
<td>N</td>
<td>112</td>
<td>163</td>
<td>95</td>
<td>103</td>
<td>97</td>
<td>93</td>
</tr>
<tr>
<td>Writing</td>
<td>.653</td>
<td>.449</td>
<td>1.000</td>
<td>.594</td>
<td>.584</td>
<td>.488</td>
</tr>
<tr>
<td>N</td>
<td>108</td>
<td>95</td>
<td>160</td>
<td>107</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>.639</td>
<td>.550</td>
<td>.594</td>
<td>1.000</td>
<td>.690</td>
<td>.591</td>
</tr>
<tr>
<td>N</td>
<td>152</td>
<td>103</td>
<td>107</td>
<td>207</td>
<td>142</td>
<td>135</td>
</tr>
<tr>
<td>Structures</td>
<td>.568</td>
<td>.545</td>
<td>.584</td>
<td>.690</td>
<td>1.000</td>
<td>.399</td>
</tr>
<tr>
<td>N</td>
<td>133</td>
<td>97</td>
<td>104</td>
<td>142</td>
<td>190</td>
<td>124</td>
</tr>
<tr>
<td>VSPT score</td>
<td>.520</td>
<td>.321</td>
<td>.488</td>
<td>.591</td>
<td>.399</td>
<td>1.000</td>
</tr>
<tr>
<td>(Banded)</td>
<td>172</td>
<td>93</td>
<td>104</td>
<td>135</td>
<td>124</td>
<td>247</td>
</tr>
</tbody>
</table>

On the whole, the correlations found in this study tell the same story as those found by Alderson (2005) and Alderson and Huhta (2005): the tests in different skills correlate moderately with each other, as could be expected, but not to an extent that there would any reason to doubt the usefulness of any of them in a diagnostic test such as DIALANG.

The Vocabulary Size Placement Test differs from the other DIALANG tests in many ways, so it makes sense to examine its correlations separately. Table 5 shows that the VSPT result correlated significantly with all the main language tests. Not surprisingly, the highest correlation was with the vocabulary test (.591), but, in fact, it may be somewhat surprising that the correlation was not that stronger since they are both supposed to measure vocabulary knowledge. The problems with the reliability of the VSPT scores for some learners, as explained above, may partly explain this. The correlations between the VSPT and the reading and writing tests were close to .5. They are, thus, somewhat, but not very much, lower than the inter-correlations of the five main tests. The rather low correlations with the structures and listening tests (.399 and .321, respectively) may suggest that the VSPT measures rather different type of language proficiency than these two tests.

Alderson (2005, 87) also studied the relationship between the different scoring options available for the VSPT and the other DIALANG tests. He found that the score resulting from
the present scoring mechanism used in the VSPT had the lowest correlations with the other DIALANG tests: around .40 for reading and structures, .54 for listening, and about .60 for writing and vocabulary. These are not very different from the ones reported for the VSPT in Table 5 above, especially if one bears in mind that Alderson could compute the correlations between the precise tests scores, not between CEFR levels or VSPT bands, as was done in the author’s study. Interestingly, Alderson (2005, 87) discovered that the more simple VSPT scores, in which the learner was not penalized for guessing, resulted in over .60 correlations between the VSPT and reading, listening, and structures, and in over .70 correlations with writing and vocabulary. Thus, the conclusion from the two studies above for the VSPT is that in its current format, it is not quite as closely related to performance on the other DIALANG tests but that if the scoring were changed, it appears to have the same level of correlations as the other tests. As could be expected, it is somewhat more closely related to the vocabulary test than to the other tests – the fact that it also correlates well with the writing test may be due to the fact that several items in the (indirect) writing tests used in DIALANG are in fact vocabulary items.

The author’s survey of DIALANG users also provides us with some data on the test takers’ score profiles across the five skills. To make the comparison of the findings of this study more comparable with the other studies, only those informants who had taken English tests (the most frequently chosen test language) and who had taken at least two different skills tests were included in the comparison. This restriction limited the number of test takers to a few dozen per test (reading = 55, listening = 39, writing = 41, structures = 55, and vocabulary = 44), which thus makes the results somewhat tentative.

Figure 7 shows the profile for English among the (mostly) Finnish and German university students surveyed. As can be seen, the average level of the learners is quite high, about B2 or close to C1, which would lead one to expect a fairly even performance across the skills. On the whole, this appears to be the case, although the profile is not totally the same across the skills. Writing seems to be slightly more difficult (or the learners’ writing skills lower) than the other skills for these learners, but the differences are not great. For the most part, these data appear to be in line with the findings of the DIALANG pilot test data for advanced learners (Figure 6 above) and, thus, provide evidence for the validity of the DIALANG tests as far as their standard setting and linking with the CEFR levels is concerned. The main difference between these results and those reported in Figure 6 appears to be that the writing
test seemed to be somewhat more difficult than the other tests for these learners, whereas that was not the case in the bigger pilot test data set.

![Figure 7. DIALANG English test results in the survey study by the present author](image)

c. Satakunta Polytechnic study

The third study that provides us with evidence about the correlations between different DIALANG tests is a fairly large-scale study involving Finnish polytechnic students at the Satakunta Polytechnic university (Jaatinen 2005). The Satakunta Polytechnic used DIALANG as a placement instrument to place students onto different levels of courses in both English and Swedish. At the same time, they also wanted to study the students’ proficiency in these two languages with the help of an external, international measure (more about the reasons for the study, see Section C-c and 3.4.5 below).

All new students entering the Satakunta Polytechnic in 2005 were administered the DIALANG reading, vocabulary, and structures tests in both English and Swedish (Jaatinen 2005). In addition, a range of different background information such as other language test results was gathered from the students – this will be reported on below when the relationship between DIALANG and external criteria is discussed.
On the basis of the assumptions of the equivalence of language tests developed from the CEFR and the shape of typical skills profiles of beginning vs. advanced language students, the following hypotheses can be drawn for the Satakunta study:

- The profile of proficiency in English should be more even than for Swedish, especially for the gymnasium background students because English is the first foreign language for the great majority of the students entering the Polytechnic. Typically, the students have studied it for ten years compared with 5 – 6 years of Swedish.

- The profile of the gymnasium background students should be more even than that of the vocational background students because they have studied both of the languages more (the difference may not be great, however) and, on average, they are academically better students in most subjects (this is one of the main reasons why they chose to enter gymnasium – actually, you have to apply for a place in that school).

An examination of the test results shown in Figures 8, 9, 10, and 11 indicates that:

- The profile of English for the gymnasium background students is rather even (Figure 8). The average proficiency of the learners is fairly high, around B2, and as was predicted and as befits advanced learners.

- In contrast, the same gymnasium students know Swedish at a clearly lower level (A2 on average), and their profile of mastery is rather uneven: the results of the reading test are clearly higher than the results of the vocabulary and structures tests (Figure 9). Although, the distribution of levels for the last two tests is somewhat similar, it differs in that there are clearly more A1 results in vocabulary than in structures.

- A very similar difference between the English and Swedish profiles can be found among the vocational background students, although, predictably, they do not know the two languages as well as the gymnasium students (Figure 10 and 11).

- The profiles of the gymnasium vs. vocational students do not appear to differ to any significant degree, neither for English nor for Swedish. Obviously, the level of proficiency of the gymnasium students is higher in both languages because of the reasons explained earlier. However, it appears that the difference between the two
groups is so small (about one CEFR level) that it does not cause a noticeable change in their profiles in the same language. The fact that both groups achieved B1 across the skills in English may exceed a kind of threshold after which it may be more typical for different aspects of language to be mastered at roughly the same level. This contrast with the command of Swedish in both groups, which remained at A2 or below. At these lower levels, it seems that the typical profile remains quite uneven.

To sum up the examination of the distribution of DIALANG results in the Satakunta Polytechnic study it could be argued that the results support the interpretation that the DIALANG reading, vocabulary, and structures tests are fairly equivalent in their difficulty. This is indicated by the fact that the average distribution of results across the CEFR levels is roughly the same when the students’ average proficiency is intermediate or advanced (at or above B1), as is the case for English among these learners. In contrast, the profile is rather uneven for Swedish, which was a language mastered at a clearly lower level (A1 – A2) by the same students.

**Figure 8.** DIALANG English test results of the gymnasium background students
Figure 9. DIALANG Swedish test results of the gymnasium background students

Figure 10. DIALANG English test results of the vocational background students
Figure 11. DIALANG Swedish test results of the vocational background students

d. HISBUS English proficiency testing project

Although the German HISBUS study (to be described in more detail in section C-b) only used the DIALANG reading and VSPT tests of English, it is of interest to compare the distribution of the reading scores found by Peschel et al. (2006) with the results of the English reading tests by Alderson (2005), Jaatinen (2005) and Huhta (this paper) reported above. The HISBUS study examined the English proficiency of German university students with a large, representative sample of over 3,000 students across Germany. The test results are similar to the other reported results in that the levels B1 and B2 are the most common ones (Figure 12). The proportion of B2 level results is however bigger in the HISBUS study with about 75% of the students receiving that grade. The other noticeable difference is the very small proportion of C1 results (C2 is very hard to achieve in English reading, as noted earlier) compared with the other studies but this may simply reflect the distribution of English reading ability among German university students. Overall, then, the HISBUS results
are not so different from the results obtained with intermediate level learners that they would cast doubt on the validity of the English reading tests in DIALANG, although the results are in line with the suspicion expressed earlier that this test may be comparatively more demanding than the other English tests in DIALANG.

![Reading Score Distribution](image)

**Figure 12.** The score distribution of the English reading test in the HISBUS study on German university students

**B. Internal structure of DIALANG self-assessment instruments**

Some studies were carried out by DIALANG Project members at different stages of the piloting of DIALANG that examine the quality of the self-assessment instruments used in the system. Two of them also examined the relationship between DIALANG self-assessment and the CEFR statements that they were derived from, which could be thought of as a concurrent validation study with an external criterion. Since DIALANG self-assessment are so closely based on the CEFR, drawing the line between a study of the internal structure of the SA instruments and a study of their relationship with another instrument is difficult to draw. As reported in e.g. Alderson (2005) and Huhta and Figueras (2004), DIALANG self-assessment
statements were ‘translated’ from the CEFR by changing the ‘can do’ formulation into ‘I can’, and for some statements, by slightly simplifying the statements or the terminology used in them. DIALANG claims that the self-assessment statements and the CEFR levels derived from users’ replies to these statements are equivalent to the original CEFR descriptors and levels. This claim can, and it has been studied empirically.

The first study, Kaftandjieva and Takala (2002), presented detailed quantitative evidence for validity of the DIALANG self-assessment statements – and by extension, of the CEFR scales, too. The study was based on the Finnish translations of the statements (descriptors) which were sorted by 12 Finnish as L2 teachers onto the six CEFR levels. The researchers found that the judges carried out the sorting extremely consistently, no matter which of the several possible indices of inter-rater reliability was examined (op cit. p. 112). The examination of the structure of the reading, writing, and listening scales indicated that “they can be successfully used as a framework for foreign language learning, teaching, and assessment” (op cit. p. 127) and that the relationship between the DIALANG descriptors and the original CEFR descriptors was very close when it comes to placing them on the CEFR scale (op cit. 108-110). It was also found, however, that the levels in three scales (reading, writing, listening) were not of equal length, some being wider and others narrower. The level boundaries for, say B1 and B2, were not at the same place on the three scales, either. The researchers could also identify some statements which were placed unexpectedly or on which the raters disagreed more than others.

Alderson’s (2005, 102) findings are very similar to those made by Kaftandjieva and Takala: he discovered a very high relationship between the DIALANG self-assessment statements in the pilot testing of English and the original CEFR level of the statements. Overall, the Spearman correlation was .930, and .928 for reading, .920 for writing, and .911 for listening.

Alderson (2005, 102) also studied the equivalence of DIALANG self-assessment statements across different languages. This is a rare empirical study in the sense that it not only sheds light on the quality of the self-assessment statements but also on the quality of translations. As was described earlier, considerable attention was paid to ensuring that the translations from English into the other thirteen languages would be as equivalent as possible. Alderson found out that the intercorrelations of the difficulties of the 18 core self-assessment statements in the eight most common interface languages in the DIALANG pilot tests were
very high, mostly over .9 (see table 8.3 in Alderson 2005, 102, for reading SA statements). Alderson also correlated the logit values of the Finnish and English SA statements and found strong, over .9 correlations for the reading (.901), writing (.904), and listening (.979) statements.

Finally, both Luoma (2004) and Alderson (2005) analyzed the relationship between two kinds of self-assessment used at the piloting stage of DIALANG: (1) overall SA where the learners placed themselves directly onto one of the CEFR levels after reading brief descriptions of each of the six levels (for reading, listening or writing), and (2) detailed SA where the learners responded to a set of specific ‘I can do’ statements describing concrete activities in reading, listening or writing. The learner’s CEFR level based on the latter type of SA was calculated afterwards when the statements were calibrated.

Luoma (2004, 150-151) examined the relationship between overall and detailed self-assessment with the Finnish pilot test data from over 300 learners and with a sample of the English data from about 200 learners. Her study covered only listening. Luoma found strong but far from perfect correlations between the two kinds of self-assessment: .720 for English and .826 for Finnish. Alderson’s (2005, 106) also studied the correlations between the two self-assessments but with the full English pilot data. His finding for listening (n = 534) was rather close to the correlation found by Luoma: .667. The correlations for the other two skills studied were similar in range: .628 for reading (n = 624), and .667 for writing (n = 677).

The studies reviewed above indicate that the detailed self-assessment instruments using ‘I can do’ statements are highly consistent internally, have a very close relationship with the original CEFR statements from which they were derived, and also appear consistent across the languages. It also is clear, however, that while detailed and overall self-assessment correlate with each other they are not the same. The operational version DIALANG uses only detailed self-assessment although the possibility of including both was considered at times in the Project.
C. Relationship of DIALANG tests with external measures of language proficiency

There are at least two types of external criteria against which the criterion-related or concurrent validity of DIALANG tests has been studied: learners’ self-assessments and certain other language tests or assessments that the informants have taken. Also, the performance on DIALANG tests of groups of learners who are known to differ from each other in terms of language proficiency can be used as concurrent (or predictive) validity evidence. It should also be noted that the DIALANG self-assessments are in need of external validation against other measures of language proficiency, such as tests. This evidence will be discussed below in Section D.

Studies that have used learners’ self-assessments as external criteria against which DIALANG tests can be compared (and vice versa), include at least the following: (1) Luoma (2004), Alderson (2005), and Alderson & Huhta (2005) study based on DIALANG pilot test data for English, and (2) the German HISBUS study of university students’ English proficiency.

Studies that have used other language tests or teachers’ grades as external criteria include (1) the present author’s survey of DIALANG users (reported in this paper), and (2) the Satakunta Polytechnic study into the English and Swedish proficiency of their students (Jaatinen 2005). Both these studies, as well as Alderson (2005) study, also include other external indicators of proficiency, which are perhaps less direct and precise but nevertheless potentially useful, such as information about the learners’ language studies (e.g. length) and frequency of language use.

a. Self-assessments as concurrent validity evidence for DIALANG language tests

Learners’ self-assessments are potentially a very useful external criterion against which language tests can be compared because self-assessment has access to information about the learner’s proficiency that no language test has, such as the learner’s entire experience on using the language in potentially a very wide range of contexts and for many different purposes (Huhta, 2003). The problems associated with self-assessment are also well known, however, the most serious of them probably being lack of experience in doing self-assessment (Oscarson, 1997; Ross, 1998).
Since DIALANG contains both self-assessment instruments and language tests, it is – perhaps somewhat paradoxically – possible to do external validation of DIALANG tests with the help of self-assessments that are internal to DIALANG. This is in fact what is reported in Luoma (2004), Alderson (2005) and Alderson and Huhta (2005).

The results published on the pilot testing of DIALANG tests and self-assessments are an example of concurrent validation studies of DIALANG tests with (DIALANG) self-assessments (Luoma 2004; Alderson 2005; Alderson & Huhta 2005). In fact, the tests were correlated with two types of self-assessment because the pilot tests were accompanied by both holistic and detailed self-assessment instruments. Self-assessments were found to correlate moderately with the test scores. The correlations between overall self-assessments on the 6-point CEFR scale and the English pilot test results turned out to be .544 for reading, .580 for writing, and .474 for listening (Alderson & Huhta 2005, 319). The correlation between the detailed self-assessments and the pilot test scores was .487 for reading, .550 for writing, and .495 for listening (Alderson 2005, 107).

Luoma (2004) had earlier studied the same question but with smaller samples but with two test languages, English (n = 175-223) and Finnish (n = 308-315) and in one skill, listening. In her study, the correlation between overall self-assessment and the raw listening score was .416 for English and .619 for Finnish. The correlations between detailed self-assessment of listening and the raw score from the listening test were .489 for English and .598 for Finnish (Luoma 2004, 150-151).

On the basis of his findings, Alderson (2005, 105) suggests that learners may be better at assessing their productive skills (i.e. writing in the case of DIALANG) than their comprehension, presumably because they normally get more feedback on their writing and speaking.

Yang’s (2003) study of test-takers’ reactions to DIALANG feedback provides probably the only direct empirical evidence there is about the correlation between the self-assessment and test results in based on the operational version of DIALANG. The study was very small-scale as it was qualitative in orientation (n = 12), but the researcher collected her informants’ VSPT score (exact score on the 1-1000 scale), their English reading test result and the
corresponding self-assessment of reading in English (Yang 2003, Appendix F). The present author calculated the correlations from those data and they are presented in Table 6 below.

**Table 6.** The Spearman rank order correlations between self-assessment and the English VSPT and reading test in DIALANG in Yang’s study (Yang 2003)

<table>
<thead>
<tr>
<th></th>
<th>VSPT</th>
<th>Self-assessment</th>
<th>Reading test</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSPT</td>
<td>1.000</td>
<td>.204</td>
<td>.680</td>
</tr>
<tr>
<td>sig</td>
<td></td>
<td>p = .526</td>
<td>p = .015</td>
</tr>
<tr>
<td>Self-assessment</td>
<td>1.000</td>
<td>.678</td>
<td></td>
</tr>
<tr>
<td>sig</td>
<td></td>
<td></td>
<td>p = .015</td>
</tr>
<tr>
<td>Reading test</td>
<td></td>
<td></td>
<td>1.000</td>
</tr>
</tbody>
</table>

The small number of participants obviously makes the correlations tentative at best. However, self-assessment and the reading test correlated quite substantially (.678) but the correlation between SA and the VSPT was very low and non-significant. The two language tests – VSPT and reading – correlated with each other at the same level as the SA and reading. In Yang’s study, the number of matches (4) between SA and reading test, as well as the number of overestimations (5) and underestimations (4) of proficiency was almost equal. None of the over or underestimations was greater than one CEFR level, so that on the whole, Yang’s informants seemed to be reasonably good at estimating their own proficiency.

There appears to be no other correlational studies on the correspondence between DIALANG self-assessments and DIALANG test results in the operational version of the system. The questionnaire survey by the present author of the users of the operational DIALANG did not collect such information about the informants’ self-assessed levels of proficiency that would allow correlational analyses between self-assessments and tests. (Note, however, that the survey study provided other types of empirical evidence about the relationships between DIALANG self-assessments and learner characteristics, as will be explained later in this paper.)

**b. HISBUS English proficiency testing project**

The German HISBUS study is the other empirical study that provides us with concurrent validity evidence for some DIALANG tests against learners’ self-assessed proficiency
The study did not use DIALANG self-assessments but a different SA instrument developed by the researchers for the purpose.

The study was conducted by HIS – Hochschul-Informations-System, which is a German non-profit organization that carries out educational research in higher education sector in order to provide decision-makers in Germany with empirical information about higher education and its outcomes (see http://hisbus.his.de/ for more information). The ‘BUS’ part of the word HISBUS stands for the fact that the organization and its studies typically cover a range of themes in their studies. HIS and DIALANG Project came to an agreement in 2004 that the HISBUS study of the English proficiency of German university students would use the DIALANG English reading comprehension tests as their data gathering instruments and the VSPT as the placement tool for the reading tests.

The study in question was conducted in 2005 on a representative sample of 3,700 German university students (Peschel et al. 2006). The students self-rated their English proficiency in different skills (reading / comprehension, speaking, writing, and understanding of subject-specific texts) using a simple self-assessment scale developed by the German researchers (see the explanatory text under Table 7 below). They also took the DIALANG English VSPT and reading tests, delivered by the HIS computerized system that simulated the functioning of the DIALANG system in terms of placement of students into an easy, intermediate or difficult reading test on the basis of the VSPT result. The scoring of both the VSPT and reading worked exactly as in DIALANG.

The report of the HISBUS study does not give the correlation coefficients between the students’ self-assessments and the DIALANG reading and VSPT results but the examination of the correspondence between the two sets of measures reported in Table 7, and the more detailed tables provided in (Peschel et al. 2006), indicates that the correlation must be quite substantial.

Thus, the HISBUS study provides concurrent validity evidence for the English reading and Vocabulary Size Placement Tests via the self-assessments done by the students who participated in the study (80% of the students in the sample completed all the instruments used in the study). Furthermore, the distribution of the reading and VSPT levels in this representative sample of German university students can also be broadly indicative of the
validity of these DIALANG tests, especially if the distribution makes sense in terms of what is known about the English proficiency of German students.

Table 7. German university students’ results in DIALANG English reading and VSPT tests and their self-assessments

<table>
<thead>
<tr>
<th>DIALANG reading test</th>
<th>DIALANG Vocabulary Size Placement Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test result</td>
<td>Percentage of students at this level</td>
</tr>
<tr>
<td>C2 1 1.21</td>
<td>Level 6 12</td>
</tr>
<tr>
<td>C1 4 1.23</td>
<td>Level 5 53</td>
</tr>
<tr>
<td>B2 76 1.83</td>
<td>Level 4 24</td>
</tr>
<tr>
<td>B1 17 2.57</td>
<td>Level 3 8</td>
</tr>
<tr>
<td>A2 2 2.69</td>
<td>Level 2 2</td>
</tr>
<tr>
<td>A1 0.3 2.60</td>
<td>Level 1 3</td>
</tr>
</tbody>
</table>

Average SA (self-assessment) was calculated by the HIS project by quantifying the simple categories used in the self-assessment instruments in the following way: 1 = very good (sehr gut), 2 = good (gut), 3 = satisfactory (befriedigend), 4 = sufficient (ausreichend), 5 = unsatisfactory / none (mangelhaft / keine).

Table 7 summarises the key results of the German HIS study on university students’ proficiency in English (Peschel et al. 2006, 33-36). The distribution of CEFR levels in the English reading has a clear peak at level B2, with very few students at the extreme levels A1 and C2. The test results are line with the students’ self-assessments: for example, the average self-assessment of the students who were assigned to levels B2-C2 by the DIALANG system was either ‘good’ or ‘very good’, i.e., very close to 1 or 2 on the self-assessment scale presented in Table 7 above. The self-assessment of the students at the two highest levels C1 and C2 in reading is almost identical (1.21 vs. 1.23), and differs significantly from the average SA for those at B2 (1.83), which may suggest that the English reading test in DIALANG is better at distinguishing B2 learners from C1 learners than between C1 and C2. This is in fact entirely in line with some of the findings reported in Section A above on the internal structure and score profiles across different DIALANG skills tests for English. The finding also makes sense in the light of what we know about the English reading tests and the lack of C1 and C2 level items, which means that the measurement of learners’ proficiency at the highest levels cannot be as accurate as one would hope. Thus, this comparison with an
external measure – students’ self-assessments – supports evidence about the internal structure of the reading tests gained via examining the score distributions and the \textit{a priori} information that we have about the construction of the English tests.

The overall finding about the German university students’ reading ability in English appears very plausible given what is known about the language education in Germany (university students typically come from the type of schools which aim at high levels of achievement in all subjects) and of the important role that English plays in the everyday lives of most young people in Europe. The HIS report (Peschel et al. 2006) gives every indication that the researchers responsible for the study trusted the DIALANG test results to give an accurate picture of the language proficiency of the target group.

The only point that seems slightly odd in Table 7 is the self-assessment of reading / comprehension of those who got A1 and A2 in the reading test. It is difficult to say why those who got A1 rated their reading skills slightly better than those who got A2. Perhaps these less advanced students found it difficult to assess their proficiency accurately with the help of the extremely simple scale (later, in Section D, we will present other findings that also indicate problems in beginning learners’ self-assessments). Also, as was noted in the previous section on the internal structure of DIALANG, it seems slightly surprising to find out that only about five percent of the students surveyed achieved C1 or C2 in reading.

The \textbf{results for the Vocabulary Size Placement Test} follow a distribution that is slightly skewed towards the higher end of the scale. The VSPT levels or score bands are not at all related to the CEFR, so it is not known how they relate to each other. The results of the HIS study seem to suggest that the top level of the VSPT scale might cover levels C1 and C2 on the CEFR scale, because the reading test peaks at the fourth level (i.e., B2) whereas the VSPT peaks at the fifth level. The second argument for that interpretation is the fact that the average self-assessment of the students placed at the highest VSPT level is about the same as for those placed at C1 and C2 for reading (around 1.2). An implication of this is that the English VSPT may not be able to distinguish learners at the highest levels of proficiency (C1-C2), at least when the VSPT results are reported on the 6-point scale (a more precise score on 0 – 1000 scale is also reported to the user, however).
The comparison of the DIALANG test results with the self-assessments also suggests that the VSPT scale might extend lower than the CEFR scale: the average self-assessment of those who were at A1 and A2 in reading was around 2.6 whereas for the two lowest VSPT levels self-assessment was 2.75 and 2.96, respectively. The fact that the self-assessments against which the test results are compared concern somewhat different skills – reading / comprehension in the case of the reading test and all skills in the case of the VSPT – obviously makes this reasoning somewhat uncertain. However, the shape of the distribution of the test takers at the three lowest levels in the reading test vs. the VSPT seems to support the conclusion that the lower end of the VSPT scale differs from the CEFR scale. Incidentally, the fact that more test takers were placed at the lowest VSPT level (3%) than at the second lowest (2%) is in line with the finding by Huhta (2007a) that the scoring mechanism of the VSPT places more test takers at the lowest level than it should. The issue with the VSPT scoring was known at the time of the HIS study and thus a warning against guessing was added to the VSPT instruction for the students participating in the study. Apparently not everybody paid attention to the warning as indicated by the slight rise in the number of test takers placed at the lowest level but on the other hand, the proportion of those assigned to an unduly low level would probably have been higher without the warning.

c. Satakunta Polytechnic study

The third empirical study shedding light on the concurrent, criterion-related validity of the DIALANG tests did not use self-assessments as external criteria but national examination results and teacher / school grades. The study was conducted at the Satakunta Polytechnic university in Finland and it used DIALANG as a placement instrument to place students on different levels of courses in both English and Swedish (Jaatinen, 2005). Besides placement, the institution also wanted to examine their students’ proficiency in these two languages with the help of an external, international measure (more about the reasons for the study (see Section 3.4.5 below).

All new students entering the Satakunta Polytechnic in 2005 were administered the DIALANG reading, vocabulary, and structures tests in both English and Swedish (Jaatinen 2005). In addition, a range of different background information was gathered from the students, including their marks in the two languages in either the vocational school or in the gymnasium, depending on the route they entered the Polytechnic. The gymnasium
background students also provided their Matriculation examination grades for English and Swedish.

Overall, the study largely confirmed what could be expected on the basis of the students’ background when it comes to the length and amount of language studies. On average, those who had studied the language for a shorter time got lower results and those who had studied it longer got better results. In practice this meant that the students coming from the vocational schools had significantly lower proficiency in both languages than those coming from the gymnasia, and that the results were lower for Swedish than for English in both groups. These results, then, provide very general positive evidence for the validity of these tests in the two DIALANG languages, by being able to clearly distinguish the two groups that should differ significantly in terms of their language proficiency.

Since the Satakunta Polytechnic was a fairly large-scale study with about 1,000 students and since it includes three external measures of proficiency, it also provides valuable, and more precise, concurrent validity evidence for the English and Swedish tests in DIALANG. The best of the three external criteria against which we can compare the DIALANG test results is the Matriculation examination. It is a centralized, national examination, which is professionally developed and administered in standardized conditions across the country. The foreign language test in the examination cover roughly the same areas of proficiency as DIALANG: both lack a speaking test but have a range of tests for the other main skills. Over a half of the points in the Matriculation exam come from a reading and listening comprehension tests, one third from a writing test (a composition), and somewhat over ten percent from an integrated vocabulary and structures test.

In the Satakunta study, the Matriculation examination results were available only for the students who had completed the gymnasium (about 750), as the exam is the final test only for that level and type of schooling. Jaatinen (2005) does not report what proportion of their gymnasium background students entered the Polytechnic the same year that they had taken the Matriculation examination, but typically the majority, but not all, students enter tertiary education within a year or two after completing their gymnasium studies.
Table 8. Spearman rank-order correlations between the three DIALANG tests and the Matriculation Examination tests of English and Swedish

<table>
<thead>
<tr>
<th>DIALANG test result</th>
<th>Matriculation examination grade</th>
<th>English (n = 744)</th>
<th>Swedish (n = 756)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td></td>
<td>.69</td>
<td>.57</td>
</tr>
<tr>
<td>Vocabulary</td>
<td></td>
<td>.63</td>
<td>.60</td>
</tr>
<tr>
<td>Structures</td>
<td></td>
<td>.68</td>
<td>.51</td>
</tr>
</tbody>
</table>

Table 9. Spearman rank-order correlations between the average DIALANG result and the school grades

<table>
<thead>
<tr>
<th>Overall DIALANG result (i.e. classification into 2 – 4 levels of courses)</th>
<th>English</th>
<th>Swedish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocational school grade (5 point scale)</td>
<td>.38</td>
<td>.27</td>
</tr>
<tr>
<td>p=.01</td>
<td>n = 235</td>
<td>n = 220</td>
</tr>
<tr>
<td>Gymnasium school grade (6 point scale)</td>
<td>.44</td>
<td>.61</td>
</tr>
<tr>
<td>p=.01</td>
<td>n = 699</td>
<td>n = 712</td>
</tr>
</tbody>
</table>

Table 8 (based on Jaatinen 2005, 23-24) displays the rank order correlations between the three DIALANG tests and the Matriculation examination grades. The correlations of the DIALANG tests with the Matriculation examination are higher for English than for Swedish, and are between .60 and .70 for the former and .50 - .60 for the latter. Assuming that the Matriculation exam tests are more or less equivalent in terms of quality, the difference may arise from the (probably) higher accuracy of the DIALANG English tests, which had been properly piloted and calibrated, unlike the Swedish tests, whose piloting was not completed before the Project came to the end in 2004. Consequently, the scoring system and the linking of the test scores with the CEFR levels of the Swedish tests are based on expert judgment only, which is a less reliable approach than the more empirically based calibration and linking of the English tests.

Table 9 summarizes the information on the relationship between the DIALANG test results and the two other external criteria, the vocational school grades in English and Swedish (for
the vocational background students) and the gymnasium grades in the two languages (for the gymnasium background students). The concurrent validity evidence from these comparisons is less straightforward than from the comparison with the Matriculation examination for at least two reasons. First, the DIALANG test results used in the analyses by Jaatinen (2005) are not the CEFR scale-based scores but a more indirect and more complex conversion of the results of the three DIALANG tests used in the study into levels of language courses onto which the students were placed on the basis of their overall result in the three DIALANG tests (see Jaatinen 2005, 9, for details on how the course levels were defined). Secondly, the number of course levels was very small, either two or four, which probably artificially lowers the correlations between the criteria and DIALANG.

The correlations between the school grades and the DIALANG results (course levels) were clearly higher for the gymnasium background students. The correlation between the English grade in the gymnasium and the course level assigned to the student was over .60 but the corresponding correlation for Swedish was somewhat lower (.44), which is probably explained by the restriction in the variation in the DIALANG test results, as they were categorized into only two levels (or levels of courses).

The correlations between DIALANG results and the school grades were rather modest for the vocational students, which may, at least partly, be explained by the fact that the English and Swedish studied in the vocational schools is not general language as in the gymnasium but specific purpose language. Thus, the language grades in the vocational school may reflect somewhat different kind of proficiency. Consequently, it may not be surprising that the English and Swedish grades given in the gymnasium relate more strongly with DIALANG, which is a test of general rather than specific language. Other possible reasons include differences in the reliability of teachers’ grading of students in different types of institutions but given the lack of research into this issue, this remains pure speculation.

d. Present author’s survey study of DIALANG users

The author’s survey study included two types of external assessments that allow us to study the concurrent validity of the operational DIALANG tests: the Matriculation examination grades and language teacher’s grades in the gymnasium. The main problem with the Matriculation examination as an external criterion for this study is that the time interval
between taking the two tests was often quite considerable, and that the interval varied from learner to learner. The majority of informants were students in their early twenties who had had 1 – 3 years of studies at a university or a polytechnic university, which means that about 1 – 4 years had passed since they had taken their school leaving exams. Some uncertainty in the results may also arise from the fact the Matriculation examination grades are based on norm-referencing (the top five percent get the top grade, the next 15% get the second best grade, the next 20% get the third best grade, and so on), whereas DIALANG ‘grades’ are based on absolute, criterion-referenced scoring.

The second external criterion available for some of the informants in the survey study was the teacher grade. Almost fifty 17 – 18-year-old students in a Finnish gymnasium (upper-secondary school) participated in the study and took one or more DIALANG tests in English as part of their English studies. The students reported both their DIALANG test results and their latest English grade given by their regular language teacher. Since the students were still in the middle of their upper-secondary studies, they had not yet taken the Matriculation examination and, consequently, those grades were not available for them.

As was the case with the Matriculation examination grades, there are issues also with the teacher grades as an external criterion in concurrent validation study. The teacher grades in Finland are mostly based on the student’s language skills but also other factors can be taken into account such as active participation in the lessons. The teacher is solely responsible for giving the grades, which are given on a scale ranging from 4 (failure) to 10 (excellent). Usually, teachers base their grading on tests that come with the textbooks, on tests that they design themselves, or on both. They can also make use of continuous assessment to complement the information obtained via formal tests. While the language teacher typically knows his or her students very well, after following their progress for at least 1-3 years, there is always some variation in how each teacher defines and operationalized the assessment of language proficiency, or indeed, if other factors than language skills affect the grades (see e.g., Black & Wiliam, 1998, on problems with teacher grades). Also the reliability of teacher grades may vary from teacher to teacher, as, for example, they may or may not co-operate with colleagues when assessing and giving grades.

Table 10 reports the correlations between the DIALANG tests and the two external criteria described above. The correlation between the mean of DIALANG test result (i.e., average of
all DIALANG tests taken by the learner) and the two external criteria were almost exactly the same (close to .520). When we look at the correlations for individual DIALANG skills tests, they appear to be somewhat higher for the teacher grade than for the Matriculation examination result. Perhaps the school grade reflected more closely the students’ current ability level compared with the Matriculation examination, which the learners had taken 1-4 years earlier. Since the number of students available for these analyses was so small, this difference cannot easily be interpreted as evidence for the quality, or lack of it, of either of the external criteria.

Table 10. Correlations between DIALANG tests and two external measures in the author’s survey study of DIALANG users

<table>
<thead>
<tr>
<th></th>
<th>Matriculation examination grade in the test language</th>
<th>School grade in the test language (in gymnasium)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean DIALANG test result across all skills</td>
<td>.519</td>
<td>.518</td>
</tr>
<tr>
<td></td>
<td>N 109</td>
<td>47</td>
</tr>
<tr>
<td>Reading</td>
<td>.544</td>
<td>.627</td>
</tr>
<tr>
<td></td>
<td>N 66</td>
<td>32</td>
</tr>
<tr>
<td>Listening</td>
<td>.533</td>
<td>.812</td>
</tr>
<tr>
<td></td>
<td>N 41</td>
<td>8 (p=.014)</td>
</tr>
<tr>
<td>Writing</td>
<td>.339</td>
<td>.609</td>
</tr>
<tr>
<td></td>
<td>N 38 (p=.038)</td>
<td>11 (p=.047)</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>.436</td>
<td>.762</td>
</tr>
<tr>
<td></td>
<td>N 51</td>
<td>24</td>
</tr>
<tr>
<td>Structures</td>
<td>.329</td>
<td>.701</td>
</tr>
<tr>
<td></td>
<td>N 53 (p=.016)</td>
<td>20</td>
</tr>
<tr>
<td>VSPT score (Banded)</td>
<td>.316</td>
<td>.550</td>
</tr>
<tr>
<td></td>
<td>N 82 (p=.004)</td>
<td>40</td>
</tr>
</tbody>
</table>

The significance of the correlations in Table 10 is at least .001 when not specifically mentioned.

There was an interesting tendency in the correlations between DIALANG tests and the Matriculation exam grades, however (see Table 10). The highest correlations were found for the two comprehension tests, reading and listening – their correlations were .544 and .533, respectively. The correlations with the teacher grades were even higher, .625 for reading and .812 for listening (only eight students took the listening test, so the latter correlation is only suggestive at best). Vocabulary had a somewhat lower correlation (.436) with the Matriculation examination result, followed by writing and structures (.339 and .329,
respectively). Given the slight emphasis on comprehension in the language tests in the Matriculation examination, it may not surprising to find out that the comprehension tests correlated best with it. This finding, in fact, appears to lend some credibility to claims about concurrent validity of the DIALANG reading and listening tests.

Concurrent validity evidence is available also for the Vocabulary Size Placement Test. The correlations between the VSPT results (as a band, on the six-point scale used in reporting the VSPT results to the test taker) and the Matriculation examination and school grades were both statistically significant, and roughly in the same range as the correlations reported above for the main DIALANG skills tests and the two external criteria.

D. Relationship of DIALANG self-assessments with other measures of language proficiency

Most of the concurrent validity evidence for DIALANG concerns the language tests used in the system. However, there is some evidence also for the self-assessment instruments, which are the other type of language proficiency measure implemented in DIALANG.

First, some the studies reported above in Section 3.3.2.2 – C – a, in which DIALANG tests were compared with self-assessments, also shed light on the self-assessment instruments, as long as the SA used in the studies came from DIALANG (in e.g. the HISBUS study, the SA was not based on DIALANG and cannot thus be used here).

Luoma (2004), Alderson (2005), and Alderson and Huhta (2005), are our main source of information as far as published research is concerned. They found out that the correlations between DIALANG self-assessment statements and DIALANG language tests in the pilot testing data were moderate in magnitude and statistically significant, slightly over .5 for the overall self-assessments and slightly below .5, on average, for the detailed, statement-based self-assessments in Alderson’s study. Luoma (2004), however, found for English that the detailed SA statements correlated somewhat better than overall SA, based directly on the 6-point CEFR scale. For Finnish, she found somewhat higher (around .6) correlations between SA and listening test score but there the overall SA correlated slightly better with the test than the detailed SA.
The correlations in Alderson’s (2005; Alderson & Huhta 2005) study were slightly higher for writing than for reading and listening, which may suggest that learners typically receive more feedback on their productive skills and may thus be more accurate in their assessment of these skills. The correlations suggest that the DIALANG self-assessment instruments allow learners to evaluate their foreign language proficiency in a way that is at least to some extent in line with the way that the (DIALANG) language tests measure their proficiency. Since the correlations are far from perfect, however, it is clear that self-assessments studied above involve other things, too, which we can only guess at.

Apparently the only other published studies that can shed light on the relationship between DIALANG self-assessment instruments and external measures of proficiency are the two small scale studies have been conducted by a group of French and Belgian researchers. They carried out two studies in the mid-2000s on learners’ ability to self-assess: to do this, they compared the results of the Test d'Evaluation de Français (TEF) and self-assessment of French used in DIALANG. The TEF is a test of French as a foreign language offered by the Chambre de Commerce et d’Industrie de Paris (see www.fda.ccip.fr), which reports test results on the CEFR scale of proficiency.

In the first study (Desroches et al., 2005), thirty-six Spanish-speaking test-takers took the reading and listening parts of the TEF and responded (on paper) to the operation DIALANG self-assessment statements for the two skills (18 statements per skill). The test-takers’ responses to the SA statements were converted to the CEFR scale in a straightforward way, according to the number of SA statements estimated to be at each CEFR level. Thus, since three of the DIALANG self-assessment statements are thought to be A1 statements, a student was assigned to level A1 if he/she responded ‘yes’ to 0 – 3 of the statements, to level A2, if she gave 4 – 5 positive responses (since there are two level A2 reading statements), and so on. Since the number of SA statements per CEFR level is slightly different for listening, the conversion to the CEFR levels is also slightly different. (See the DIALANG Project website at http://dialang.org/project/english/ProfInt/can_statements.htm and Appendix C in the Common European Framework for information on the CEFR levels each self-assessment statement is estimated to belong at.)
Descroches et al. (2005) found a modest but significant correlation between the test-takers’ result in the TEF listening test and the DIALANG self-assessment for listening ($r = .038$, $p = .021$), but for reading, the correlation was non-significant ($r = .05$, $p =.765$). Most of the test-takers assessed their proficiency considerably higher than what their TEF result indicated. In reading, for example, almost all of the 15 test-takers who received the lowest level (A1) on the TEF test self-assessed themselves as high as B2 and C1. A similar tendency was found in listening: of the nine test-takers placed on A1 in the TEF listening test, one had self-assessed him/herself at A2, one at B1 and seven at B2. The self-assessments and test results of the intermediate and advanced learners did not differ as much as the beginners’, which indicates that the beginners’ gross overestimations were the source for the low correlations between SA and test results.

The second study (Demeuse et al., 2006) was based on the same TEF tests but all the 36 reading and 44 listening SA statements developed for the piloting of DIALANG were used in this study (see the DIALANG Project website for the full list). This time a group of 18 learners that included native speakers of different languages (Spanish, Portuguese, English, and German) was studied. Again, the correlation was higher for listening ($r = .30$) than for reading ($r = .014$), but this time neither of them was statistically significant.

Since both studies were carried out with relative small numbers of learners – especially the latter study – the conclusions from these studies are somewhat uncertain. It is not easy, for example, to say what they tell us about the validity of DIALANG self-assessment statements, except that the fact that the researchers chose to use DIALANG self-assessment statements indicates that they valued these as potentially valid instruments for the purpose. The different conversion of SA statements into CEFR levels in these studies may also have affected the results (in DIALANG the conversion is based on estimating the learner’s self-assessed proficiency in terms of a value on an IRT ability scale, which is then converted into a CEFR level by applying standard setting based cut-off points). Perhaps the most interesting conclusion from the two studies is that they show, once again, how difficult self-assessment can be for (apparently) untrained learners. They also suggest that self-assessment can be particularly difficult for learners at lower levels of proficiency.

Since so little research has been carried out on how DIALANG self-assessments relate to other measures of proficiency, the present authors carried out more detailed analyses on the
DIALANG English pilot data. The purpose of these analyses is to examine the correlations between SA and test results that are reported in Alderson (2005) and Alderson & Huhta (2005). Of particular interest is to try to identify factors that are related to the match or mismatch between self-assessment and test results. The factors studied here are the level of language proficiency, mother tongue, and level of education. Mother tongue is a proxy for a culture and educational system that differ between countries and may affect how members of a particular culture and/or educational system approach self-assessment (e.g. familiarity with it, tendency to be confident vs. modest of one’s skills). Logically, these analyses might also be presented under the topic of internal structure of self-assessment instruments; however, the line between internal and external validity evidence is difficult to draw in this case, and thus the findings can be discussed here without digressing too far from the main line of presentation.

Tables 11 A, B, and C below display the crosstabulations between the DIALANG English pilot test results (converted into CEFR levels) and self-assessments (with the help of the detailed ‘I can do’ statements, also converted into CEFR levels). The tables allow us to examine, for example, if the learners at lower levels of proficiency – according to the test results – tend to be more accurate in their self-assessments than more advanced learners, or vice versa.

Overall, the learners tended to overestimate their reading and listening skills but slightly underestimate their writing skills (as was mentioned above, Alderson 2005 found that the correspondence between SA and test result was best for writing). In reading, the vast majority of learners at A1 to B1 overestimated their skill, often by two or more levels. The match was best at around B2, whereas the C1-C2 readers tended to slightly underestimate their skill.

For listening, the picture is somewhat different. Here, too, the beginners (A1-A2) tend to seriously overestimate their skill but there is an equally clear tendency for the advanced C1-C2 learners to underestimate how well they can listen to English. Intermediate B1-B2 learners mostly seemed to have a more accurate idea of their listening abilities than the others.

For writing, the pattern is somewhat similar to listening. Again, the beginners tend to, for the most part, overestimate their skills. Interestingly, the vast majority of B1 learners self-
assessed their writing ability at that very same level. B2 learners, however, underestimated their writing by exactly one level. The more advanced learners, too, underestimated their writing skills but more seriously: the underestimation was at least two CEFR levels in most cases.

**Table 11 A.** Crosstabulation of test results and self-assessments for reading in the English pilot data

<table>
<thead>
<tr>
<th>Self-assessed level (Reading, detailed SA)</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading test result</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>2</td>
<td>12</td>
<td>17</td>
<td>10</td>
<td>4</td>
<td>4</td>
<td>49</td>
</tr>
<tr>
<td>A2</td>
<td>5</td>
<td>21</td>
<td>22</td>
<td>38</td>
<td>7</td>
<td>0</td>
<td>93</td>
</tr>
<tr>
<td>B1</td>
<td>3</td>
<td>16</td>
<td>42</td>
<td>104</td>
<td>22</td>
<td>2</td>
<td>189</td>
</tr>
<tr>
<td>B2</td>
<td>1</td>
<td>8</td>
<td>11</td>
<td>90</td>
<td>39</td>
<td>11</td>
<td>160</td>
</tr>
<tr>
<td>C1</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>63</td>
<td>44</td>
<td>13</td>
<td>125</td>
</tr>
<tr>
<td>C2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Total # of learners</td>
<td>11</td>
<td>59</td>
<td>95</td>
<td>305</td>
<td>121</td>
<td>33</td>
<td>624</td>
</tr>
</tbody>
</table>

**Table 11 B.** Crosstabulation of test results and self-assessments for listening in the English pilot data

<table>
<thead>
<tr>
<th>Self-assessed level (Listening, detailed SA)</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening test result</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>7</td>
<td>24</td>
<td>28</td>
<td>12</td>
<td>1</td>
<td>4</td>
<td>76</td>
</tr>
<tr>
<td>A2</td>
<td>4</td>
<td>14</td>
<td>23</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>57</td>
</tr>
<tr>
<td>B1</td>
<td>2</td>
<td>15</td>
<td>63</td>
<td>39</td>
<td>13</td>
<td>4</td>
<td>136</td>
</tr>
<tr>
<td>B2</td>
<td>0</td>
<td>3</td>
<td>46</td>
<td>63</td>
<td>23</td>
<td>10</td>
<td>145</td>
</tr>
<tr>
<td>C1</td>
<td>0</td>
<td>2</td>
<td>15</td>
<td>40</td>
<td>19</td>
<td>9</td>
<td>85</td>
</tr>
<tr>
<td>C2</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>10</td>
<td>12</td>
<td>3</td>
<td>35</td>
</tr>
<tr>
<td>Total # of learners</td>
<td>13</td>
<td>58</td>
<td>185</td>
<td>177</td>
<td>70</td>
<td>31</td>
<td>534</td>
</tr>
</tbody>
</table>
### Table 11 C. Crosstabulation of test results and self-assessments for writing in the English pilot data

<table>
<thead>
<tr>
<th>Self-assessed level (Writing, detailed SA)</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Writing test result</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>9</td>
<td>25</td>
<td>23</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>61</td>
</tr>
<tr>
<td>A2</td>
<td>4</td>
<td>23</td>
<td>46</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>77</td>
</tr>
<tr>
<td>B1</td>
<td>2</td>
<td>16</td>
<td>143</td>
<td>8</td>
<td>11</td>
<td>8</td>
<td>188</td>
</tr>
<tr>
<td>B2</td>
<td>1</td>
<td>5</td>
<td>130</td>
<td>11</td>
<td>30</td>
<td>8</td>
<td>185</td>
</tr>
<tr>
<td>C1</td>
<td>1</td>
<td>0</td>
<td>55</td>
<td>10</td>
<td>34</td>
<td>7</td>
<td>107</td>
</tr>
<tr>
<td>C2</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>13</td>
<td>5</td>
<td>26</td>
</tr>
<tr>
<td><strong>Total # of learners</strong></td>
<td>17</td>
<td>69</td>
<td>403</td>
<td>32</td>
<td>89</td>
<td>34</td>
<td>644</td>
</tr>
</tbody>
</table>

The second factor whose influence on the match between self-assessment and test results is studied here is the **mother tongue** of the pilot test takers – which stands for their background in terms of **culture** and **educational system**. Obviously, the first language is a somewhat uncertain proxy for culture, because not all speakers of the same language share the same culture. The match between a language and a culture varies depending on the language. For example, the vast majority of those speaking Finnish as L1 probably live and are educated in Finland. However, the same does not apply to languages such as French, German, and Spanish, although the information about the location of the pilot test sites suggests that the majority of the test takers with those L1s came from France, Germany, and Spain, respectively.

Table 12 displays rank order correlations between self-assessment and test result broken down by the mother tongue of the test takers; statistically non-significant correlation coefficients are in brackets. The table also shows the mean level of proficiency in the three tested skills (R = reading, L = listening, W = writing) in terms of the 6-point CEFR level. This enables us to see if the test takers from a particular country / language were clearly different from the others in terms of proficiency (e.g. we can see that, on average, the speakers of Icelandic tended to be more proficient than the other pilot test takers; the mean of 4 equals B2). Languages with a total of more than 70 pilot test takers are included as separate entries in the table. The ‘other D languages’ refers to those languages of the DIALANG
project that had fewer than 70 test takers and which are here grouped together. ‘Other languages’ means that the L1 of the test taker was not among the 14 DIALANG languages.

Table 13 shows the mean test results and self-assessments on the 6-point CEFR scale, thus a mean close to three indicates level B1 and a four indicates level B2. The comparisons suggest that the overestimation of reading happened in all language groups, and that sometimes the average overestimation was close to one CEFR level. In contrast, the general overestimation in listening was due to clear overestimation in certain groups only (the other languages, Spanish, and to an extent Danish), and there were three groups whose members actually underestimated their listening slightly. It was noted above that in general, writing was underestimated, and that in particular the B2-C2 learners underestimated their writing ability (in English). Seen from the first language perspective, the ones who underestimated most their writing skills were the Icelandic, Norwegian, and Dutch speakers.

Since the number of test takers is in many cases rather small, it is impossible to draw firm conclusions from the results displayed in Table 12 and 13. However, the overall tendency seems to be that these pilot data do not support the hypothesis that learners coming from different cultures and language backgrounds differ in terms of their ability to do self-assessment. Most of the correlations do not vary very much across the language groups, at least for writing and reading. Listening has more variation, however. The potentially interesting findings in Tables 12 and 13 that might deserve further study in the future include the following:

- Listening appeared to be the most difficult skill to self-assess, at least for some groups;
- Speakers of Finnish and Spanish (expect for reading) had the best match between SA and test results;
- Speakers of Dutch found it difficult to self-assess listening, and the Spanish found it difficult to self-assess reading (in contrast to the other skills);
- Speakers of Spanish tended to overestimate their skills across the board but only in reading did this result in a significant drop between the rank order correlation between the test and self-assessment.
Table 12. Rank order correlations between test results and self-assessments in the English pilot test data across different first language groups

<table>
<thead>
<tr>
<th>Mother tongue</th>
<th>mean</th>
<th>st. deviation</th>
<th>Correlation between test result and self-assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>L</td>
<td>W</td>
</tr>
<tr>
<td>Danish</td>
<td>3.5</td>
<td>3.6</td>
<td>4.1</td>
</tr>
<tr>
<td>Dutch</td>
<td>3.6</td>
<td>3.8</td>
<td>3.7</td>
</tr>
<tr>
<td>Finnish</td>
<td>3.2</td>
<td>3.4</td>
<td>2.9</td>
</tr>
<tr>
<td>French</td>
<td>3.1</td>
<td>2.9</td>
<td>3.1</td>
</tr>
<tr>
<td>German</td>
<td>3.8</td>
<td>3.9</td>
<td>3.9</td>
</tr>
<tr>
<td>Icelandic</td>
<td>4.0</td>
<td>4.4</td>
<td>4.1</td>
</tr>
<tr>
<td>Norwegian</td>
<td>3.9</td>
<td>3.6</td>
<td>4.0</td>
</tr>
<tr>
<td>Spanish</td>
<td>2.7</td>
<td>2.4</td>
<td>3.0</td>
</tr>
<tr>
<td>Other D language</td>
<td>3.5</td>
<td>3.3</td>
<td>3.5</td>
</tr>
<tr>
<td>Other language</td>
<td>3.0</td>
<td>26</td>
<td>2.8</td>
</tr>
</tbody>
</table>

Table 13. Comparison of the CEFR level of the test results and self-assessments across different language groups

<table>
<thead>
<tr>
<th>Mother tongue</th>
<th>Reading</th>
<th>Listening</th>
<th>Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test</td>
<td>SA</td>
<td>Test</td>
</tr>
<tr>
<td>Danish</td>
<td>3.5</td>
<td>4.3</td>
<td>3.6</td>
</tr>
<tr>
<td>Dutch</td>
<td>3.6</td>
<td>3.9</td>
<td>3.8</td>
</tr>
<tr>
<td>Finnish</td>
<td>3.2</td>
<td>3.4</td>
<td>3.4</td>
</tr>
<tr>
<td>French</td>
<td>3.1</td>
<td>3.8</td>
<td>2.9</td>
</tr>
<tr>
<td>German</td>
<td>3.8</td>
<td>4.2</td>
<td>3.9</td>
</tr>
<tr>
<td>Icelandic</td>
<td>4.2</td>
<td>3.9</td>
<td>4.4</td>
</tr>
<tr>
<td>Norwegian</td>
<td>3.9</td>
<td>4.1</td>
<td>3.6</td>
</tr>
<tr>
<td>Spanish</td>
<td>2.7</td>
<td>3.8</td>
<td>2.4</td>
</tr>
<tr>
<td>Other D language</td>
<td>3.5</td>
<td>3.7</td>
<td>3.3</td>
</tr>
<tr>
<td>Other language</td>
<td>3.0</td>
<td>4.0</td>
<td>2.6</td>
</tr>
</tbody>
</table>

Thirdly, the effect of the educational background to the accuracy of self-assessment of studied. The pilot test takers were divided into groups: one comprised those who had reported that they had either university or non-university higher education, the other included all the others (primary, secondary, vocational, and other types of education). As Table 14 indicates,
the higher education background learners had clearly higher level of proficiency in all three skills, which may not be very surprising. The match between self-assessment and tested ability is slightly better for the more highly educated learners of English, as far as reading and writing are concerned. However, the reverse is true for listening. None of the differences is very big so the conclusions must be very tentative. Perhaps the fact that reading and writing in English is very important in higher education, in particular, might mean that learners at that level of education have a slightly more accurate idea of how well they can read and write in English. This hypothesis needs to be studied further in other contexts and with other datasets, however.

**Table 14.** Correlations between test results and self-assessments in the English pilot test data for learners with higher education vs. non-higher level of education

<table>
<thead>
<tr>
<th>Education</th>
<th>mean</th>
<th>st. deviation</th>
<th>Correlation between test result and self-assessment</th>
<th>Reading</th>
<th>Listening</th>
<th>Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>L</td>
<td>W</td>
<td>R</td>
<td>L</td>
<td>W</td>
</tr>
<tr>
<td>Higher</td>
<td>3.6</td>
<td>3.5</td>
<td>3.7</td>
<td>1.2</td>
<td>1.4</td>
<td>1.3</td>
</tr>
<tr>
<td>Lower</td>
<td>3.1</td>
<td>3.2</td>
<td>3.0</td>
<td>1.2</td>
<td>1.4</td>
<td>1.2</td>
</tr>
</tbody>
</table>

### 3.3.2.3 Learners’ reactions to DIALANG, its tests, self-assessment instruments and feedback

Learners’ reactions to DIALANG form the second major type of empirical evidence for the construct validity, authenticity, and interactivity of DIALANG. These are presented in the articles by the present author on the VSPT (Huhta 2007a), self-assessment (Huhta 2007b; Huhta, submitted), and in Appendix 1 of the present document. In addition to these studies, there has been very little systematic research into users’ attitudes or reactions to DIALANG. The only exceptions to this seem to be the studies by two MA students at Lancaster University (Floropoulou 2002; Yang 2003). In the following, the main findings of all these articles and studies will be summarised.

Traditionally, research based on test takers’ views and reactions has been classified as study into the face validity of a test, and it has been considered less important than other types of validation. That view can be – and has been – challenged on different grounds. Also, what
has been called face validity can in fact mean quite different things. From an ethical point of view, it has been argued that it is only fair to give the test takers a voice in, e.g., test design, as they are important stakeholders in the testing process and are affected by the test results (see e.g., McNamara & Roever 2006). Asking test takers’ opinions is not only fair, however, because it can provide insights into the qualities and problems in assessment instruments that could otherwise remain undetected (e.g. think-aloud studies; see e.g., Green 1998; Banerjee 2004). Gathering comments from users seems to be quite common in the piloting phases of language tests – and in the piloting of computer software, too. Whatever information is collected from test takers can be considered as one kind of validity evidence. It is good practice in testing to use as many different types of evidence to triangulate the quality of assessment procedures to see where the evidence converges and where it does not, so not using input from learners would in fact be quite odd in test development and validation nowadays.

A. Present author’s survey study of DIALANG users

The following sections summarize the results of the present writer’s survey of over 550 users of DIALANG which have been reported in Huhta 2007a, (on the VSPT), Huhta 2007b (on self-assessments), and in Appendix 1 to this paper.

a. Users’ overall reactions to DIALANG

Appendix 1 gives an account of 557 users’ responses to several open-ended questions concerning the quality of DIALANG as a whole and the quality of its feedback. The content of the responses was analysed and categorised with the help of Atlas-ti programme designed for analysing qualitative data, and the findings described in Appendix 1 are summarised below.

Although the learners studied were not randomly sampled, they nevertheless included a substantial number of students who were not taking DIALANG voluntarily but as part of their course requirement. The sample thus included both those who had chosen to try out DIALANG because of whatever advance information they had been given by their teacher or institution, as well as those who may not have bothered to try it out had they not been compelled to do so in order to pass their language course. Overall, then, it could be argued
that the informants of the study probably represent a range of typical DIALANG users, at least in Finland and probably also in Germany, which are among the countries where DIALANG has been most popular. With this in mind, it is obviously encouraging to see that the overall trend in the users’ responses: positive views on DIALANG clearly outnumber negative views and that the list of different positive features of the system is longer than the list of problems (see Appendix 1).

The most often mentioned benefit of DIALANG by these users was that it gives them a chance to see what their proficiency level is, or as many informants put it, to see how good their language skills are. Many learners valued the possibility of seeing where their problems were; the immediate feedback on individual items was thought to be particularly useful for this. The multi-sidedness and comprehensiveness of the system were also appreciated (see Appendix 1 for details). One gets the impression from these comments that, by and large, the system appears to have achieved one of its main aims, namely, that it provides feedback to its users on their level of proficiency, as well as on the strengths and weaknesses in their proficiency.

The most often mentioned specific problems with DIALANG concerned the Vocabulary Size Placement Test and a range of technical problems (Appendix 1). Huhta (2007a) elaborates on the issues concerning the VSPT, which seemed to divide users’ opinions more than any other part of the system. As described above, the VSPT problems were a combination of issues with reliability, validity, and interactiveness but there were also users who felt the test content (i.e., isolated words) was problematic, thus implying some authenticity issues, as well. Other problems mentioned by a substantial number of users concerned two major aspects of the system, namely test content and feedback. Since both are so extensive it is not surprising to see them among the top five problem areas in DIALANG, as different people would find fault with a certain, but different, aspect of test content or feedback. What stood out in these complaints were the limited range of different test items (or lack of different test versions) and the too general and impersonal feedback provided by the system. None of the above criticisms came as a surprise because similar feedback had already been received from users over the years preceding the study. However, the fact that it converges with previous information from mainly teachers and institutions, gives these learner data added credibility. It also confirms that while DIALANG has taken steps to provide as detailed and personal...
feedback as possible, there is still quite some way to go in developing truly diagnostic foreign language assessment systems that give highly personalized feedback.

b. Learners’ reactions to the VSPT

The problems with the reliability of the VSPT scores were already discussed above, so it is not repeated here. However, the learners also had a lot to say about the validity of the test (see Huhta 2007a for details), when they were specifically asked about the VSPT in the author’s survey (the questionnaire had sets of questions for all the different parts and types of feedback provided by DIALANG). Two inter-linked themes stood out in their comments. First, the VSPT divided people’s opinions, even if we disregard the criticism caused by the over-sensitive scoring mechanism. The informants’ views on the content and format of the VSPT also differed: Some learners thought the VSPT was the best part of DIALANG while others found the scope of proficiency it was testing insufficient.

The second theme that can be discerned underlying many comments was the unfamiliarity of the VSPT as a test method and lack of knowledge about its function as a placement instrument in DIALANG. For example, it is difficult to say if the learners criticizing the VSPT for focusing only on isolated words did not know that its main role was to serve as a quick placement tool or whether they were questioning the validity of predicting a person’s general proficiency from such a limited sample of language. A few comments also suggested that the VSPT might function somewhat differently across different languages and proficiency levels, which would undoubtedly deserve to be studied more. Interestingly, analyses carried out by the Project during the development of the VSPT indicated that there were some differences between the two language versions of the VSPT that were studied, Finnish and English (Luoma & Jäppilä 2001). The discrimination of the real words used in the Finnish test was higher than the discrimination of the pseudo-words, but the reverse was true for the English test. Also, the relationship between the piloted 150-item versions of the VSPT and the shortened, 75-item operational version was somewhat different in the two languages.

It appears that while the construct validity of the VSPT (i.e., knowledge of words) was questioned by some test takers, the interaction between the VSPT as a test task and the language ability – and possibly other competences – made the VSPT stand out from the other
DIALANG tests. Some users’ comments appeared to focus on precisely the interactiveness of the VSPT. Especially the users’ reports on being engaged in metacognitive activities when taking the VSPT (Huhta 2007a, 53) seem to indicate that specific type of interaction took place between these learners’ cognitive system and the VSPT. For the most part, such comments were positive rather than negative. When we remember that the VSPT was the part of DIALANG that attracted by far the most enthusiastic comments from some test takers, one may wonder if the reason for such rave reviews was the VSPT’s unique type of interaction between the task and the cognitive system of the learner. That unusual interaction may also explain why other learners found the VSPT so off-putting: To engage in an activity that was so different from taking other language tests may have been an adequate reason to become suspicious about it.

The VSPT deserves more studies, but in the short term, it is important to change the scoring algorithm of the present version of DIALANG. For example, the current scoring mechanism causes many problems for the gathering of fair and reliable data on not only on learners’ reactions to the test itself but also on the scores users receive from it, and possibly from the main DIALANG tests, if the placement procedure causes them to be given inappropriately easy or difficult language tests.

c. Learners’ reactions to self-assessment and self-assessment feedback

Self-assessment is an essential element of DIALANG and two of the empirical articles that are part of the present study (Huhta 2007b, 2008) focus on it. The two articles provide us with somewhat different kinds of evidence of the construct validity of self-assessment.

The first article, Huhta (2007b), is based on the author’s survey of over 550 users of DIALANG, and it reports on their reactions to the two types of feedback on self-assessment of language proficiency, i.e., a comparison of learners’ self-assessments and test results, and information about possible reasons for a mismatch between the two. In the study, the learners replied to questions about their use and perceived interest and usefulness of the feedback.

The article reports on how users’ reactions to the self-assessment feedback compared with their views on other types of feedback. It turned out that feedback on the match between self-assessment and test result was considered one of the most interesting and useful types of
feedback on offer in DIALANG. Only the traditional types of test feedback, i.e. the test result and (immediate) feedback on individual test items, were regarded better in this respect. The latter finding is in line with the informants’ free responses, according to which the best aspects of DIALANG included the chance to know one’s proficiency level and the immediate item feedback (see Appendix 1). The fact that self-assessment feedback was received quite positively is obviously encouraging considering that DIALANG was aimed to develop learners’ self-assessment skills, which are important for independent and life-long study of languages.

The survey study found out that both major national groups of learners included in the study, Finns and Germans, shared a positive attitude to self-assessment of language proficiency. However, attitudes to the usefulness of doing self-assessment and getting feedback on it may vary depending on the background and educational system of the learners, as was found out in an earlier, small scale study by Floropoulou (2002b), which will be described later in this paper.

In contrast to the findings of the survey study concerning feedback on the match between self-assessment and test result, learners’ reactions to the information about the reasons why self-assessment and test result may not match were more mixed. Overall, it was considered the least interesting and useful part of feedback, which may not be surprising as it is less relevant to many test takers: only those whose self-assessment and test result do not match may be expected, on average, to spend much time studying it. It is also one of the least personalised parts of feedback in DIALANG. However, there were also users who found it interesting and useful to dwell on the information found in this feedback.

The informants’ unfamiliarity with feedback on self-assessment (both types) was a likely reason for the finding that these two types of feedback were the ones of whose usefulness the learners felt the least sure about (Huhta 2007b, 381). As they had never before received such feedback, they were often somewhat uncertain in their opinion, irrespective of whether the opinion was positive or negative.

Huhta’s (2007b) study also examined whether the users’ background (e.g., age, gender, proficiency) was related to their perceived interest in and usefulness of self-assessment feedback. Did learners of a particular age, gender, or level of proficiency regard SA feedback
more vs. less useful and interesting? The findings are clearest when it comes to the gender: women tended to consider SA feedback more useful and interesting than men. There was a tendency for the youngest learners (under 18-year-olds) to react to the SA feedback more negatively than the (somewhat) older learners but the rather small number of younger learners in the data makes this finding somewhat uncertain. Level of language proficiency or experience in learning languages was not related to the way these informants thought about the SA vs. test result feedback. However, the less proficient and less experienced learners tended to regard the more general information about SA and language proficiency as more useful than the more experienced or more proficient learners.

In addition to the information on DIALANG users’ self-assessment reported above, the users’ reply to an overall question concerning the test result (CEFR level) is worth mentioning here because it is related to the informants’ idea of their proficiency. The survey questionnaire asked the informants to say, in their own words, if they got the results in the DIALANG test(s) that they had expected. 349 of the 557 informants replied to the question (one questionnaire version lacked this question). Of those who replied, 219 said the result(s) they received matched their expectations, 50 said their results were worse and 31 that they were better than expected. Another 31 either simply stated they did not get the expected results or that the results varied, and 18 replied something else, for example, that they did not expect anything particular. Overall, thus, two thirds of those who replied stated that the DIALANG test results had matched their expectations and one third that they had not or that the results were mixed.

The second empirical article on self-assessment (Huhta, submitted) is in a very different kind of study. It aims to increase our understanding of SA of language proficiency by examining the advantages of the self-assessment in DIALANG. What are the advantages of self-assessment used in DIALANG? How do certain groups, such as language education experts, structure the concepts related to these advantages? The study also explores the use of concept mapping and multidimensional scaling in the study of such concepts, with the aim of producing a perceptual map that displays how the dozens of different advantages of SA cluster in relation to each other.

Based on a sorting exercise by 15 language education specialists, it appears that four main clusters account for most of the advantages of DIALANG self-assessment. According to the
importance judgments by the 15 informants, the most important of these was the fact that the system combines self-assessment and tests and that the learners have a chance to compare their SA and test results. The second major cluster turned out to relate to the general quality of SA in DIALANG, the SA statements, and the range of languages available. The third concerned the awareness raising effects of the SA, and the fact that it is based on considerable, international expertise. Finally, the fourth cluster was a chain of inter-linked concepts related to independence, motivation, novelty, computerization, and anonymity of DIALANG self-assessment. It is not possible to fully compare the advantages ascribed to DIALANG self-assessment by language learners and teachers because of the nature of the study, but the available data suggests that the teachers considered SA to have more varied types of advantages to learning than what the learners considered. This is based on comparing the advantages of DIALANG self-assessment mentioned by the teachers and learners from whom the list of advantages was solicited before the teachers (i.e. language education experts) conducted the sorting exercise. The finding is perhaps not entirely surprising, as teachers should be experts in language education who should have a more detailed and comprehensive knowledge about self-assessment. The benefits of SA gathered from language learners were mostly related to the first two clusters, that is, the comparison of SA and test result and to the general quality of SA in DIALANG.

In addition to the article on SA-related concept mapping (Huhta, submitted), the present author has also studied whether ordinary language learners conceptualise the advantages of DIALANG self-assessment differently from the language education experts. A group of eleven students majoring in non-language subjects at a university or other type of institution sorted the same advantages as the teachers. The map that resulted from the analysis of the learners’ sorting data is shown in Appendix 3. As can be seen, the non-language experts studied here did indeed sort the advantages somewhat differently from the experts but that, overall, the clusters are not entirely different. The first cluster is the same as with language teachers: the chance to compare SA and test result. The second cluster combined statements referring to the international, expert-based nature of the system and the fact that it was motivating (this is somewhat similar to teachers’ cluster three but learners did not include awareness raising in this cluster). The third cluster related to awareness and reflection and to the fact that language proficiency consists of different areas, so this is similar to the other half in teachers’ cluster three. However, learners included statements concerning the quality of the SA statements in this cluster. The fourth cluster was about the general high quality of self-
assessment in DIALANG (easy to respond, clarity, and also the choice of language in which to reply). The fifth major cluster in learners’ classification was very similar to teachers’ cluster four: a set of interlinked themes such as independence flexibility and anonymity. In contrast to the teachers, these learners place motivation into a different cluster – cluster two, as mentioned above, thus connecting it with the international, expert-based nature of the system.

The concept mapping studies provide evidence about the construct validity and perhaps also about the authenticity of DIALANG self-assessment, at a fairly general level. The main clusters identified in the study suggest a high quality of both the overall DIALANG system and the self-assessments (teachers’ cluster 2, but also some aspects of cluster 3 and 4). Also, the fact that some of the clusters (teachers’ cluster 1 and partly 3) relate to some of the purposes of DIALANG provides support to the arguments that the system indeed meets those aims – particularly that the self-assessment applied in the system raises learners’ awareness. As far as the comparison of the experts’ (teachers) and non-experts’ (learners) classification is concerned, the overall results were rather similar: the clusters identified by both groups were often the same, although there were some interesting differences (e.g. the different association of statements relating to motivation) that may be worth studying in more detail in the future.

d. Learners’ reactions to other DIALANG feedback

This last section on learners’ reactions to DIALANG focuses on the different types of feedback that the system offers to its users. The three preceding sections have already covered feedback either from a very general point of view (users’ overall reactions to DIALANG) or from a very specific angle (feedback related to the VSPT and self-assessment). Since feedback is such an important element of DIALANG, it makes sense to present the main results of the present author’s survey of the 557 learners as far as their reactions to the eight main parts of feedback are concerned. This thus complements the previous three sections and the two articles (Huhta 2007a, 2007b) that are based on the same survey but cover the VSPT and self-assessment only.

Huhta (2007b) includes two tables that are reproduced here, as they summarize in the form of a graph the users’ view of the eight types of feedback in DIALANG (Figure 13 and 14
below). The survey questionnaire included three Likert-scale questions that charted the learners’ opinions of the feedback. For each type of feedback these three questions were asked:

*Do you usually read this part of DIALANG feedback?*

*How interesting is this part of feedback in your view?*

*How useful is this part of feedback for you?*

The response options for the first question were: *never – sometimes – usually always*, and for the other two they were: *very – somewhat – only a little – not at all* (interesting / useful). In addition, there was a ‘cannot say’ option in the last two questions. Respondents’ opinions were also tapped with open-ended questions; these were analysed in section ‘a’ above on users’ overall reactions to DIALANG and they are more fully elaborated in Appendix 1.

Figure 13 provides an *overall picture* of the users’ *reactions to the eight types of feedback* in DIALANG, and it combines the users’ responses to the last two questions (interest and usefulness). ‘Cannot say’ responses are excluded from the figure, and the vertical numerical scale on the left is based on quantifying the responses onto a 0 – 3 scale where ‘not at all’ equals zero and ‘very’ equals three. The figure shows that in the users’ view the best part of feedback was the traditional overall result of the test (bar number 1), which is called ‘Your level’ in DIALANG but which means the same as the grade, mark or total score in other tests. On average, its rating was over 2.5, which is close to the highest possible scale value ‘very’ interesting / useful. The ratings for all the other parts of feedback were not as high but the average ratings for almost all of them were close to 2, which corresponds to the option ‘somewhat’ interesting or useful on the scale.

Figure 14 tells basically the same story as Figure 13 but it displays the distribution of users responses over the response options (except ‘cannot say’, which is excluded here). The main additional information that it provides compared with the other figure is that with only two exceptions the most common reaction to the different types of feedback was that it was ‘somewhat’ interesting or useful. The exceptions are the main test result, which received more ‘very interesting’ and ‘very useful’ ratings than any other part of feedback, and the information about self-assessment (reasons for possible mismatch) where somewhat negative ‘only a little’ type of ratings were the most common response. (In Figure 14, very positive
ratings are those where the mean of the ratings of interest and usefulness are close to three (i.e., ‘very’), somewhat positive ratings are close to two (‘somewhat’), somewhat negative ones close to one (‘only a little’), and very negative ones close to zero (‘not at all’).

**Figure 13.** Summary of the learners’ overall reaction to the different types of feedback in DIALANG (ratings of interest and usefulness combined in each bar)

Table 15 below displays a detailed breakdown of users’ responses to the three key Likert-scale questions on the use (i.e. reading) of the different types of feedback, and on their attitude towards them in terms of perceived interest and usefulness of the different feedback.
In the main, the table presents in the form of numbers the information displayed in Figure 13 and 14 above – at least as far as the ratings of interest and usefulness are concerned. The table however gives more information. First, it shows the users’ responses to the question about whether they actually read the feedback; this is likely to give us an idea of what feedback users of DIALANG typically pay attention to and what they do not read so often.

**Table 15.** Distribution of users’ responses to the three Likert scale questions on reading different feedback and on their perceived interest and usefulness

<table>
<thead>
<tr>
<th>Read in general?</th>
<th>How interesting?</th>
<th>How useful?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Never</td>
<td>Sometimes</td>
</tr>
<tr>
<td>Test result (CEFR level)</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Immediate item review / feedback</td>
<td>18</td>
<td>21</td>
</tr>
<tr>
<td>Post-test item review / feedback</td>
<td>13</td>
<td>36</td>
</tr>
<tr>
<td>Self-assessment feedback (SA/ test result match)</td>
<td>7</td>
<td>27</td>
</tr>
<tr>
<td>Information about self-assessment</td>
<td>25</td>
<td>58</td>
</tr>
<tr>
<td>Extended level descriptions</td>
<td>17</td>
<td>54</td>
</tr>
<tr>
<td>Advice on improving proficiency</td>
<td>20</td>
<td>43</td>
</tr>
<tr>
<td>Placement test result</td>
<td>8</td>
<td>27</td>
</tr>
</tbody>
</table>

The informants’ responses to the question on reading the feedback give a partly different picture than the responses to the questions on interest and usefulness. Again, the test result dominates as the most often read part of feedback – and it is also rated as the most interesting and useful one, as was described above. However, the two other types of feedback that resemble the test result in that they are brief and focus on one or two overall type of results were also used very frequently, according to the informants, namely the feedback on self-
assessment (the match vs. mismatch) and the VSPT result. The latter is in fact automatically displayed to the user who takes it, unlike all the other feedback that needs to be accessed via the feedback menu or by switching it on (the immediate item feedback). Thus, although only about a third of the informants rated self-assessment feedback and the VSPT result as ‘very’ interesting and useful, two thirds of them nevertheless said they usually read them always. This is an interesting discrepancy that deserves further research. Perhaps many learners who do not necessarily consider these types of feedback useful just want to read after each DIALANG test just in case they might learn something from them. Also the use of both types of item level feedback was reported to be somewhat more common than what their ratings of interest and usefulness might lead us to expect.

The remaining feedback (advice, extended descriptions of CEFR levels, reasons for mismatch between SA and test result) does not change that much depending on the test result but stays more or less the same from one test session to the next. Thus, it can expected that they are not read after every test – and this is exactly what the informants’ responses to the question about reading the feedback tell us in Table 15.

An analysis of the number of informants who actually replied to the questions on reading the particular type of feedback supports their actual responses (this is not displayed in Table 15, however): the number of informants responding to the question on test result was the highest (462, as not everybody replied to this question, especially if they had taken only one test), and the number was almost as high for the two types of item review and VSPT, but only 300 – 320 for all the others. The proportion of informants who replied to the question concerning their interest in different feedback and their perception of its usefulness was, however, very high and ranged from 470 to 550 depending on the feedback (the learners could read about all feedback and see screenshots of it in the questionnaire and so they could usually form an opinion of them even if they had not read the particular feedback, for whatever reason).

Another additional piece of information that Table 15 contains concerns the proportion of learners who were not sure how to judge the interest and usefulness of the different feedback (the ‘cannot say’ categories in Table 15). The final column in Table 15 also concerns informants’ uncertainty in forming an opinion when asked to rate the usefulness of each feedback. As explained in Huhta (2007b), a question was added to the last version of the questionnaire that asked the learners to say how strongly they felt that the feedback they had
just rated in terms of usefulness was either useful or useless for them. These responses are available from only 130 respondents but they nevertheless provide us with more precise information about the certainty of the learners’ views on the usefulness of feedback. The number in the ‘strength’ column indicates the percentage of those who responded to this question who said that they were quite certain in their opinion on the degree of usefulness of the particular type of feedback.

Again, the test result stands out as the type of feedback that the learners have a firm opinion on – typically a very positive one. Hardly anybody said they could not say if that feedback was interesting or useful, and 87% of those who estimated the certainty of their judgement of usefulness said they were quite certain in their view. The ‘cannot say’ and ‘strength’ responses do not, however, always give the same picture. Generally, not many respondents chose the ‘cannot say’ option but when they did, they concerned the information about SA, advice, the extended descriptions of levels, and the VSPT (4 – 7% of the respondents chose the ‘cannot say’ option). The examination of the responses in the last column in Table 15 that indicates the certainty of the usefulness ratings suggests that there were two types of feedback in DIALANG whose usefulness was difficult to judge by a considerable proportion of DIALANG users: they both concerned self-assessment, where only 61 – 67% were quite sure about their opinion.

The informants’ reactions to the feedback were triangulated in the questionnaire in several ways. As has been described above, the informants were asked to rate the different types of feedback with Likert-scale questions, and that they were also asked to evaluate the feedback and other aspects of DIALANG by answering open-ended questions. In addition, the informants were asked to compare the different types of feedback by selecting the three most useful parts of feedback and by marking which one of these was the most useful feedback, which the second most useful, and which the third most useful. Furthermore, they were asked to pick the three least useful types of feedback, and rank them from 1 to 3 in the same way as the most useful feedback. A total of 393 informants selected the most useful feedback and 373 the least useful feedback (not all questionnaire versions contained these comparative questions).
Table 16. Ranking of feedback: the three most useful vs. the three least useful parts of feedback

<table>
<thead>
<tr>
<th>Feedback Type</th>
<th>Most useful</th>
<th>2nd most useful</th>
<th>3rd most useful</th>
<th>Not mentioned</th>
<th>Least useful</th>
<th>2nd least useful</th>
<th>3rd least useful</th>
<th>Not mentioned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test result (CEFR level)</td>
<td>50</td>
<td>26</td>
<td>12</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>96</td>
</tr>
<tr>
<td>Immediate item review / feedback</td>
<td>24</td>
<td>22</td>
<td>12</td>
<td>42</td>
<td>12</td>
<td>7</td>
<td>5</td>
<td>76</td>
</tr>
<tr>
<td>Post-test item review / feedback</td>
<td>10</td>
<td>19</td>
<td>11</td>
<td>60</td>
<td>10</td>
<td>11</td>
<td>14</td>
<td>65</td>
</tr>
<tr>
<td>Self-assessment feedback (SA / test result match)</td>
<td>3</td>
<td>11</td>
<td>21</td>
<td>65</td>
<td>8</td>
<td>18</td>
<td>12</td>
<td>62</td>
</tr>
<tr>
<td>Information about self-assessment</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>94</td>
<td>30</td>
<td>24</td>
<td>17</td>
<td>30</td>
</tr>
<tr>
<td>Extended level descriptions</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>83</td>
<td>11</td>
<td>18</td>
<td>13</td>
<td>58</td>
</tr>
<tr>
<td>Advice on improving proficiency</td>
<td>6</td>
<td>10</td>
<td>12</td>
<td>72</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>64</td>
</tr>
<tr>
<td>Placement test result</td>
<td>4</td>
<td>10</td>
<td>14</td>
<td>72</td>
<td>15</td>
<td>11</td>
<td>18</td>
<td>56</td>
</tr>
</tbody>
</table>

Table 16 summarizes the rankings of the most useful and least useful type of feedback in DIALANG. The comparisons of feedback support the information obtained with the other types of questions. Again, the winner was the test result, which was chosen as the most useful type of feedback by 50% of those who ranked the usefulness of feedback, and very few selected it as one of the three least useful parts of feedback. However, twelve percent did not choose it among the top three feedback, which means that they considered it only ‘average’ among the eight forms of feedback in terms of usefulness. The usefulness of item review, especially the immediate item review, is more pronounced in these rankings than in the other types of questions reviewed above. 24% chose the immediate item review as the most useful feedback, and only 42% failed to include in among the top three. The post-test item review was the next most useful but is clearly behind the immediate review with 10% of top ratings. Information on self-assessment and the extended CEFR level descriptions were the feedback that was least often chose among the top three.
The selections of the three least useful types of feedback differ less from each other than the rankings of the three best feedback. The exceptions to this are the test result, which was almost never chosen among the bottom three, and the information on SA, which 30% chose as the least useful feedback. However, another 30% did not choose the information on SA among the bottom three, which shows that it divided learners’ opinions somewhat.

Finally, two more specific findings related to the feedback are presented. The survey questionnaire also included two further questions that concerned feedback on the individual test items. The first asked the informant to state which item review – immediate or post-test – he or she preferred. A total of 338 learners replied to this question, and 58% of them preferred the immediate item review, whereas only 30% considered the post-test item review better, and 9% thought they were equally good. This is in line in particular with the ranking information presented in Table 16, which also indicates clear preference for the immediate item review. Also the ratings of interest and usefulness in Table 15 discussed earlier display this tendency, especially when one looks at the percentages in the ‘very’ category for these two feedback. The overall conclusion, however, from users’ opinions on the two types of item-level feedback is that both of them are needed and that the possibility of choosing which one to use adds to the usefulness of DIALANG because users’ preferences were somewhat divided with some clearly preferring one and others preferring the other. It appears that the preference of either immediate or post-test item review is related to the interactiveness aspect of usefulness. Whether one uses an immediate item feedback, in particular, appears to affect the way in which the test-taker interacts with the test as a whole, at least, and possibly with each individual items. Some users commented that immediate item feedback disturbed and interrupted their test-taking process and even depressed them, when they got items wrong. Others, however, specifically wanted to know right away if they had replied correctly or not because they felt that was the optimal way for them to work their way through the test.

The second question reported here concerned usefulness of the categorization of the test items into subskills (of reading, writing, and listening) or areas (of grammar and vocabulary) on the screen for post-test item review. The question was included in only some questionnaire version, thus only 253 informants replied to it. A clear majority, 78% of them considered it ‘useful’, 6% ‘somewhat useful’, and only 6% ‘not very useful’ or ‘not useful’ (4%).
B. Other studies on DIALANG users’ reactions

Floropoulou (2002b) investigated five Greek and five Chinese students’ the attitudes to self-assessment and their reactions to DIALANG. Her findings indicate that informants’ culture was clearly related to their opinions: While most of them tended to deem self-assessment as a worthwhile activity, the Chinese were more positively disposed to DIALANG than the Greeks who tended to be indifferent or even negative towards the system. She also found out that the self-assessment implemented in the system did not automatically help learners untrained in SA to assess themselves accurately, as the accuracy of their SA both against the DIALANG test results and other information about their proficiency varied considerably. The informants had never really thought about self-assessment before the study, which made it difficult for some of them to do the SA task. Two of the ten students mentioned that it was difficult to make the binary yes/no decisions when answering the statements. Floropoulou argues that many users of DIALANG need support in using the system and doing self-assessment with it.

However, Floropoulou also reported that some informants appeared to change the way they viewed their language proficiency (in English, the test language) and that could identify their strengths and weaknesses with the help of the programme. Thus, the system appeared to increase some users’ awareness and it made them think about language proficiency, and also about the CEFR and how proficiency is conceptualized in it (all informants were completely ignorant of the CEFR before taking the test).

Floropoulou (2002, 31) asked the students to rank several aspects of DIALANG in terms of their helpfulness. The most helpful part was considered to be the ‘advisory feedback’, which consists of extended descriptions of the CEFR levels and advice on how to make progress towards the next level from the one you got on the test (5 out of the 10 informants selected this part). Immediate feedback on items was considered the next most helpful feedback followed by the explanations of why one’s SA and test result may not match, and the CEFR levels (apparently the test result expressed as a CEFR level). The self-assessment was considered somewhat less helpful than those already listed. By far the least helpful in the opinion of Floropoulou’s informants was the VSPT.
Finally, Floropoulou (2002b) also noticed that the students tended to compare DIALANG with the proficiency tests they had taken in the past such as the TOEF and the IELTS. Even though they were “well informed about the function of DIALANG i.e. self-evaluation, they seem to forget about it immediately after they get the results” (op cit., p. 35).

In another study, Yang (2003) investigated how test takers use DIALANG feedback. In her study, twelve postgraduate students at Lancaster University took the English reading test in DIALANG. Yang then interviewed them and used pictures of the feedback screens to support their memory when discussing their reactions to feedback.

Yang attempts to explain learners’ reactions to feedback on the basis of their goals and motivation for learning English, and their views about feedback and language tests. Overall, she found out that most of the subjects reacted positively to DIALANG feedback. However, their goals appeared to have an effect on, for example, whether they were willing to make the effort to follow the advice on language learning that they could read in the advice section of the feedback. Students whose main aim was to get the degree from the university rather than to improve their English were less likely to pick up the advice, as Yang found out after re-interviewing some of her informants two weeks after the test.

The majority of the subjects in Yang’s study failed to recognize the aims and functions of the DIALANG test (Yang 2003, 43). Their prior knowledge and beliefs formed from such proficiency tests as the TOEFL and IELTS prevented them from fully understanding the difference between DIALANG and other tests. Especially those learners who were doubtful about the test and its feedback tended to compare it with the international language certification examinations (see also Floropoulou 2002b above). For example, the role of the VSPT and self-assessment did not seem fully clear to them, nor the need to have such information in the test as the explanations of the reasons for possible mismatch between self-assessment and test result.

The informants seemed to consider that the main purpose of feedback is error correction, which is a very commonly hold view (see e.g. Kulhavy, 1977). According to Yang (2003), this led them to pay special attention to such feedback as the item review and the references to weaknesses in learners’ proficiency (rather than what they can do) in the extended descriptions of CEFR level found in the advisory feedback in DIALANG.
Yang also found out that the explanatory and advisory feedback made her informants reflect upon their learning process and helped them realize the factors involved in their self-assessment. The comparison of their current level with the next higher level in the extended level descriptions seemed to provide some learners with goals for learning, and the suggestions for improvement appeared to motivate them to make efforts by providing them with strategies with which to reach the goals. Yang argues that this reflection by the learners shows how the feedback can function as a catalyst for self-regulated learning, as argued by Butler and Winne (1995, 246). In the context of the present paper, Yang’s findings suggest that DIALANG feedback may enable DIALANG to meet one of its key aims, namely improving its users’ autonomous, self-directed learning of languages. This, then, also implies that the feedback serves to promote the kind of interactiveness between the DIALANG test tasks and the learners’ characteristics that the designers of the system hoped it to promote. Thus, these findings provide evidence for the interactiveness aspect of usefulness of DIALANG.

Yang (2003, 37) also came to the conclusion that the VSPT feedback was found to be the least effective type of feedback, because the majority of the subjects did not agree with the VSPT score. This finding is in line with the present author’s finding that the VSPT was considered the most problematic part of DIALANG in the survey of over 550 users of the system (Appendix 1 and Huhta 2007b).

3.2.2.4 Other empirical evidence about DIALANG feedback

An on-going project at the CALS provides empirical evidence that is relevant to the validity of the ‘Information about self-assessment’, which is part of DIALANG feedback (referred to as ‘Explanatory feedback’ in Alderson 2005). That feedback is aimed at learners who assess their proficiency with DIALANG self-assessment instrument but find out that their test result does not agree with their self-assessment. The feedback screen that provides the users with information about the match or mismatch between their SA and test result also provides them with direct access to ‘Information about self-assessment’, which is a set of inter-linked pages that describe why language learners’ self-assessment may not always match the outcome of language tests. The pages also contain more general information about language tests by briefly stating the main differences between DIALANG and most other Internet-based
(typically small scale and very one-sided) language tests, other large-scale international language tests, and the tests that language learners encounter at school. In addition, this section provides information about the more general differences between language tests and the so-called real-life language use, which also may help one to understand why the language test results may not match the learners’ expectations based on using the language in the ‘real world’. There was no one source to the content of this part of feedback but results from a range of different studies as well as experience of language teachers and learners were combined to produce these explanations.

The project that is relevant to the discussion of the validity of feedback is called ToLP (Towards Future Literacy Pedagogies. Finnish 9th graders’ and teachers’ literacy practices in school and out-of-school contexts), which is a Academy of Finland funded (2006–2009) study carried out at the University of Jyväskylä, Centre for Applied Language Studies (see www.jyu.fi/tolp for more information). The project deals with literacy practices of Finnish 9th graders (about 15-year-olds) and their L1 and L2 teachers in the comprehensive school, both in-school and out-of-school contexts. The project conducted a survey of the 9th graders based on a statistically representative sample of students on that grade, and a small section of the questions explored the students’ views on which factors they thought had contributed to their idea of knowing a foreign language either well or not so well. Thus, the question concerned the basis of their self-assessment of foreign language proficiency.

Some of the factors presented to the learners for consideration were very similar to the points included in ‘Information about self-assessment’ in DIALANG, which is why the students’ answers inform us, for example, about the relevance of these points. Are some of the points suggested in the DIALANG feedback on self-assessment factors that young, but experienced (typically, they had studied one foreign language for at least seven years and at least one other language for a shorter period) learners consider factors that have affected their view of how well they know the foreign languages that they have studied?

The precise question was this (translated from the original Finnish; see also Table 17):

_How do you know that you know a foreign language well or poorly? What affects your own view of your language proficiency? Select one option in each row._
Table 17 reports the responses of the 9th graders studied in the TOLP survey. Probably not unexpectedly, the results appear to confirm the assumptions made in the explanatory feedback in DIALANG on self-assessment that the experience and feedback that learners have in formal language education (lower-secondary school in this instance) has a powerful effect on how good vs. poor they perceive their proficiency to be. In particular, school tests and examinations appear to have a clear impact on how learners perceive themselves, at least according to their own opinion. The vast majority of all respondents estimated that school-related factors had had at least some effect on their views of their proficiency. Interestingly, however, the single most important factor in this respect, according to Finnish comprehensive school students, were how well they had actually managed to use the language for real-life purposes (point 2 in Table 17), so the school is clearly not the only influential factor.

Table 17. 9th graders’ view of what has affected their view of their foreign language proficiency in the TOLP survey

<table>
<thead>
<tr>
<th></th>
<th>This has affected my view of my proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a lot</td>
</tr>
<tr>
<td>1. How easily I learn the language in question at school.</td>
<td>35</td>
</tr>
<tr>
<td>2. How I manage to use the language (e.g. when traveling, using Internet, reading magazines).</td>
<td>57</td>
</tr>
<tr>
<td>3. What my teachers have said about my language proficiency.</td>
<td>28</td>
</tr>
<tr>
<td>4. What my friends and family have said about my language proficiency.</td>
<td>16</td>
</tr>
<tr>
<td>5. What the foreigners that I have met have said about my language proficiency.</td>
<td>23</td>
</tr>
<tr>
<td>6. How well I do on the examinations at school.</td>
<td>43</td>
</tr>
<tr>
<td>7. How well I know the language compared with the others in my class.</td>
<td>18</td>
</tr>
</tbody>
</table>

The figures in Table 17 are percentages of respondents who chose each response option (n = 1687).
Information about self-assessment in the DIALANG feedback section also suggests that some learners may compare themselves with other learners – at school or in the workplace, for example – and form their idea of themselves as good or poor users of the language based on that comparison. The importance of comparing oneself with others is also suggested by research in psychological research into personality (e.g., McFarland & Miller, 1994). The TOLP survey indicates that quite a few (youngish) language learners indeed say comparison with the other learners affects their view of the mastery of a another language, although that factor was not as important as the other factors related to school or real-life use. Interestingly, a closer analysis of the responses suggests that boys were more likely to compare themselves with their schoolmates than girls – or that they admitted doing that more readily than the girls.

Finally, a study by Luoma and Tarnanen (2003) should be mentioned although it does not directly relate the piloted or operational versions of DIALANG. For practical reasons, the Project had to limit the range of item types implemented in the system and leave out many, potentially interesting and innovative tasks (Alderson 2005). However, the Project designed and published a set of ‘experimental’ items that illustrate alternative approaches to utilizing the computer or to provide feedback. Some of the experimental items demonstrate how speaking and writing could be assessed by using self-assessment, and Luoma and Tarnanen (2003) describe the rationale and development of one such item that tests writing of Finnish as L2 (see item 12 in the Experimental Item Types Demonstration Programme available at http://www.lancs.ac.uk/fss/projects/linguistics/experimental/start.htm). The findings of the study suggested that the learners involved in it were able to self-assess their writing proficiency quite accurately with the help of the self-assessment task that was designed for the purpose.
3.4 Impact of DIALANG

According to Bachman and Palmer (1996, 30) the impact of test use operates at two levels: micro and macro. The micro level concerns individuals who are affected by the test, such as learners and their teachers, whereas the macro level relates to the educational system and society at large. The impact of assessment is often discussed in terms of how positive or negative it is, and the designers and administrators of assessments typically hope that the impact will be positive or that there will be no unintended consequences, especially negative ones.

Before analyzing the impact of DIALANG according to the Bachman and Palmer framework, it is important to consider certain factors that are relevant to the discussion of the impact of this particular language assessment system. A highly significant premise for any treatment on the effects that DIALANG has had is the fact that the system was not just another local testing system designed at the initiative of a few individuals or institutions. It was a high-profile venture – at least in the area of European language education – that came into being because of the political will of Directorate-General for Education and Culture of the European Commission. As described by Alderson (2005, 36-37), the system was conceived of as one of the assessment systems serving the European skills card scheme (European Commission 1995, 35) proposed in the mid 1990s by the Commission. The skills card was planned to be a document that EU citizens could use in order to demonstrate all the different kinds of skills they had acquired either through formal education or in some other way. Obviously, such a system needed procedures to certify citizens’ skills in different areas of life, and plans were developed to design instruments for certifying the different skills. Since there were already several international examinations for certifying foreign language skills, it was not politically feasible to design a new pan-European language examination system for foreign languages, DIALANG finally became a diagnostic rather than a certification system.

Because of the high political profile of DIALANG it was thoroughly reviewed and monitored by the European Commission during its development (see e.g., the evaluation report by Leney et al. 1998). From the beginning, DIALANG also attracted considerable attention among language education specialists in many European countries and beyond. The system was also presented in numerous seminars and conferences, and articles were written about it.
in journals, language teachers’ magazines and even in newspapers, at least in some European countries. Naturally, DIALANG has also been promoted by the European Commission in their language policy documents (e.g., European Commission, 2003). Thus, it can be expected that DIALANG has had a reasonable impact on language learners, teachers and educational institutions in Europe in particular. What the impact has been like and how many people are likely to have been affected will be examined next.

3.4.1 Impact on individual learners

The individuals most directly affected by test use are in most cases the learners taking the test. Bachman and Palmer (1996, 31) argue that test takers can be affected by three aspects of the testing procedure:

1. The experience of taking and, in some cases, of preparing for the test;
2. The feedback they receive about their performance on the test, and
3. The decisions that may be made about them on the basis of their test scores.

The impact of DIALANG on individuals, institutions or society has not been systematically studied. The research reported in the articles comprising the present thesis, and the supplementary findings reported in this paper, shed some light on the effects of DIALANG on individual test takers, including how they felt about taking DIALANG tests. The article on the VSPT (Huhta 2007a), in particular, contains learners’ comments on how they felt when taking those vocabulary tests that form part of DIALANG. To summarize, the VSPT divided users’ opinions like no other part of the system. It inspired some of the most positive reactions to DIALANG: taking the VSPT was felt to be a very different and an interesting experience compared with typical language tests. Some learners, however, did not like the experience at all because, for example, it seemed to them that the VSPT was testing something else than what they believed it should, or tested them in a way that they felt to be inappropriate – as is described in more detail in Huhta (2007a) and in previous sections in this paper.

There appears to be no published research on any preparation that language learners may do prior to taking DIALANG tests. This is perhaps not surprising because typically DIALANG is not – and should not be – used for high-stakes purposes, and there should be little need for
anyone to prepare for it unless preparation means familiarizing learners with the technicalities of the computer-based system that underlies the assessment system. However, DIALANG itself is sometimes used as a way to prepare for more high-stakes tests, as reported by some teachers at the Language Centre of the University of Jyväskylä during the collection of the survey data reported in Huhta 2007a and 2007b and in this paper. Those teachers considered it useful preparation for their end-of-course language tests that their students took DIALANG tests in the skills covered by those achievement tests, even if the medium of delivery was different, i.e., computerized rather than paper-and-pencil.

The second aspect of impact on individuals defined by Bachman and Palmer (1996, 31) concerns the impact of the feedback that a test delivers. As far as DIALANG is concerned, the impact of its feedback is highly important because feedback is such an essential element of the whole system. The individual user is hoped to be affected by the feedback in several ways, as explained at the beginning of this paper: he/she should receive information about the strengths and weaknesses in proficiency, become more aware of language and learning, and to be able to plan future language learning on the basis of the feedback.

Section 3.3.2.3-A-d above (page 106-114) above and Appendix 1 present a summary of learners’ comments on DIALANG collected in the survey of over 500 learners. These comments give us an idea of the impact of DIALANG feedback on typical users of the system. Users’ views on what is best in DIALANG clearly indicated that the feedback they received on their level proficiency was the most important thing they got out of the system. Many also wrote that the more detailed feedback on errors in specific test items was useful. In fact, all types of feedback in DIALANG were considered interesting and useful by significant portions of learners who participated in the survey, although the popularity of different feedback varied (see Huhta 2007b for details).

The third aspect of the impact of the feedback concerns the decisions made on the basis of that feedback. In the case of DIALANG, the most typical decision-makers are the learners themselves and the teachers who place students on their courses. Only rarely can somebody else such as a researcher or an educational authority use the information provided by DIALANG. It is somewhat difficult to know whether learners or language teachers are the most common users of DIALANG test results and other feedback. The system was initially designed with the individual learner in mind but it is clear that since its launch in 2001, an
ever-increasing number of institutions such as universities, adult education centres and language schools, and their teachers, have started using the system for placement purposes. In that type of use, the learners can obviously see their CEFR level, at the very least, when they view their test result but the main user and decision-maker is the teacher or other representative of the institution responsible for organizing the placement procedure. On the whole, however, the fact that the learners themselves are often the only stakeholders making decisions on the basis of feedback sets DIALANG apart from more high-stakes tests. Also, the type of decision is different: in DIALANG, the conclusions drawn from the feedback typically concern the overall stage and details of learning or the way forward. In higher-stakes and teacher-led testing, the decisions generally concern such matters as acceptance or denial of entry into studies, or awarding of grades and certificates (cf. Bachman and Palmer 1996, 32).

What we know of the impact of DIALANG feedback on individual learners who take the tests for other than placement purposes is, thus, that the vast majority of them appear to consider at least some of the feedback interesting and useful (see Section 3.3.2.3 and Appendix 1 for details). That the learners surveyed were able to form a view on DIALANG feedback had obviously required that they had at least read and/or used the feedback they received but it should be made clear that neither the present author’s survey of users nor any other available information can tell us much more about the impact of the feedback. More specifically, we do not know directly and accurately what kind of decisions these learners make on the basis of the test results and other feedback they received from DIALANG. In the following, the kinds of decisions that learners can potentially make after considering DIALANG feedback are briefly outlined.

The ‘decisions’ or conclusions that learners might make from DIALANG feedback are obviously related to the goals of the system. Learners are informed about their level of proficiency on the CEFR scale, and they also receive more detailed information about their proficiency, as well as information about their self-assessment. Whatever conclusions learners draw from this information is dependent on how believable they find the results they receive. If the results seem implausible, the learners are likely to reject the information as useless. However, if the feedback appears believable, there are several ways in which the learners might want to use such information. Probably the most straightforward type of decision concerns feedback on individual items and points of language: If the learner notices
that he/she made a particular error, it is quite possible that he/she learns that point of language or avoids the same error in the future, and he/she may check the problem point later in a dictionary or a grammar book. These were in fact specifically mentioned by some learners surveyed or interviewed by the present author at various points in the development of DIALANG.

It is more difficult to say what conclusions the learner draws from the general feedback on the level of proficiency, for example. The learner’s view of his/her proficiency may be confirmed or not, and this may, in a very general way, help him or her in setting goals for further learning. Perhaps the learner makes a decision to try one or more of the advice for making progress towards the next CEFR level. Indeed, a few learners surveyed mentioned that they might want to try to do as DIALANG advised them to do in order to improve their proficiency but whether they actually did so is not known.

One of the goals of DIALANG is to increase learners’ awareness of their own proficiency and of language proficiency, language tests, and language learning in general. How this relates to decisions the user makes from the feedback is not straightforward. Probably the only thing we can say about this is that being more aware of language and learning helps one to make more justified and realistic decisions and conclusions about one’s language skills and about what to do next. Some learners’ comments in the survey study suggest that they had indeed become (more) aware of certain things after reading the feedback. This was also found out by Yang (2003) in her study reported earlier. How this might affect language learning in the longer run is obviously beyond the scope of such cross-sectional studies that are reported in the current thesis. Only truly longitudinal research (see e.g., Ortega & Byrnes, 2008; Ortega & Iberri-Shea, 2005) could reveal how learners’ learning, beliefs, and other behaviour may have been affected by their use of DIALANG.

Yang’s (2003) small scale study in which she interviewed, via e-mail, some of her 12 informants two weeks after taking a DIALANG test, showed that some learners had indeed made an effort to do as advised by DIALANG but that others had not. Interestingly, her study also suggested that willingness to be affected by such feedback may be related to the goals and motivation for language learning of the learner.
3.4.2 Impact on language teachers, schools and other educational institutions

The studies reported in the articles that comprise the present author’s thesis focus on individual learners rather than language teachers and language-teaching institutions. However, in considering DIALANG’s impact on teachers and institutions, it is appropriate to review briefly what we know of this more general aspect of impact.

DIALANG is a battery of language tests that can operate independently of a language teacher; in fact, it was aimed at individual language learners rather than for groups studying under the supervision of a teacher. For this reason, there was no systematic effort to develop procedures and advice for language teachers on how to make use of DIALANG in their teaching. Nor was the work and research on teacher-based assessment (e.g. formative assessment) actively used in the development of the system, although it is clear that the goals of DIALANG and the goals of typical formative assessment are very similar (see Huhta 2008). During the pilot testing phase some advice on how to make use of the pilot tests in teaching was placed on the Project website for the benefit of teachers who helped the Project by administering pilot tests to their students (see Information for Language Teachers at http://dialang.org/project/english/index.html). Towards the end of the Project, some advice language teachers on using the operational version of DIALANG was developed but mainly due to lack of funding it was never translated and placed on the new DIALANG website (www.dialang.org).

However, dozens if not hundreds language teachers use DIALANG not only as a placement tool but also in a more versatile and integrated fashion on their courses. Such uses include at least preparation for other language tests and examination, and the training of self-assessment. Some teachers involved in the present author’s survey study of DIALANG users also used the system as a way to increase their students’ awareness of language and learning – which is obviously very compatible with the goals of DIALANG. It is very useful for many learners that a language teacher is available to explain what DIALANG is and to help them understand the results and other feedback. The need for teacher support in using DIALANG clearly arose in Floropoulou’s (2002b) study, in which some of the learners found the very idea of doing self-assessment alien, for example. Also, some of Yang’s (2003) informants did not fully understand the purpose of certain parts of the system such as the VSPT, self-assessment and some of the feedback. In both studies the learners view and understanding of
DIALANG was strongly affected by their previous experience of certification examination that have very different purpose than DIALANG. Also, some participants in the present author’s survey of DIALANG users mentioned that teacher support is needed in interpreting some of the information and in helping learners to draw conclusions for their own learning from the results. This all can obviously be seen to be at odds with the very common aim in language education, namely increasing learners’ independence and self-regulation, which is also one of DIALANG’s aims. However, not all learners are alike and some probably need a lot of teacher support, or, at least, they need it at certain points in their learning.

Taking into consideration all the information collected by the DIALANG Project and all the communications between the Project and hundreds of language teachers and institutions across Europe, it is clear that there was, and still is, widespread interest in DIALANG among the language teaching profession in Europe. Dozens, if not hundreds of institutions have used DIALANG for placement purposes, often on a fairly large scale, i.e., with hundreds of students per year, and for several years. As already mentioned, DIALANG is also used in many institutions as a self-study component, as a way to practice language and test-taking for more high-stakes tests, and as a way to understand and practice self-assessment – and for several other purposes. Again, we lack precise information of the number of institutions and teachers using the system for these purposes.

Although we do not know how many institutions use or have used DIALANG, we have some information why different institution use the system and what they think about it, which all shed light on the impact of DIALANG on institutions. A project-internal survey of typical institutional users of DIALANG across Europe was carried out in 2005 by the then Project Manager Gillian McLaughlin. The survey covered 58 institutions, 49 of which had already used DIALANG. The institutions were mostly located in the countries that are known to use the system most frequently (cf. statistics on tests taken and visitors to the website below), namely the Netherlands (14 institutions), Germany (9), Belgium (5), Italy (4), Switzerland, Finland, France, and the UK (3 each). Most of the institutions represented higher education (16 institutions), vocational training (13), and secondary education (12) but some also came from the private education sector or companies’ in-house training sections.
The institutions were asked to rate certain features of DIALANG on a 1 – 5 scale, where 1 indicated very weak and 5 excellent. The institutions’ mean ratings of the system were as follows:

- range of languages (4.64)
- clarity of interface, instructions, and navigation (4.27)
- content of the tests (4.18)
- quality of feedback and advice (4.07)
- reliability of the test results (3.89)
- ease of running DIALANG (3.73)
- ease of download and installation (3.71)
- efficiency of the system (3.58)

The purposes that the institutions were using DIALANG for – or were hoping to use it, if not yet using – included these: as a placement test (35 institutions), to monitor progress (31), as a pedagogical aid (29), and as an entrance test (23), or, more rarely, as a recruitment test (6). Thirty-three institutions gave an estimate of how many students they expected to use DIALANG during the year when the survey was carried out. On the average, each institution expected to have about 800 users, with a grand total of almost 26,000 students in the 33 institutions that provided an estimate on the number of students expected to use DIALANG.

Overall, then, the institutions surveyed by McLaughlin had a very positive view of DIALANG, which of course provides a good basis for the system having an impact on what and how these institutions do when it comes to testing the language proficiency of their students. If this sample of institutions generalizes into the institutions that use DIALANG, we can conclude that such institutions appear to value the unique range of languages offered by the system, and on the whole, they seem quite satisfied with its content, ease of use, as well as the technical aspects of the system. The number of students affected must be quite considerable (more on this below), and DIALANG appears to give language teaching institutions a useful tool that is typically used by a particular institution for more than one purpose. In particular, DIALANG seems to have replaced the earlier placement procedures in the institutions that have started to use it, and it offers them new options to support and monitor their students’ learning.

A further indication of the impact of DIALANG on language education in Europe and on organizations that have an important role in European language education is the use by the Council of Europe of DIALANG. It has already been mentioned that some of the procedures
developed in the DIALANG Project to link tests to the CEFR were later adopted in the Council’s Manual on linking examinations with the CEFR (and several members of the authoring group of the Manual and its Supplement were in fact former members of the Project: Figueras, Kaftandjieva, Takala, and Verhelst).

However, DIALANG contributed to the Council of Europe’s work in other ways, too. By far the most visible contribution is the inclusion of DIALANG scales in the new CEFR that was published in 2001 and has since then been translated into dozens of languages (i.e., Appendix C of the CEFR). Another DIALANG contribution is the inclusion of DIALANG reading and listening items for English, French, German and Spanish in the set of comprehension items published by the Council of Europe to illustrate the Common Framework levels (see http://www.coe.int/T/DG4/Portfolio/?L=E&M=/main_pages/illustrationse.html).

### 3.4.3 How many people are affected by DIALANG?

A way to get an idea of the number of people affected by DIALANG worldwide is to examine the publicly available statistics on the visitors to the DIALANG website (www.dialang.org) and the project-internal information on the number of tests taken.

The current web-counter has been in place since October 4, 2005, and it records a total of about 570 000 unique visitors to the DIALANG website by mid May 2009. This means an average of 440 visitors per day during the past four years. In 2008 there were 158 238 unique visitors. Most visitors have come from the Netherlands (about 100 000 or 17% of all visitors), Germany (61 000 / 11%), France (55 000 / 10%), Finland (44 000/ 8%), Spain (43 000 / 7%), Denmark (37 000 / 7%), Poland (29 000 / 5%), Italy, Belgium, and Switzerland (each 23 000 / 4%), the Czech Republic (11 000 / 2%), Sweden, the UK, the USA, and Austria (each over 9 000 / 1.7%), and Norway (about 5 800 / 1%). All in all, there are visitors from practically all countries of the world.

All the 14 different language versions of the website are visited on a daily basis. The widely spoken languages English, French, German, and Spanish naturally receive most of the visitors but the Dutch language site is equally popular due to the widespread use of DIALANG in the Netherlands and to a lesser extent in Belgium. However, also the smallest
languages such as Gaelic and Icelandic are visited by at least some interested individuals on most days of the year. This, then, testifies to the need among European users of access to information about such diagnostic tools in a wide range of different languages.

McLauing’s project-internal report from February 2005 showed that there had been 174,565 visitors to the DIALANG website between February 2001, when the system prototype and the website were launched, and February 2005. There is then a steady increase in the people visiting the system website, as can be seen in Table 18. Unfortunately, the data for 2005 is incomplete and the available statistics do not allow us to determine the precise number of visitors in 2006 and 2007. It appears clear, however, that the number of visitors was rather modest in the first years of the system, although it rose steadily till 2004, when the system was officially launched (in March 2004). After that, the yearly number of visitors almost doubled to about 150,000 by 2006 and seems to have remained at that level for the past three years.

Table 18. The number of visitors to the DIALANG web site (www.dialang.org)

<table>
<thead>
<tr>
<th>Year</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006 and 2007</th>
<th>2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visitors</td>
<td>11,247</td>
<td>18,332</td>
<td>37,074</td>
<td>87,031</td>
<td>about 100,000?</td>
<td>about 130,000–150,000/year</td>
<td>158,238</td>
</tr>
<tr>
<td>website</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total visits between Feb 2001 – Feb 2005: 174,565</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total visits between Oct 4, 2005 – Dec 31, 2008: 520,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td>2001</td>
<td>2002</td>
<td>2003</td>
<td>2004</td>
<td>2005</td>
<td>2006</td>
<td>2007</td>
</tr>
<tr>
<td>sessions</td>
<td>no info</td>
<td>no info</td>
<td>about 200,000</td>
<td>Jan – Sept 171,895</td>
<td>probably 200,000</td>
<td>203,000</td>
<td>218,000</td>
</tr>
<tr>
<td>started</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>summer 2003 – Sept 2005: over 340,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

According to McLaughlin (McLaughlin, 2005a), there were over 340,000 DIALANG test sessions started between the summer of 2003 and September 2005, which did not include almost two months in late 2004, when the system was down because of a technical failure. During a nine-month period between January and September 2005, a total of 171,895 tests were accessed.
Autumn months (September to November) are the busiest months of the year on the DIALANG website followed by February / March. These coincide with the first months of the autumn and spring terms in the educational institutions that provide the largest number of users of the system.

The DIALANG website in fact comprises 14 language specific sites. Of these, the English language pages have been the most popular: about a third of the visitors visit these pages, followed by the German and French pages (about 15% each), and Finnish and Dutch pages (9-10% each).

The total number of visitors to the DIALANG website is thus at least 700 000 so far, mostly from Europe. This means that a very large number of language learners and teachers are aware of the existence of the system. But how many learners actually take DIALANG tests?

As regards the number of test takers and the number of tests taken, the information is not as accurate and complete as for visitors to the website. Furthermore, that information is not publicly available, and thus we have to turn to some project-internal reports prepared soon after the development project came to an end at the end of 2004 and to more recent reports extracted from the system. The reason for the lack of accurate information is that users do not register to take DIALANG tests but they can use the system anonymously. In addition, the only reports that can be extracted on the use of the DIALANG system concerns the number of test sessions started on each day; information on the number of completed test sessions or the number of tests taken during each test session is not available. Thus, the statistics that we can access do not directly tell us how many learners use the system each year or each month, or how many separate tests are taken.

The number of test takers and tests can, however, be roughly estimated by putting together various pieces of information gathered during the DIALANG Project and in subsequent research studies. Visual examinations of the log files by the project members indicated that in most cases tests that were started were also completed by the users of the system (e-mail from McLaughlin 24 / 09 / 2004), which suggests that the typical test session results in a completion of at least one language test.
It is likely that there are fewer test takers than started testing sessions, however. Some sessions are aborted by the users – they may have just wanted to take a quick look at the system before deciding whether to take tests later. Other sessions end prematurely in a technical failure. A further reason why there must be fewer test takers than test sessions is that if a user wants to take several tests, he or she is unlikely to take them all in a single session simply because of lack of time available or fatigue. A typical DIALANG test takes a fair amount of time to complete, as there are 30 items in every test. It can be estimated that one test takes about 20 – 40 minutes, depending on the skill tested and whether the learner opts to take the VSPT and self-assessment before the actual language test. One session can, however, take even longer if the learner wants to study the feedback they receive carefully. The results of the survey of over 550 users of DIALANG by the present author indicate that the users spent typically spent 5-10 minutes reading feedback after the first test they had taken and somewhat less after subsequent tests (Table 19 below).

Furthermore, some learners take more than one DIALANG test session because they want to focus on different languages or different skills at different times – or they want to re-take tests to try to get a better result or to see if they have made progress in the language(s) and skill(s) they study. The fact that DIALANG has only one set of language tests (although at three different levels) probably limits the number of re-test takers because in most cases a re-test taker is likely to get exactly the same test as before.

Table 19. The length of time the users took to read DIALANG feedback in the author’s survey study

<table>
<thead>
<tr>
<th>How long the informants reported reading DIALANG feedback after the first test</th>
<th>How long the informants reported reading DIALANG feedback in general</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>valid %</td>
</tr>
<tr>
<td>1-2 minutes</td>
<td>140</td>
</tr>
<tr>
<td>about 5-10 minutes</td>
<td>340</td>
</tr>
<tr>
<td>about 15-30 minutes</td>
<td>40</td>
</tr>
<tr>
<td>over 30 minutes</td>
<td>6</td>
</tr>
<tr>
<td>total</td>
<td>526</td>
</tr>
<tr>
<td>missing</td>
<td>31</td>
</tr>
</tbody>
</table>
The informants in the author’s survey had taken on average two tests prior to responding to the questionnaire and that they intended to take further four tests later. As part of that survey, the author observed four groups of learners of Finnish as a foreign language taking Finnish tests during a two-hour session; the learners could take as many tests during the time available for them. Typically, they completed three or four tests, but some (generally very proficient) learners managed to complete all five skills / areas available for the language. This suggests that the total number of tests that a typical DIALANG test taker might take varies between two and six.

Most language schools and other educational institutions that use DIALANG for placement and other purposes typically require their students to take more than one skills test. For example, the Satakunta Polytechnic university (Jaatinen 2005) made their students take a total of six tests in two languages.

The discussion above suggests that the minimum number of test takers may be obtained by dividing the number of test sessions by three or four. This is based on the assumption that the average test taker wants to take six different tests and divides them between three sessions, i.e. he/she takes two tests in one session. (If we divide the number of sessions by four, we allow a quarter of the started test sessions to end in a technical failure or to represent just somebody’s quick peek at the system.)

To estimate a possible maximum number of test takers from the number of sessions we divide the latter by two. This assumes that the average user only takes two tests, which he or she takes in one session.

According to McLaughlin (2005), there were over 340 000 DIALANG test sessions started between the summer of 2003 and September 2005, which did not include almost two months in late 2004, when the system was down because of a technical failure. During a nine-month period between January and September 2005, a total of 171 895 tests were accessed. The busiest month was September with almost 39 000 tests, followed by March (almost 22 000), April (20 000), February (almost 20 000), and August (almost 19 000).
Applying the above rules to calculate the minimum and maximum number of learners taking DIALANG tests, we can estimate that between 40,000 and 85,000 learners used DIALANG during the first nine months in 2005, and that about 85,000 – 170,000 used it between 2003 and 2005.

The most recent system statistics available on the number DIALANG test sessions started (Gavin Smith, personal communication, March 2009) shows that there were slightly over 200,000 sessions in 2007 and almost 220,000 in 2008. It appears, then, that the yearly number of test sessions initiated has stabilized at around 200,000, which suggests that the number of learners who take DIALANG tests is about 50,000 – 100,000 per year, and the number of tests taken each year is somewhere between 100,000 and 600,000. Since the yearly number of test sessions appears to have been around 200,000 since 2004, the total number of learners who took DIALANG tests between 2004 and 2008 is around 250,000 – 500,000, and the number of tests taken around 500,000 – 3,000,000.

The comparison of the statistics in Table 18 on the number of visitors to the DIALANG website and the number of test sessions started shows an interesting pattern: there may be more people who actually take DIALANG tests than visit the website. This is most likely due to the fact that many educational institutions use the system for placement purposes and administer the tests to dozens or even hundreds of students each year who may never visit the website.

McLaughlin (2005) analyzed test data for certain selected periods in 2005 in detail to find out which languages and skills were tested and which interface languages were selected by the test takers. The periods thus analyzed were the whole month of February, a four-day period between September 15 and 19, and one day in March (2nd of March). For estimating the proportion of the languages as test languages, she had access to a larger data set, which covered at least 2004 but probably a somewhat (but unspecified) longer period. A summary of the information concerning the test and interface languages is presented in Table 20 below.
Table 20. The distribution of test languages and interface languages in DIALANG

<table>
<thead>
<tr>
<th>Language tested</th>
<th>Percentage of tests accessed in February 2005 (total number of tests accessed = 19,726)</th>
<th>Percentage of tests accessed in 2004 (exact dates unknown) (total number of tests = 226,539)</th>
<th>The most common interface languages for each test language (in percentages)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English (EN)</td>
<td>51</td>
<td>50</td>
<td>EN, FR, NL (over 20% each)</td>
</tr>
<tr>
<td>French (FR)</td>
<td>15</td>
<td>12</td>
<td>FR (30%), DE (25%), NL (25%)</td>
</tr>
<tr>
<td>German (DE)</td>
<td>11</td>
<td>12</td>
<td>DE (30%), NL (30%), FR (20%)</td>
</tr>
<tr>
<td>Spanish (ES)</td>
<td>5</td>
<td>5</td>
<td>ES (20%), NL (20%), IT (15%), DE (10%), FR (10%)</td>
</tr>
<tr>
<td>Dutch (NL)</td>
<td>4</td>
<td>5</td>
<td>NL (55%), FR (30%)</td>
</tr>
<tr>
<td>Swedish (SV)</td>
<td>4</td>
<td>5</td>
<td>FI (75%), SV (20%)</td>
</tr>
<tr>
<td>Finnish (FI)</td>
<td>2</td>
<td>4</td>
<td>FI (80%), EN (10%)</td>
</tr>
<tr>
<td>Italian (IT)</td>
<td>3</td>
<td>3</td>
<td>IT (35%), DE (25%), FR (20%)</td>
</tr>
<tr>
<td>Danish (DA)</td>
<td>1</td>
<td>3</td>
<td>DA (almost 70%), EN (10%)</td>
</tr>
<tr>
<td>Norwegian (NO)</td>
<td>0.40</td>
<td>1</td>
<td>NO (50%), NL (10%), DE (10%)</td>
</tr>
<tr>
<td>Portuguese (PT)</td>
<td>0.44</td>
<td>1</td>
<td>FR (20%), DE (20%), ES (20%), PT (15%), IT (10%)</td>
</tr>
<tr>
<td>Greek (EL)</td>
<td>0.40</td>
<td>0.47</td>
<td>EL, DE, EN, FR, NL (10-15% each)</td>
</tr>
<tr>
<td>Gaelic (GA)</td>
<td>0.18</td>
<td>0.24</td>
<td>EN (50%), DE (15%), IT, FI (10% each)</td>
</tr>
<tr>
<td>Icelandic (IS)</td>
<td>0.49</td>
<td>0.23</td>
<td>IS (40%), DE (40%)</td>
</tr>
</tbody>
</table>

Note: Unusually for the information that has typically been available from the DIALANG system, the Table 20 refers to tests rather than test sessions. It may be that the information available to McLaughlin in fact indicated which test / skill the user first accessed in the session he / she started.

Table 20 shows that English is by far the most popular test language in DIALANG: half of all the tests taken are tests of English. French and German are the next two most common test languages, followed by Spanish, Dutch, Swedish, Finnish, Italian and Danish. The first four languages are also among the most widely studied languages in Europe, which obviously explains their popularity as DIALANG test languages. The fact that Dutch, Swedish, Finnish and Danish have reasonably many test takers relates to the fact that the Netherlands and Finland in particular, and also Denmark to some extent, are countries in which DIALANG has been promoted quite extensively. The test takers in these countries include a lot of immigrants and other learners of the majority language of the country.

The interface languages used by DIALANG test takers tell us more directly about the countries in which particular language tests are typically taken, because many, if not most users select their mother tongue as the interface language – if that is available – unless they
are very proficient in the language they are taking tests in. Thus, French as an interface language for, say, German, Spanish, or Dutch tests tells us that the test taker probably comes from France or from the French-speaking part of Belgium or Switzerland. Similarly, the fact that the interface language for most tests of Swedish is Finnish indicates that most of the Swedish tests in DIALANG were taken by Finns, at least during the period in 2004-2005, when these statistics were gathered.

The distribution of the language skills tested via DIALANG was studied by analyzing the 18 451 tests accessed during a period of about a month in early 2005 (McLaughlin 2005). The most often accessed skill was listening (4 967 times), followed by writing (3 525), vocabulary (3 515), reading (3 375), and structures (3 069). The usage of the system was also analyzed during a shorter, four-day period in September 2005, during which 5 265 tests were accessed. The ranking of the skills was almost the same as in the first, longer period of monitoring: listening (1 422), writing (1 237), reading (966), vocabulary (915), and structures (725).

The two periods analyzed above are fairly short so it is difficult to say how accurate a picture they give about the skill typically tested in DIALANG. What is striking in the available data, however, is the interest shown by the users in taking listening tests. All the other skills or areas of language appear to be equally popular, with perhaps structures being slightly less popular than the other aspects of language proficiency. The great interest shown in listening may be explained by the fact that the vast majority of the other free, Internet-based language tests and quizzes lack listening. Thus, having a chance to test one’s listening skills may be particularly attractive to learners and teachers who come across with DIALANG.

On the whole, there is a lot in DIALANG that has the potential to cause an impact on an individual learner that differs from the impact of most if not all language tests they have taken before. Computer-based language tests covering most all the major skills except speaking were not very common when DIALANG was introduced to the world in 2001. The most common language tests, or test-like exercises, that one could find on the Internet were – and still are – short sets of tasks focusing on grammar and vocabulary. Thus, the experience of taking a full-scale language test on reading and listening, in particular, must have been a novel experience for most learners who took DIALANG for the first time (see the above discussion of the popularity of listening). Furthermore, the feedback from DIALANG was
likely to be unlike anything they had met before. First, language tests did not use the CEFR scale for reporting purposes in the early years of the 2000s – although this has now become much more common. Secondly, the range of feedback offered by DIALANG was, and remains, unique in language testing.

To sum up, it is clear that **DIALANG has had some impact on at least hundreds of thousands of language learners (and teachers) across the world. It is possible that the number of people affected is more than a million.** The countries where the system has had the strongest impact include the Netherlands, Germany, France, and Finland. **So far, there have been probably over 700,000 different visitors to the DIALANG website since early 2001 and several hundred thousand learners have taken DIALANG tests.**

### 3.4.4 Scientific impact

DIALANG has had a significant effect on language testing research. Before DIALANG, systematic work on diagnostic testing and assessment of second and foreign language proficiency was quite rare. In fact, there was no universally accepted definition of diagnostic testing in L2 testing, and no well-developed diagnostic instruments, grounded on theories of L2 development that could even remotely compare with those developed and regularly used for diagnosing development in first language proficiency (see Alderson 2005, 2007, and Huhta 2008, for a detailed discussion). The size and length of the DIALANG Project and its high profile in European language education made it much more visible than most other test development projects and thus attracted considerable attention as such. In addition, numerous presentations and demonstrations of the system, a fair number of articles in national and international publications on language education, and finally the book detailing the development and significance of DIALANG for language testing (i.e., Alderson 2005), all contributed to raising awareness of this almost totally neglected but potentially a very important area of assessment.

Too little time has passed since the development of DIALANG to change the overall picture: we still lack uniformly accepted definitions of diagnostic testing, we do not yet have theories to adequately guide the development of diagnostic instruments for L2, and we do not have properly developed and validated diagnostic tests and assessment procedures for L2
proficiency. However, a significant shift has occurred in language testing research in the sense that diagnostic testing has emerged as a new and interesting area, and the challenges of diagnostic testing and assessment are being tackled in several different ways (see e.g., Jang, 2005; Kunnan & Jang, forthcoming).

There is an increasing awareness of the need to develop our understanding of the construct of L2 proficiency to develop better procedures for diagnosing it (Alderson, 2005, 2007; Huhta 2008). It will probably take quite a long time and more work on diagnosis for a universally accepted definition of diagnostic testing in L2 to be established but work towards a better understanding of the concept is definitely on the way. Experience from general education suggests that defining diagnosis clearly is a considerable challenge (Huhta 2008).

Some new research specifically aims to contribute to our understanding of how foreign language abilities develop, and thus also how to diagnose it, by combining the expertise of second language acquisition research and language testing, as endorsed already a decade ago by Bachman et al. (1998). A good example of this type of approach is the work done by members of the SLATE network (Second Language Acquisition and Language Testing in Europe), which consists of researchers from both fields, who share an interested in understanding better the linguistic characteristics of the communicatively defined levels of the CEFR (see http://www.jyu.fi/hum/laitokset/kielet/cefling/en/desc, Section 3.2 for more information on SLATE).

Research on the relationship between CEFR levels and linguistic features is emerging in many European countries, partly related to the SLATE network, partly independent of it. A Finnish example is the CEFLING project, which focuses on analyzing Finnish and English writing by Finnish and immigrant adults and teenagers (see www.jyu.fi/cefling) and partly uses language examination data for the purpose. A Norwegian project studies Norwegian teenagers’ written English (Hasselgreen & Moe 2006) and another examines young Norwegians development in reading and writing English (Hasselgreen 2008). A British study examined Spanish and Chinese learners writing in English with the help of data from the IELTS test (Banerjee et al. 2007). Another study in Britain focused on the development of speaking by Greek learners of English with reference to the CEFR levels (Davou 2008). Cambridge ESOL has embarked on a major research programme that aims at designing detailed descriptors of grammar, vocabulary and functions needed at the different CEFR
levels; they use the Cambridge English examinations performance data as the basis of their research (see Kurtes & Saville 2008, and www.englishprofile.org for more information).

It should be added that one aspect of the work on linking linguistic features with the communicatively defined CEFR levels in many of the above mentioned projects is a study of task characteristics and the relationship between tasks and CEFR levels. Typical research questions include, for example: Can tasks be placed on specific CEFR levels? What is the effect of the task on the linguistic features that the learner is required (or can) to display, and how does that relate to the CEFR level? Do certain tasks allow the learners to demonstrate their proficiency at a certain CEFR level? Do other tasks prevent them from doing that (e.g. can an easy task prevent an advanced learner from displaying their true proficiency)?

A somewhat different initiative in understanding language proficiency in more detail is the Linguistic Correlates of Proficiency project at the University of Maryland, USA, which focuses on understanding and diagnosing advanced language proficiency in terms of sounds, vocabulary, grammar, dialects, and registers that pose problems for English-speaking learners of languages that are critical to the US national security such as Arabic (see http://www.casl.umd.edu/work/ProjectDetail.cfm?project_id=159 for more information). This is unusual motivation to attempt to diagnose L2 learning and proficiency but not unique in the history of language teaching and testing (remember the role of the CIA, FBI, and the US government sections responsible for foreign relations in the development of speaking tests and rating scales in the 1950s and 60s).

In addition to these recent developments in understanding L2 learning and proficiency, the contribution of the Educational Testing Service (ETS) in the USA to L2 diagnosis should be examined more closely.

The ETS has been particularly active in research into the cognitive diagnosis of the different tests that they design, including the influential TOEFL English language examination. Their line of research has been running in parallel with the work on DIALANG, and the two are the longest and best-known programmes that address the issues in diagnosing foreign language proficiency. The main differences between the two lines of work include at least the following:
• The ETS has analyzed its existing language and other tests that are used for certification purposes in order to find out what diagnostic information could be extracted from them. So far, they have not developed tests for purely diagnostic purposes. In contrast, DIALANG was, from the very beginning, an attempt to create tests that provide diagnostic information to its users;

• The ETS has studied tests that tap a range of different skills (e.g., Graf 2008); (Hartz & Roussos 2008), not only (foreign) language skills, although these, too, have been the focus of several studies (e.g., Freedle & Kostin 1996) (Sheehan 1997); (Buck & Tatsuoka 1998) (Schmitt 1999); (Qian & Schedl 2004); (Kostin 2004); (Jang 2005). In contrast, DIALANG is specifically a foreign language assessment system;

• The ETS has focused on the detailed analysis of test items and tried to understand what cognitive and other (e.g. linguistic) features underlie performance on different items by using very advanced statistical procedures and very large data sets. DIALANG had a much broader view of what is included in diagnosis. Feedback on individual items and the subskills that are hypothesized to relate to each item is only one aspect of the diagnosis that DIALANG offers, and there are many other types of feedback are available in the system..

• A key aspect of DIALANG is that it relates much of its broader diagnostic feedback to the CEFR scale, which has not been part of the research on diagnosis at the ETS. The fact that the ETS has carried out retrospective linking studies with the CEFR (Tannenbaum & Wylie 2008) has more to do with the more general need for the TOELF scores to be interpretable with reference to the CEFR because the test takers in Europe have increasingly started to use the CEFR scale as the yardstick with which to interpret the results of language tests even if the tests originate outside of Europe.

• DIALANG was mostly a development project with extremely tight timelines and limited resources to carry out research into the details of the largely uncharted area of diagnosing FL proficiency. The research activities were mainly restricted to the linking of test scores and self-assessment statements with the CEFR, the analyses of the pilot test data, the design of scoring systems to the tests, and the development of standard setting procedures. In contrast, the ETS line of research has been much more like a real research programme (or at least part of larger research programmes) that has continued for more than a decade, as far as work on diagnosing different abilities is concerned.
It was stated above that the ETS has not designed tests specifically and only for diagnostic purposes. However, their work on understanding what information different test items might provide has clearly paved way for expanding the feedback that the users of the ETS tests can get nowadays. The tests continue to serve the long-established purpose for providing users with certificates of the mastery of different skills and areas of knowledge, but there appears to be a new trend to provide the learners with more information about their performance than just the overall and sub-test scores. This is evident in the new TOEFL iBT system, where the test takers are provided with more detailed feedback in the form of performance descriptors that are very much like the ones found in the CEFR and in DIALANG (Wang, Eignor, & Enright, 2008 308-309) and that describe what a typical learner in any given TOEFL iBT score range can do in English.

Given that two major international lines of work on diagnosing language proficiency exist, obvious questions arise. Were the two groups aware of each other’s work? Was there cross-fertilization across the two groups of researchers? In the context of the present discussion, it is of interest to ask: Did DIALANG have any impact on how, for example, the TOEFL iBT was developed?

The decisions about the design of DIALANG were taken between 1997 and about 2000 / 2001. After that it was impossible to implement any major new features into the system due to the limited time and resources available to the Project. Thus, the DIALANG Project could only make use of work carried out by the ETS before the new millennium. The Project was aware of the research on the diagnostic potential of language test items by e.g. Buck and Tatsuoka (1996; 1998) but the technical complexity and the time it requires to analyze the test item content in such detail made it impractical for the Project to adopt the approach to diagnosing proficiency taken by Buck and his colleagues. However, the Project obtained the permission from the ETS to use their questionnaire on computer familiarity and implemented it as part of the pilot testing procedure. The analyses of the responses to these questions were carried out within the Project and they have not been reported anywhere. The overall conclusion of the analyses was that computer familiarity did not appear to be related to performance on DIALANG pilot tests, not at least among the pilot test population.
The DIALANG Project could, thus, make only very limited use of the research carried out at the ETS. Has the ETS made use of DIALANG in any way? This is quite hard to answer for somebody who is an outsider to the ETS. In theory, it is quite possible that DIALANG has had some impact on the development of the TOEFL iBT – certainly, the fact that it provides its users with quite detailed descriptors of proficiency indicates at the very least that the thinking of the developers of the new TOEFL has proceeded along very similar lines to the creators of DIALANG. It is certain that the TOEFL development team knew of DIALANG because the system was presented in a symposium at the Language Testing Research Colloquium at St Louis, Missouri, in 2001. One of the editors of a recent volume on the validation of the TOEFL iBT (Chapelle et al., 2008) also wrote a review of DIALANG in the major language testing journal in 2006 (Chapelle, 2006). However, none of the articles in Chapelle et al. (2008) refer to any publications on DIALANG, nor do they refer to the CEFR or to the work on CEFR scales by North. The other ETS publications on diagnosing language proficiency listed above do not appear to refer to DIALANG either, probably because their approach to diagnosis has been so different from DIALANG’s, as was described earlier in this section.

3.4.5 Use of DIALANG as a research instrument

The more general scientific impact of DIALANG was discussed in the previous section. DIALANG has also played a somewhat different role in advancing science as far as language education is concerned: It has been used as a research tool in several studies where the researchers have needed a way to gather information about learners’ proficiency. DIALANG has been considered to be sufficiently valid and reliable to serve as one or sometimes the only data gathering instrument in these studies.

The two biggest studies that have relied on DIALANG as their main research instrument are the German HISBUS study (Peschel et al. 2006) and the Finnish Satakunta Polytechnic university study (Jaatinen 2005), which have already been described above. Here, they are described only very briefly, with some relevant additional information about the studies.

The HISBUS study of the English proficiency of German university students was based on DIALANG reading tests in combination with students’ self-assessments. Previously, the HIS
studies had used only self-assessment as the instrument to gather information on the students proficiency in English (and in the other languages that the students might know). The repeated HIS studies allow the German educational authorities to follow the development of proficiency in English and French, in particular, since 1994 when the study was first conducted.

The DIALANG reading test was clearly the main data gathering instrument in the HISBUS study. In order to make the tests as dependable as possible, the researchers actually replicated the DIALANG testing procedure in their computerized, web-based test-delivery system. In co-operation with the DIALANG partnership, they programmed a mechanism that consisted of the Vocabulary Size Placement Test of English that assigned each test taker to one of the three levels of reading test, exactly in the same way as in DIALANG. The system copied both the VSPT scoring algorithm, the scoring mechanism of the reading tests (including item-specific weights), and the cut-off points for the different CEFR levels. The results of the study as far as the English proficiency of the students is concerned were described in Section 3.2.2.2 – C above. Suffice it to say here that the results generally appear believable on the basis of what is known about the background of the students and by comparing the test results with their self-assessed proficiency.

The Satakunta Polytechnic university in the south-western part of Finland also used DIALANG as a research instrument but for somewhat more complex purposes than the German HISBUS study. On the one hand, its purpose was to find out about the level of proficiency in English and Swedish of the students entering the institution (Jaatinen, 2005) but on the other hand the educators who designed the study had another purpose in mind, namely to use DIALANG results to demonstrate how varied, and in many cases inadequate, their students’ language proficiency was with respect to the official target levels. They also hoped that the study would support their arguments for obtaining more resources to the language education in their institution or for the lowering of the official target levels in certain cases.

The language teachers in the Satakunta polytechnic (and in other such institutions) were, and apparently still are, faced with considerable problems when having to try to bring a very heterogeneous group of students to the official level of proficiency set for English and Swedish for the students graduating from the Polytechnic. Particularly problematic are the
students (one quarter of all) who have only completed the compulsory education, i.e., the 9-year comprehensive school and typically also a 2–3 year vocational school, which means a total of 7 + 2 or 3 years of studies of English and only 3 + 2 or 3 years of Swedish. These students typically include many who have a low motivation to study languages. The other group of students (three quarters) have completed, after finishing the comprehensive school, the three-year academically-oriented upper-secondary school (the gymnasium), which has more hours of English and Swedish than the vocational schools. The students who go to the gymnasium are also typically more academically talented and more motivated to study languages.

In order to make the administration of the Polytechnic realize that the comprehensive/vocational school background students really started their studies at the Polytechnic with such low levels of proficiency that it was practically impossible for many of them to reach the official target levels, the language teachers decided to test all new students with the DIALANG English and Swedish tests (Scheinin, personal communication, January 19, 2004).

The findings of the study were as the teachers expected – or feared (see Section 3.3.2.2 – C for details). The comprehensive school background students’ proficiency in both languages, but especially in Swedish, was quite low and a sizable proportion of them were assigned at the two lowest CEFR levels. In Swedish, 50-80% of these students got A1 in the structures and vocabulary tests, for example, which means that they will have a long way to go to reach the official target B1.

DIALANG has also been used as a way to estimate learners CEFR level in studies that aim at establishing how proficiency develops across CEFR levels and what linguistic features characterize each level. Davou (2008) used the English DIALANG tests together with another language test and CEFR-based self-assessments to investigate how the speaking proficiency of Greek learners of English develops. Her specific interest was the formulaic sequences in spoken language, which affect how fluent and idiomatic the speaker is perceived to be. A similar study on the development of the French language is also using DIALANG to establish the CEFR level of the learners studied (Prodeau, personal communication, March 2009).
3.4.6 Reviews of DIALANG for other purposes

The creation of DIALANG was related to larger European Commission initiatives on education, training, and mobility of workforce, as explained above. Thus, it is probably natural that DIALANG has been paid serious attention to by the EU when European-wide language and language-related policies have been planned and implemented in the 2000s.

One of the most important activities that the European Commission has engaged in the 2000s is the gathering of information about the key competences of the EU citizens and to develop indicators (i.e. statistical indexes) that give decision makers and other interested parties an overview of the range and/or quality of people’s competences across Europe. Such key competences include, for example, literacy in the mother tongue, numeracy, ICT skills, and foreign language skills. The need for different indicators became more urgent because of the Lisbon Strategy drawn up by the heads of the European Union in 2000 that aims to make Europe more competitive (see http://ec.europa.eu/growthandjobs/key/index_en.htm for more information). A skilled population is seen as a key factor in maintaining and increasing the economic competitiveness of the EU but there is fairly little reliable information about the current skill levels of the population of Europe. Foreign and second language skills are one of the areas where the lack of dependable information is almost total. Thus, the European Commission engaged in the development of an indicator for foreign language skills, and it was eventually decided that the indicator should be the percentage of learners at the end of lower secondary education (ISCED2 level) in Europe who are at different levels of the CEFR in a range of foreign languages. Thus, the data for the language indicator will be gathered from about 15-16-year-olds rather than adults, although the indicators of other skills may focus on adults.

The gathering of information about language proficiency for the purpose of an indicator – or for any other purpose, for that matter – requires the use of some type of assessment instruments, which are typically either formal tests or self-assessment tasks. To find out if existing language tests could be used as those measurement instruments, the European Commission commissioned reviews of European and other language testing systems. West (2003) analyzed three major international systems: DIALANG, the examinations belonging to the ALTE (Association of Language Testers in Europe) organization, and the mother
tongue language tests used in the PISA (The Programme for International Student Achievement) studies of the OECD. In addition, she reviewed a number of national language examinations. It is indeed an indication of the importance of DIALANG that it was one of the systems examined for potential use in this major European study. The review came to the conclusion that the best approach to collecting data for the indicator would be to design completely new tests because none of the existing systems could meet all the criteria set for the instruments because they had been designed for different purposes (West 2003). (For more information about the European Survey of Language Competences, or ESILC, which is the study that is being conducted in 2008-2011 to provide data for the European indicator of language skills, see www.surveylang.org.)

Although the Commission selected teenagers as the focus group for the development of the foreign language indicator, adults are obviously highly important, too, and information and indicators of adults’ competences are needed in Europe. Thus, in parallel to West’s (2003) review of language tests for the European language indicator study, the Commission funded broader analyses of the options in testing adults’ skills. Haahr et al. (2004) conducted an analysis to aid the Commission to define a strategy for the assessment of different skills of the adults in the EU. Since the command of foreign languages is important for adults, they, too, were included in the analysis, and DIALANG was in the only language assessment system reviewed in the report. The analysis came to the conclusion that DIALANG appeared to be valid and reliable enough for the purpose of surveying adults’ foreign language skills but that a number of issues had to be solved if it were to be used as part of a standard household survey based on the researchers visiting individuals in their homes. For example, the number of skills to be tested and the time it would take to test them reliably were among the main issues. The report outlined different scenarios, one of which was to conduct a separate survey of foreign language skills rather than try to integrate it into a study of several other skills.

In 2005, Commission commissioned another, more detailed analysis of the options in assessing the skills of the adult population in the European Union, focusing on comparing the use of self-reports (self-assessments) and formal tests, analysing the feasibility of using computer-based assessments, and the relevance of international qualification frameworks (Haahr & Hansen, 2006). The review covered a range of skills in addition to foreign language skills such as literacy in L1, numeracy, and ICT skills. The recommendations included the
combined use of both tests and self-assessments because of their different but often complementary strengths and weaknesses as assessment instruments. The report also recommended the use of computerized measures whenever possible. DIALANG was reviewed in the report as an example of how an assessment system could be linked with an external, independent standard – the CEFR – and how tests and self-assessments could be combined in one system. At the time of the writing of this paper, the Commission has apparently not yet decided on the development of instruments to assess different skills of European adults, and it is thus not clear if and how adults’ foreign language skills might be tested (the ESLC study, as was mentioned above, concerns only teenagers at ISCED2 level of education).
3.5 Practicality of DIALANG

According to Bachman and Palmer (1996), the final aspect of test usefulness – practicality – differs from the other test qualities in that it relates to availability of resources and to decisions to use or develop a particular test – or not to develop. A test is practical if the resources that it requires (time, money, work, personnel) do not exceed the available resources; a test that is too long, too expensive or that requires special skills to administer, or to take, is impractical. Practicality is a relative concept, however, and a test may be perfectly practical for one purpose but not for another. It is probably the case that the most common tension that occurs between the different aspects of usefulness takes place between practicality and the other aspects.

The *a priori* design of DIALANG paid considerable attention to the practicality of this new assessment system because the main purposes of the system were related to diagnosing and helping individual learners, including learners studying languages outside formal education and, thus, without access to a teacher to explain the feedback to him or her. Since it was clear from almost the beginning of the project that the system would be computerized and that it would be delivered via the Internet, the IT system within which the language assessment system would run needed to be as easy to use as possible. Taking the language tests should be easy and straightforward. Taking the self-assessment instruments should be easy. And, finally, the feedback should be comprehensible even to a lay person without theoretical knowledge about language and language learning. The fact that DIALANG has been so much used for other purposes such as placement on language courses by educational institutions, in particular, adds a new twist to the consideration of practicality of the system. As it was not designed for that purpose, the system clearly lacks features that would be highly practical from the point of view of a language teacher or an institution testing whole groups of language learners.

Practicality has already been touched on when the other aspects of test usefulness have been discussed in the preceding chapters. The following briefly summarises them, area by area, and also elaborates on some of them whenever necessary. The treatment of practicality proceeds area by area, and within each area, the design (*a priori*) aspect of practicality is
discussed first, followed immediately by the presentation of whatever empirical evidence exists for or against the practicality of DIALANG in that area.

3.5.1 Practicality of the computerized system underlying DIALANG

The decision to make DIALANG into a computer-based assessment system rather than a paper-and-pencil based system was arguably the most important decision in terms of practicality. Extracting information from learners’ responses to language test tasks and delivering that information fast and with reasonable costs to thousands of users in 14 languages would have been impossible with a traditional, paper-based assessment system. If a large scale diagnostic system were to be developed, the only platform that made any sense was the computer. In this decision, then, the consideration of practicality was the unavoidable starting point – and thus the key aspect of usefulness as it dictated the overall approach to test delivery. It should be added, however, that using the computer offers other considerable benefits when it comes to diagnosing performance (see Alderson 2005), so this practical decision did not necessarily distract from the other aspects of usefulness discussed above.

The practical benefits of using the computer and the Internet as the basis of DIALANG include at least the following:

1. Item writing, translation, and review

   - Inputting items directly, via the Internet, into a common item database by item writers located in different countries saved time and costs (e.g. mailing costs). Errors in inputting could be minimised as the item writers themselves did that rather than e.g. a secretary or a student hired to do that ‘mechanical’ part of item writing.
   - Inputting of items with the help of common item templates is relatively easy and straightforward and it automatically ensures the same layout for the items, so this approach was adopted in DIALANG.
   - Item writers could review the input item immediately with the help of a Review Tool, which displayed the items exactly as it would appear to the learners taking the (pilot) test.
• Designated item reviewers could review items from their work place or home, across Europe.

• Using an on-line translation system (Translation Input System, TIS), the translators could input their translations needed in the system (interface, instructions, help screens, warnings, button names) from their work place or home, across Europe.

• Reviewers of translations could review the translations on-line via the TIS, from their work place or home, across Europe.

2. Piloting and standard setting

• Piloting of the test items and self-assessment statements, and the gathering of background data from the pilot test takers took place with the help of a programme called Pilot Testing Tool. It was downloaded by the institutions organising piloting and installed on their computers. During pilot sessions, the Pilot Tool established contact via the Internet to the DIALANG database and retrieved one of the several pilot test versions that had been selected for the purpose, and administered it to the test taker. The responses were transmitted via the Internet to the DIALANG database. The whole process saved time and resources as the transmission of data and recording of learners’ responses were done automatically by the system.

• The first type of standard setting used in DIALANG was partially done via the Pilot Tool. The experts doing the standard setting of the English test items could view the items on-line, as they would appear to the test takers, but they marked their judgments of the items on paper.

3. Operational test delivery, provision of feedback

• The operational version of DIALANG operates like the Pilot Tool. The system is downloaded and installed on the computers on which it is to be used, and then the use of the system takes place via the Internet. The obvious practical benefits of such a system are that taking the tests is not tied to a particular place or time, as is the case with traditional paper-and-pencil tests, but the test taker is free to use DIALANG at home, for example, if he / she has a computer and a broadband connection to the Internet.
• The system interacts with the users by recording their choices (interface and test language, skill to be tested) and by delivering the selected tests to them. It records their answers, marks them against the scoring key on the fly, and then delivers the results – by the item and in total (as a CEFR level) – and the other parts of feedback. The practicality of this is that interaction and feedback are far more immediate than with most human-scored tests in which the results are delivered after days or even weeks after the test. The amount of feedback is flexible and depends entirely on the needs and interests of the user of DIALANG.

• The upgrading, bug-fixing, and other changes to the operational system can be done simply by changing either the system that resides on the DIALANG server at Lancaster University or – in major upgrades – by changing the installation file that users then have to download and re-install from the DIALANG website at www.dialang.org. Such upgrades are thus very practical for the designers AND maintainers of the system and also relatively easy for the users, at least compared with a system that depends on CD-ROMs, for example, for delivery rather than the Internet.

A specific working group in the Project was assigned the task of designing the computer and software system underlying DIALANG: designing, programming, testing and documenting the software (e.g., item authoring, review, piloting, test delivery, and translation tools), selecting hardware and software for the workstations used by the programmers and for the servers used for storing and delivering the system to users around the world (e.g., items, scoring algorithms, and translations) (see also Alderson 2005, 38-41). The group included not only the programmers but also a coordinator who worked as an intermediary between them and the rest of the project coordinating team and who had the main responsibility for designing the specifications of the system for the programmers to implement. In the early stages of the working group also included other members of the Project who were experienced in applying IT for language testing or teaching. This group, then, had the main role in the a priori plans to make DIALANG system (software, hardware, interface, navigation) as practical as possible. As was the case with the other strands of work in the Project, the plans were formulated in a number of documents and memos, only the most important of which were circulated among the non-IT members of the Project (e.g., Figueras et al., 2000; Fligelstone, 2001; Fligelstone & Treweek, 2002).
3.5.2 Practicality of the content (assessment instruments and feedback)

The practicality of the content of a computerised testing system is inevitably closely linked with the practicality of the delivery mechanism itself, i.e., the software. However, the test tasks, self-assessment instruments, and feedback should themselves be practical in the sense that they should be easy, clear and straightforward to take and to interpret.

The issue of the clarity and intelligibility of both the software but, importantly, also such key aspects of content as the self-assessment instruments and feedback was tackled in two ways in the DIALANG Project: by translating them into all DIALANG languages and by using as simple language as possible.

Everything in DIALANG comes in 14 different languages, except the test items themselves. There were political reasons for this (see Alderson 2005 on the origins of the system) as one of the conditions for the project was that it covered all the official EU languages of the time, plus some languages with a special status in the particular Directorate General of the European Commission that funded the development. However, very practical considerations also underlay that language policy decision – as is the case with many other EU language policies – namely the aim to make the outcomes of the project as accessible to the citizens of the EU as possible by using a wide range of European languages. This political aim nicely coincided with other, project-internal considerations of practicality: it also ensured that very many learners can use of DIALANG via an interface language that they master very well. Admittedly, the requirement to have everything translated into 14 languages also caused considerable practical problems in the design stages of the project but from the point of view of the users the translations obviously increase the usability and overall usefulness of DIALANG very significantly.

The other main approach to increasing the intelligibility of the content of DIALANG was the use of as simple language as possible in the instructions, feedback and self-assessment. Huhta & Figueras (2004), Luoma (2004), and Alderson (2005) describe and illustrate how the Project derived its self-assessment statements from the CEFR scale descriptors. Although the original statements were already fairly clear, some of them were further simplified by splitting them into two and by using even simpler vocabulary and expressions. The use of
straightforward language and the provision of illustrations and concrete examples were also important considerations in the construction of the feedback texts (e.g., advice on how to improve one’s proficiency and the extended descriptions of the CEFR levels) because, as mentioned earlier, learners outside formal language education and teachers’ support should also be able to understand the feedback they receive from DIALANG.

No large-scale studies have been conducted specifically to find out about the practicality of DIALANG as a computer system or about the practicality of its content. We thus have to resort to reasoning and to what little empirical evidence we have about the practicality of DIALANG.

As reported above in the section on the reliability of DIALANG technology (itself an important aspect of practicality), Floropoulou (2002a) studied the technical aspects of DIALANG such as navigation and interface by observing and interviewing six informants. Her informants experienced certain navigation problems when using DIALANG for the first time. For example, they were unsure about how to proceed at certain points in the test taking process and they had difficulties in interpreting the meaning of some of the navigation buttons. They also may not have fully understood the role of such parts of the system as the Vocabulary Size Placement Test. On the other hand, however, they also liked the general layout of the interface and the fact that they could read the instructions in their native language (see Section 3.2.2 above).

The present author’s survey of over 550 users of DIALANG also yields some information about the practicality of DIALANG (see Appendix 1 for details). The survey questionnaire did not include specific questions about the interface and navigation, for example, since it focused on soliciting the informants’ views on the feedback. However, the questionnaire contained open questions about the best and worst features of the system, which allows us to see what proportion of users decided to comment on things that somehow relate to the practicality of the system.

Table 1 in Appendix 1 reports which features the informants considered to be best in DIALANG. The majority of the comments concerned content matters and different aspects of feedback but 7% of the informants included clarity of instructions and information among the best aspects of the system and another 7% singled out ease of use. Concreteness and quick
delivery of feedback, free choice of the time and place of test-taking, lack of fees and time limits, and freedom from the teacher were also mentioned by a few percent of the respondents.

Table 2 in Appendix 1 lists the most problematic features in DIALANG, according to the informants. It is apparent that the technical problems were among the top three problem areas as 11% of the respondents mentioned them. While technical issues cover a range of very different problems, it is clear that probably most of the technical problems in a computerized system have something to do with practicality. Length of some aspects of the system such as the tests, instructions or feedback was considered a problem by 6% of the informants, 5% found fault with the interface or layout and 4% with the overall functioning of the system. Some also specifically mentioned navigation problems or problems with the instructions.

The two studies above and the feedback received by the present author when he worked in the Project indicate that there is clearly room for improvement in DIALANG as a computer system: The system does not always work or it does not work for everyone. Installation of the programme can be difficult under certain circumstances (e.g. high-security computer labs). There are also some smaller, but annoying technical issues with the functioning or display of certain items with certain operating systems. The navigation in such a complex system can be a challenge to some first-time users. However, as mentioned earlier, most complex computer programmes take more than one session to become familiar enough to their users to navigate without problems.

Some of the technical problems experienced by users of DIALANG are apparently no fault of the system itself but are rather caused either by the local network in the institutions using the system or by the local or more general bottlenecks in their Internet connections. For example, the slow functioning of the programme mentioned by some users over the years is typically caused by underperforming local network and / or the local Internet service provider. These problems seem to affect the listening tests the most, as the audio files are the biggest single elements that are downloaded from DIALANG servers onto users’ computers while test-taking is in progress. As an example of this, the present author was very recently reported of attempts by an educational institution in one of the Persian Gulf states to administer DIALANG listening tests to its students but the trials failed because of the network could not handle the simultaneous downloading of many audio files. (Group listening tests are

154
successfully conducted in many other places and countries, however, where the speed and bandwidth of local connections are greater.

Despite the technical practicality problems faced by some users of DIALANG, other users have found the system easy to use. Since we lack more comprehensive studies of the practicality of the system it is impossible to say how widespread or serious the technical and other practical problems that the users face are in reality. The fact that tens of thousands of learners use DIALANG every year speaks at least for a reasonable level of practicality of the system: The benefits (overall usefulness) of the system apparently outweigh whatever technical and other obstacles the institutional and individual users of the system face.

It is clear that some aspects of practicality as the amount of information (e.g. feedback) and the lack of time limits can be viewed either as positive or negative features of the system. The lack of time limit in DIALANG was considered a problem by Floropoulou’s informants whereas it was one of the good things about the system for some of the respondents in the present author’s study. In the latter study, the considerable amount of information found in DIALANG divided users: while many appeared to value the chance to get a variety of different feedback there were also some who thought there was too much of it. Possibly learners’ different experiences and expectations as regards language tests affect their views on what is practical vs. impractical in a test (cf. Yang’s 2003 findings about the influence of other tests on how learners perceived DIALANG).

Before concluding discussion of the practicality of DIALANG we should briefly consider whose view of practicality ought to be considered. The above discussion has focused on the practicality of the system for the language learners who use it. There are, however, at least two other stakeholder groups for whom practicality can be important: the people who design and maintain a system such as DIALANG and the institutions (schools, universities, companies) that use it for placement purposes, in particular.

As was briefly mentioned above and in Alderson (2005), the construction of such a complex system as DIALANG caused considerable practical problems for the Project both in terms of managing a large, multinational and multi-part project and in terms of designing an extensive, high-quality system with the limited time and resources available. However, DIALANG managed to pass the crucial practicality test suggested by Bachman and Palmer.
(1996) namely that despite the considerable practical challenges the decision to develop DIALANG in the first place was made and that it was indeed successfully designed. It has also been successfully maintained since the end of the Project in 2004 despite minimal resources available for maintenance in the DIALANG partnership and at Lancaster University where the system database and server are located. This indicates that the basic system design underlying the content is quite robust despite the occasional technical problems encountered by some users. The challenge in the future is to find a permanent solution to the hosting, maintenance and further development of DIALANG, which has turned out to be the most difficult long-term problem that the developers of the system encountered – to this day, this issue remains unresolved.

The practicality of DIALANG for language teaching institutions should also be considered. DIALANG was never designed to be used for placement purposes by institutions; it was intended for individual learners for individual diagnosis and feedback. Thus, DIALANG lacks many features that would make it a practical tool for institutions and for placement purposes. The test result is only reported to the user and it is not stored, nor can it be printed directly from the programme. It is not possible to retrieve the results for groups of learners as a single results file, for example, but the users must be asked to tell their result to the teacher supervising the testing session or they must be asked not to quit the programme before the teacher can come and collect each individual’s test result. There is no way the teacher or institution can automatically assign test takers to particular tests – they can only ask or tell the students to choose particular tests and count on their cooperation. Thus, DIALANG is clearly far from an optimal tool in terms of practicality for institutions.

With the above in mind, it is quite remarkable that DIALANG is used so much for placement purposes across Europe. The system appears to have enough practicality in it to be considered useful by many institutions: it is free of charge and it is quite easy to access once it has been installed on the institutions computers. Indeed, McLaughlin’s (2005b) survey of institutions using DIALANG found out that the institutions considered the clarity of interface, instructions and navigation the second best aspect of the system. Also the ease of downloading, installing and running the system were regarded as relatively easy. However, the content of DIALANG is also important when institutions make decisions to use it. McLaughlin found that the range of languages available was the most important
consideration by the institutions using or planning to use DIALANG, and that the content of the tests and feedback were also important factors. (see Section 3.4.2 above)

Finally, the practicality of DIALANG for some quite different purposes has also been considered. Because the European Commission funded the development of the system, they have evaluated the system for potential use for gathering information about the foreign language skills of EU citizens, either adults or students at school (West 2003; Haahr et al., 2004; Haahr & Hansen, 2006). All reviews came to the conclusion that although DIALANG was apparently a high-quality system it was not feasible (practical, economical) to use it in its current shape for the purposes of gathering data for educational and policy-making purposes. Considerable, and expensive, changes to the system would have been required. This does not obviously diminish the practicality or usefulness of DIALANG for its intended purpose; the above reviews evaluated a range of other assessment systems and came to the conclusion that systems designed with very different goals in mind could not be used for those new purposes without considerable modifications. This is not surprising and simply testifies to the common sense and professionalism of the persons conducting the evaluation of DIALANG and the other assessment systems.
4. SUMMARY AND CONCLUSION

This paper has summarized and synthesized the content of the six articles on the DIALANG assessment system that comprise (together with this paper) the doctoral thesis of the present writer: Huhta & Figueras (2004), Huhta & Alderson (2005), Huhta 2007a, Huhta 2007b, Huhta (2008), and Huhta (submitted). The paper has also given a more comprehensive overview of what we know about the usefulness of DIALANG, mainly based on studies by both the present author and many other researchers. The thesis as a whole has aimed to shed light on these broad questions:

- What do we know about the usefulness of DIALANG on the basis of the research and development work carried out and reported by the present author (both in the six articles and in this paper)?
- What do we know about the usefulness of DIALANG on the basis of research results, statistics on the users of the system, and other relevant information gathered by others working in the field of language education?
- What is the overall picture of the usefulness of DIALANG when we combine the two sources of information above?

The analysis of test usefulness in this paper was organised in terms of the framework for test usefulness proposed by Bachman and Palmer (1996), in which test usefulness is evaluated with reference to the purpose(s) of the test. According to Bachman and Palmer, several factors contribute to the usefulness of a test for its intended purpose(s): reliability, construct validity, authenticity, interactiveness, impact, and practicality. In principle, the better the quality of each of these aspects the more useful the test is. However, there is usually a tension between the different factors since maximising the quality of one is often detrimental to the quality of another feature. Thus, it is essential to consider the purpose of the assessment to be able to decide which aspects of quality are the most crucial ones for any given purpose. In fact, the overall usefulness of the assessment is what really counts, and that cannot be evaluated without considering all factors that contribute to its usefulness.

It would have been possible to apply some other framework to examine the quality or validity of DIALANG, such as the one developed by Messick. However, the application of the Bachman and Palmer framework of test usefulness to analyse the quality of DIALANG
turned out to work quite well and while it was not always straightforward to decide where in the framework to place each argument or piece of evidence this was not an insurmountable problem. Furthermore, the framework is quite comprehensive and probably covers all the essential aspects of test quality. However, the approach also led to some repetition in the presentation of the arguments and evidence for the usefulness of DIALANG, as the same studies and sometimes even the very same evidence (or argument) contributed to more than one aspect of test usefulness.

The main purposes of DIALANG are to enable learners of foreign languages to find out (1) their level of proficiency in the skills they are interested in, and (2) their profile of strengths and weaknesses across skills and sub-skills, (3) to increase their awareness of their proficiency and language learning in general (including self-assessment), and (4) to help them take more control of their language learning. Also (5) language teachers can use the above information, for example, to plan their teaching of the particular learners. Although not among its main intended purposes, DIALANG has also become to be used extensively by (6) language teaching institutions as a placement test for their courses and programmes. The evaluation of the usefulness of DIALANG should, then, consider all these interlinked purposes.

The discussion of the usefulness of DIALANG in this paper made use of both theoretical arguments and empirical evidence. Thus, in addition to the Bachman and Palmer (1996) framework of test usefulness, a broad division into \textit{a priori} and \textit{a posteriori} validation of assessment has been used to organize the present discussion. \textit{A priori} validation includes all non-empirical work and theoretical, rational argumentation that test designers engage in during the development of the test, before it is piloted or used for real (Weir 1991, 1993, 2005; Messick 1989). \textit{A posteriori} validation uses empirical data (both test scores and qualitative data) that are gathered when the test is actually administered to real language learners, either when it is piloted or when it is has reached the operational stage – or, ideally, both. A well-grounded test design project relies on both \textit{a priori} and \textit{a posteriori} type of evidence about the quality of the test.
4.1 *A priori* theoretical arguments for the usefulness of DIALANG

The key factors in the theoretical arguments for the usefulness of DIALANG for its intended purposes are the documents that guided the construction of the assessment system. DIALANG was the first major language assessment system based on the Common European Framework of Reference (CEFR), which means that the CEFR is a key document in terms of *a priori* validation of DIALANG. In order to define the content of the tests (key consideration for the authenticity and construct validity of the tests), the DIALANG Project complemented the CEFR with its own Assessment Framework that further detailed, e.g., the communicative activities to be covered in the tests.

However, since the CEFR had some clear gaps in it – in particular in the definitions of comprehension but also of texts – the Assessment Specifications that defined the actual type and content of test items had to be supplemented with information from many other sources. Furthermore, it was necessary to design many additional, more focused documents for the development of such crucial aspects of the system as the underlying IT system, self-assessments, feedback, translations, as well as the different phases of test design (piloting, analyses, standard setting). All these documents and developments were crucial for the reliability, construct validity, authenticity, interactiveness, and practicality aspects of test usefulness. Typically, the planning work and design of documents and guidelines was carried out in Working Groups set up specifically to develop and carry out work in the different activity areas listed above.

The main challenges in the *a priori* stage faced by the Project included at least the following:

- The sheer size and length of the project (14 languages, over 20 partner institutions of varying experience, several strands of work with their own Working Groups, eight years of work) made communication, agreements, proper allocation and timing of resources, and monitoring of the work difficult;
- The lack of previous research on diagnostic assessment of foreign language learning and proficiency, and even commonly agreed definitions of diagnostic assessment of foreign language proficiency;
The fact that DIALANG was the first assessment system to be based on the CEFR meant that the Project had to design from scratch all procedures for using the CEFR as the basis for test content development and for linking the test results to it.

How persuasive are the *a priori*, theoretical arguments for the usefulness of DIALANG? To answer this question properly would benefit from a thorough knowledge of typical design phases of other large-scale language tests. Since no other large-scale diagnostic L2 tests exist, we cannot compare DIALANG with them. However, my experience in designing language proficiency tests (e.g. National Certificates: Huhta & Takala, 1997; Huhta, 1997) and an examination of the typical validation procedures used in international language examinations (see e.g. Luoma 2001), suggest that the quality of the development of DIALANG is comparable with the quality of the development of these more high-stakes examinations. On the contrary, the Project was quite ambitious, which partly explains why it lasted so long; especially the procedures developed for making use of the CEFR, analyzing the pilot test data, and linking the test results to the CEFR scale involved cutting edge research and development. The Project was fortunate in having a number of world-class experts in several key areas of development, which made such research and development possible. The fact that the Council of Europe made such extensive use of the products and procedures developed in the Project, as described above under the impact of DIALANG, testifies to the high quality of the work carried out the by the Project.

The main *a priori* limitation of DIALANG derives from the lack of knowledge about diagnosing foreign language learning and proficiency. DIALANG incorporates a range of different approaches to diagnosis but we still do not know very much about how effective and appropriate the approaches are, despite some empirical evidence summarized in this paper. Importantly, we do not know what other, more appropriate, types of diagnoses of learners’ proficiency could be used or what the long-terms effects of DIALANG are on its users. Only future research, for example, along the lines of on-going studies reported in Section 3.4.4 on DIALANG’s scientific impact can address this gap.

Another issue in the *a priori* design of DIALANG that should be mentioned that there is an obvious contradiction between how DIALANG measures language proficiency and how language competences are defined in the CEFR, which may be problematic as regards the construct validity, authenticity, and interactivity of DIALANG tests. DIALANG follows
the traditional division of tests into reading, listening, and writing, and furthermore, into vocabulary and structures, whereas the models of language proficiency developed by Canale and Swain (1980) and Bachman (1990), and included in the CEFR, define and categorize language abilities somewhat differently. This may not be a serious issue, however, as such aspects of proficiency as sociolinguistic appropriacy and functional abilities, which are essential elements in Canale and Swain’s and Bachman’s models of communicative competence are in fact operationalized in DIALANG tests.

4.2 *A posteriori* empirical evidence for the usefulness of DIALANG

No comprehensive study has been carried out on any of the aspects of test usefulness in the Bachman and Palmer framework but we have several studies and other information that shed light on, in particular, the construct validity and impact of DIALANG. There is also some evidence on the authenticity, practicality and reliability of the system. The aspect of usefulness for which we do not have much evidence is the interactiveness, that is, the extent to which the test tasks engage the test taker’s areas of language ability, personal characteristics, topical knowledge, strategic competence, and affective schemata.

It is important to also bear in mind that most of the empirical evidence only concern certain languages or aspects of the system. And even if roughly the same information is available for all or several of the 14 languages in the system, the usefulness of all its language tests cannot logically be equally high simply because of the different state of readiness of the languages in the system. While the tests in some languages have been properly piloted and the results analyzed, other languages remain non-piloted, which means that the reliability and construct validity, in particular, of the 14 different language test systems of DIALANG are bound to differ. English is the most studied and most thoroughly analyzed DIALANG test language because the number of pilot test takers was biggest for it (Alderson 2005) and because it is also the most popular test language in the operational system. This means that most of the studies using operational DIALANG in fact concern the English tests and related self-assessments, VSPT, and feedback. The fact that Danish, Gaelic, Greek, Icelandic, Norwegian, Portuguese, and Swedish tests are not based on empirically piloted data, due to lack of pilot test takers, automatically means that their test results – and thus much of their feedback – is less reliable and probably also less valid than the results and feedback on the
other languages. To alert users to this fact, the system warns them about this if they select one of these non-piloted languages as a test language.

There is evidence of a fair level of reliability of the English pilot tests in the sense of their internal consistency (Alderson & Huhta 2005; Alderson 2005), which suggests that the operational English tests in DIALANG Version 1 are also quite reliable. There is also some evidence in the same study about a fair level of reliability of the standard setting for English and German. Evidence about the reliability of the VSPT, however, suggests that it is an unreliable vocabulary size test and an unreliable placement test if the learner resorts to too much guessing when taking the test (Huhta 2007a). This could however be remedied by changing the scoring algorithm of the VSPT, if there were resources to implement such a change to the system.

The other reliability issues of DIALANG about which we have at least some empirical evidence include the technical reliability of the IT system that underlies DIALANG assessment procedures and the reliability of the scoring of open-ended questions found in its language tests. Feedback from users during the project as well as some studies (Huhta, this paper; Floropoulou 2002a) indicate that a certain but unknown proportion of users experience some problems with the technical reliability of DIALANG (e.g. breakdown or slowness of service or layout problems). The fact that a great number of learners use the system on a daily basis indicates that significant numbers of users do not experience serious technical problems in the reliability of DIALANG service.

Finally, the questionnaire survey by the present author suggests that the scoring of the open-ended questions is sometimes unreliable for at least some users (see Appendix 1). Again, it appears that the vast majority of test takers and items are not affected by this problem, simply because such open-ended items which may suffer from an inadequate scoring key are a clear minority of all items in the system.

There is also some indication of a fairly high level of internal consistency type of reliability of DIALANG self-assessment statements (Alderson 2005).

There is some empirical evidence about the construct validity of DIALANG, particularly from the point of view of the concurrent validity of DIALANG: the relationships (1) among
different DIALANG tests, (2) between the DIALANG tests and external measures of proficiency, and (3) the ability of DIALANG tests to distinguish between different, well-defined groups of learners.

The first type of concurrent validity evidence traditionally concerns the correlations between different tests in the same battery of tests, as they indicate to what extent the tests measure the same or different aspects of, for example, language proficiency.

The relationships between different DIALANG tests in several studies (Alderson 2005; Jaatinen 2005; Huhta, this paper) turned out to be quite strong but far from perfect. The correlations are of the magnitude that could be expected from a test battery that includes tests that tap different areas of proficiency that can be theoretically and logically regarded as different skills. On the other hand, as is the case with most language test batteries, the tests are quite clearly associated with each other, which is in line with the idea of an underlying ‘general language’ competence that is often found in empirical research into language proficiency. These findings are thus in line with theoretical expectations and previous research. They also suggest that it is meaningful to provide separate diagnostic feedback on each of the skills tested in DIALANG.

The VSPT in its current format is not quite as closely related to performance on the other DIALANG tests but that if the scoring were changed, it probably would have the same level of correlations as the other tests. As could be expected, it is somewhat more closely related to the vocabulary test than to the other tests – the fact that it also correlates well with the writing test may be due to the fact that several items in the (indirect) writing tests used in DIALANG are in fact vocabulary items.

The results of some of the studies also suggest that the profile across different skills tended to be more even with advanced (and also intermediate) students than with beginners, which is also in line with the previous research results on the structure of foreign language proficiency. The studies also suggest that the English reading test may not produce results that are entirely in line with the other English tests, especially for the C1/C2 level learners.
Existing evidence also indicates that the DIALANG self-assessment instruments have a highly consistent internal structure but that overall and detailed self-assessments, while clearly related, produce somewhat different results.

There are at least two types of external criteria against which the criterion-related or concurrent validity of DIALANG tests has been studied: learners’ self-assessments and certain other language tests or teacher assessments of the informants. Also, the performance on DIALANG tests of groups of learners who are known to differ from each other in terms of language proficiency can be used as concurrent (or predictive) validity evidence.

The studies reviewed in this paper (the present author’s studies; Alderson 2005; Jaatinen 2005; Peschel et al. 2006) indicate that DIALANG tests correlate moderately but significantly with such external criteria as nationwide school leaving examinations, teacher grades, and self-assessments by learners. It is difficult to draw specific conclusions from these correlations other than we should expect them to correlate to some extent if they and DIALANG are reasonably valid and reliable operationalizations of foreign language proficiency. To find that they indeed correlated significantly with each other is thus reassuring at a very general level. In the case of the Finnish Matriculation examination the pattern of correlations may suggest something more specific, however: The highest correlations were found between the exam and the DIALANG reading and listening tests, which are also areas of emphasis in the Matriculation examination, and this appears to indicate that the DIALANG comprehension tests manage to tap specific aspects of comprehension that are separable from the other aspects of language proficiency.

Perhaps a more robust indicator of the construct validity of DIALANG tests is the fact that the tests appear to be able to separate learners who are known (or strongly expected) to be at clearly different proficiency levels because they have studied the language for different lengths of time and/or clearly different number of hours. This was very clear in the studies on learners of English (and Swedish) in Finland: students who only had completed the comprehensive school achieved distinctly lower levels in DIALANG than students who had also completed the upper secondary school (gymnasium; see Jaatinen 2005 in particular). A similar finding between the length of language studies and test results was also found in the DIALANG pilot test data for English.
The studies reviewed in this paper also shed light on language learners’ ability to do self-assessment and the validity of DIALANG self-assessment instruments. Alderson (2005) and Alderson and Huhta (2005) found out that the correlations between DIALANG self-assessment statements and DIALANG language tests in the pilot testing data were moderate in magnitude (about .5) and statistically significant. Alderson (2005) suggested that learners may be better at assessing their productive skills (i.e., writing in the case of DIALANG) than their comprehension, presumably because they normally get more feedback on their writing and speaking.

In contrast, the two small-scale studies with learners of French (Desroches et al. 2005; Demeuse et al. 2006) found only rather weak or non-existent correlations between DIALANG self-assessments and an external (French) examination. The small number of learners and the apparent inability of the beginners in the studies to self-assess their proficiency appear to be the likely reasons for these findings.

More detailed analyses of the English pilot data by the present author in this paper suggested that, overall, the learners tended to overestimate their reading and listening skills but slightly underestimate their writing skills. The pattern varied slightly depending on the skill in question but in general, a significant number of learners at A1 to A2 (sometimes also B1) overestimated their skill, often by two or more levels. The match was best at around B2, whereas the C1-C2 readers tended to underestimate their skills, especially in writing and listening.

The English pilot test data gave only limited support the hypothesis that learners coming from different cultures and language backgrounds differ in terms of their ability to do self-assessment. Most of the correlations did not vary very much across the language groups, at least for writing and reading. Listening had more variation, however, and some language groups appeared to overestimate their level (although they did so rather consistently because the rank-order correlations were still quite robust). Learners with university or other higher education did not differ very much in their ability to self-assess from learners with lower levels of education, expect perhaps slightly in reading and writing, which are skills that learners in higher education practice a lot.
Learners’ reactions to DIALANG have been studied by linguistics students at Lancaster university (Floropoulou 2002a, 2002b; Yang 2003) and the present author in his survey of DIALANG users in Finland and Germany.

The overall reactions of the 557 learners to DIALANG in the present author’s study were mostly positive. The most often mentioned benefit of DIALANG was that it gives learners a chance to find out what their proficiency level is. Also the possibility of seeing where (in which items) their problems were, and the multi-sidedness and comprehensiveness of the system were mentioned by many learners. The findings thus support the conclusion that the system has achieved one of its main aims – provision of feedback to users on their level, as well as strengths and weaknesses in their proficiency.

The most often mentioned problems with DIALANG concerned the Vocabulary Size Placement Test and a range of technical problems. The problems with the VSPT related mostly to its reliability and some learners had in fact noticed the main reason for the unreliable results – a test-taking strategy based on guessing. However, the learners also commented on the validity of the test (see Huhta 2007a for details). The content and format of the VSPT were controversial and clearly divided users’ opinions, and while some learners thought the VSPT was the best part of DIALANG, while others found the scope of proficiency it was testing insufficient. It was also clear that the users’ found the VSPT unfamiliar as a testing format and many did not appear to fully understand its function as a placement tool. With this background of somewhat conflicting views on the VSPT in mind, we should also note the discrepancy between the reported frequency of reading VSPT score feedback, which was quite high, and the clearly lower level of perceived interest in and usefulness of the VSPT score by the learners.

It appears, on the whole, that many learner comments on the VSPT referred to the interactiveness aspect of the usefulness of DIALANG (especially guessing but also the fact that it engaged some learners in metacognitive activities). In this respect the learners’ comments on the VSPT differ from their remarks on the other parts of DIALANG because only a few other reactions could be seen to relate to interaction between test tasks and learners’ skills and other characteristics.
Learners’ reactions to DIALANG feedback showed that the traditional test result feedback – the CEFR level and item-level results – were read most frequently and were also considered the most interesting and useful parts of feedback. However, feedback on the match between SA and test results, and the advice for improving proficiency, were also quite highly rated. Also the other types of feedback (VSPT result, elaborated CEFR level descriptions, and information on SA) were considered interesting and useful by a significant proportion of respondents but the informants’ opinions were somewhat divided. Different types of questions – Likert-scale assessments, open-ended questions, and rankings of feedback types – all produced basically the same results about the users’ perceptions of DIALANG feedback.

Feedback on the match between self-assessment and test result was considered one of the most interesting and useful types of feedback by the informants in the survey. Only the traditional types of test feedback were considered better. The informants were obviously unfamiliar with feedback on SA and many were unsure about its usefulness. Women tended to consider SA feedback slightly more useful and interesting than men. The youngest learners (under 18-year-olds) tended to react to the SA feedback more negatively than the (somewhat) older learners. The level of language proficiency or experience in learning languages were not related to the informants’ views on comparative SA / test result feedback but the less proficient and less experienced learners tended to regard the more general information about SA and language proficiency as more useful than the more experienced or more proficient learners.

When the informants in the survey were asked if their DIALANG results were what they expected them to be, two thirds of those who replied said yes and about a third felt they had received unexpectedly high or low, or mixed, results, either when they were comparing the results with their self-assessed level in DIALANG or (probably more often) when they were comparing the descriptions accompanying their CEFR level results with what they felt they were able to do in the language. It was somewhat more common to get lower test results, and some learners’ elaborations suggested that low VSPT results were the cause of some such views.

The author also studied the learners’ views on the advantages of self-assessment in DIALANG with the help of a sorting task that resulted in a perceptual map showing how groups of informants structure the concepts – advantages of SA in this case. Two groups did
the sorting task: language education experts (language teachers) and non-experts (students specializing in non-language subjects). The concept maps of the two groups had several similarities. The connection between SA and test result – a unique feature of DIALANG – emerged as a clearly distinct, and important cluster of advantages in both group’s sorting data. Other such clusters of concepts included the high general quality of both DIALANG and its SA instruments, the international, expertise-based nature of the SA system, and the increasing awareness it creates of the nature and structure of language proficiency among its users. The motivating aspects of SA were, however, classified somewhat differently by the groups.

Floropoulou’s (2002b) study at Lancaster University indicated that the learners’ culture was clearly related to their opinions on self-assessment and DIALANG in general. While most considered self-assessment worthwhile, Chinese students liked SA more than Greek students. She also found that some informants’ view of their language proficiency changed during the study and that they could identify their strengths and weaknesses better. Yang (2003), also at Lancaster University, discovered that her informants’ reactions to DIALANG feedback appeared to vary depending on their goals and motivation for learning the language, and on their views about feedback and language tests. Overall, she found out that most of the subjects reacted positively to DIALANG feedback.

The large-scale Finnish survey of 15-year-old comprehensive school students also provided some support to the assumptions incorporated in the DIALANG feedback on the possible reasons for mismatch between self-assessment and test results (Luukka et al. 2008). Real-life experience in using a foreign language was considered the most important factor in formation of learners’ idea of themselves as good vs. poor language users, followed by their experience and feedback at school (e.g. from teachers and test results). Also, comparing oneself with their peers’ skills was an important factor for certain users, particularly boys.

According to Bachman and Palmer (1996, 30), the impact of test use operates at two levels: micro and macro. The micro level concerns individuals who are affected by the test, such as learners and their teachers, whereas the macro level relates to the educational system and society at large.
The evidence we have about the impact of DIALANG on the test-taking experience and the effect of the test feedback mostly comes from the studies on learners’ reactions to DIALANG, which were just summarized above. The impact on the test-taking experience mostly concerns the VSPT, in which some users had problems but which others found an unusually interesting test-taking experience. The perceived impact of DIALANG feedback was overwhelmingly positive in the present author’s survey study; as reported above, the test result feedback was considered the most useful but also the whole range of other feedback was felt to be quite interesting and useful. The long-term effects of the feedback remain unknown, however, despite some, mostly anecdotal evidence of some users taking on board some of the advice offered by DIALANG, for example. Additionally, the information we have about the technical problems some users have experienced obviously relates to the impact on the test-taking experience.

The decisions based on DIALANG results can be divided into those made by the learners themselves and to those made by somebody else such as their teacher or institution. We do not possess much information about the decisions that individual learners make based on DIALANG feedback other than that their mostly positive reactions to it are obviously a very encouraging starting point. The decisions made by language teachers and institutions most often relate to the use of DIALANG as a placement test or even as an entrance or recruitment test, but the system is also used for other purposes and types of decision such as to practice self-assessment and test-taking (in preparation for other language tests) or as a way to improve students’ learning-to-learn skills.

The number of people affected by DIALANG worldwide can be estimated by examining the publicly available statistics on the visitors to the DIALANG website (www.dialang.org) and the project-internal information on the number of tests taken.

The number of visitors to the DIALANG website was rather modest in the first years of the system, although it rose steadily, from 2001 to 2004, when the system was officially launched (in March 2004). After that the yearly number of visitors almost doubled to about 150 000 and it seems to have remained at that level ever since.

The yearly number of test sessions has stabilized at around 200 000, which suggests that the number of learners who take DIALANG tests is about 50 000 – 100 000 per year, and the
number of tests taken each year is somewhere between 100,000 and 600,000. Since the yearly number of test sessions appears to have been around 200,000 since 2004, the total number of learners who have taken DIALANG tests between 2004 and 2008 is probably around 250,000 – 500,000, and the number of tests taken around 500,000 – 3,000,000.

**English** is by far the most popular test language in DIALANG: probably a half of all tests taken are tests of English. French and German come next, followed by Spanish, Dutch, Swedish, Finnish, Italian and Danish. The first four languages are also among the most widely studied languages in Europe. Most visitors to the DIALANG website come from the **Netherlands, Germany, France, and Finland**, where DIALANG has been promoted more than in most other countries. This also suggests that most DIALANG tests are taken in these countries.

**Listening** appears to be the most popular skill tested in DIALANG followed by reading, writing, and vocabulary. Structures tests seem to be somewhat less popular than the others. This probably reflects the somewhat different skill profile available in DIALANG compared with the other free, Internet-based language tests.

DIALANG has had a significant **impact on language testing research**. Before DIALANG, systematic work on diagnostic testing and assessment of second and foreign language proficiency was quite rare. Work on DIALANG has alerted the testing community to the fact that we actually know very little about diagnosing foreign language proficiency. The concrete development work and the operationalization of different ideas relating to diagnosis and feedback in DIALANG provide interesting starting points to future research.

Probably the most significant and promising line of research emerging in post-DIALANG era are the studies that combine language testing and second language acquisition research in order to understand, in particular, how proficiency develops across the CEFR levels and which linguistic features or combinations of features characterize these functionally defined levels. Another interesting strand of research into diagnosing language proficiency is the detailed examination of item characteristics and test-takers’ skills carried out at the Educational Testing Service.
DIALANG has also played a somewhat different role in advancing science as far as language education is concerned: It has been used as a research tool in several studies where the researchers have needed a reliable and valid way to gather information about learners’ proficiency. The two biggest studies that have relied on DIALANG as their main research instrument were the German HISBUS survey of German university students’ English proficiency (Peschel et al. 2006) and the Finnish Satakunta Polytechnic university study (Jaatinen 2005) on the English and Swedish proficiency of all new students entering that institution.

The high profile of DIALANG in the European official language policy arena has obviously helped it to achieve the considerable impact summarized above. Not surprisingly, then, DIALANG has also been reviewed for possible use and as a source for ideas in the planning stages of the European Commission studies of language proficiency of both students in secondary education and the general adult population in the EU.

According to Bachman and Palmer (1996), the final aspect of test usefulness – practicality – differs from the other test qualities in that it relates to availability of resources and to decisions to use or develop a particular test – or not to develop. A test is practical if the resources that it requires (time, money, work, personnel) do not exceed the available resources. Practicality is a relative concept, however, and a test may be perfectly practical for one purpose but not for another. Practicality is often at odds with the other aspects of test usefulness.

The practicality of DIALANG among its users has not been studied extensively and specifically. The studies by the present author and Floropoulou (2002a) provide a somewhat mixed picture: some users appear to consider DIALANG quite easy to operate and understand, while others face problems in either navigating the system (especially when doing it for the first time) or find it too long and overwhelming. Different reactions probably relate to different purposes of using DIALANG and to different expectations of what it can deliver and what language tests typically do and look like (see e.g Yang 2003).
Overall usefulness of DIALANG

We do not have a balanced picture of all the aspects of test usefulness for DIALANG, especially when it comes to empirical evidence. The a priori theoretical arguments incorporated in the numerous design and planning documents and implemented in the procedures used for item writing, piloting, analyzing, translating, standard setting, and linking, as well as the design of the IT system cover all the aspects of usefulness in the Bachman and Palmer framework. In that sense, the usefulness of DIALANG is likely to be reasonably good for the purposes that concern individual learners and, to an extent, their teachers. The system was not designed for institution-wide use for placement purposes and thus, by design, it does not have many features that an institution would find useful.

It is possible that the a priori planning of DIALANG was strongest when it comes to the authenticity and practicality of the system. To the extent that the CEFR manages to capture the essential elements of language use, it could be argued that DIALANG has considerable authenticity because, overall, DIALANG test content was very systematically built on the CEFR and other Council of Europe definitions of language use and proficiency. DIALANG’s main claim to practicality is based on the careful design of the software and hardware system that underlies the assessment system and allows the users to navigate through the various stages of the assessment procedure and to receive all the feedback they choose to study.

The reliability, construct validity, and interactiveness of DIALANG were obviously covered in the design documents and – probably more importantly – in the review, statistical analysis and standard setting stages of the development process. For example, uniform item writing guidelines across languages are one of the key points in ensuring that these qualities are achieved. However, as has been pointed out earlier, the project had to resort to a lot of additional sources of information for actual item writing and test design because the CEFR is not a test construction recipe. While such additions were necessary they may also have had unpredictable consequences to the quality of the system. In particular, the interactiveness aspect of test usefulness appears to be difficult to capture and ensure at the design stages of a language test.

Impact is a characteristic of a test that is not directly addressed in the documents that typically guide test design although achieving a positive impact is obviously one of the
considerations in many decisions made in any test development project. Certainly, meeting the purposes set for DIALANG – i.e., helping individual learners – was seen in the Project as a key to achieving any impact on language learners in Europe. For the most part, the \textit{a priori} plans for impact were presented in the dissemination sections of the DIALANG Project applications for funding to the Commission – and for the extensions of funding in the latter stages of the Project. It was clearly the case that the main funding body of the Project wanted it to have a significant impact on European foreign language education but probably nobody had a very clear idea of what forms the impact was going to take since what was to be developed was unlike anything that had previously been done on such a large scale. Thus, it is rather straightforward to describe the types of impact that DIALANG seems to have had so far but it is much more difficult to evaluate to what extent it has achieved the aims and impact originally envisaged for it. However, it can be said about the \textit{a priori} impact plans that there were considerable grounds to expect the system to have at least some European-wide impact because of the involvement of the European Commission in the Project and the dissemination of information by the Project about the system, for example, in numerous meetings, seminars and conferences with potential users.

What does the available empirical evidence tell us about the usefulness of DIALANG? As pointed out earlier, there is no comprehensive study into all aspects of test usefulness. The best covered areas seem to be the impact and construct validity (in terms of correlational studies and group comparisons) of DIALANG. There is also some evidence about the authenticity and practicality of the system. Empirical evidence for the interactiveness of the system is, however, quite limited.

Direct evidence of reliability, construct validity, authenticity, interactiveness and practicality mostly indicates that DIALANG is reasonably useful for the purposes it was developed, although the amount of evidence varies considerably depending on the language. There is also clear evidence of certain, specific problems in the system, most notably in the VSPT and in some technical aspects of the system (these problems vary significantly depending on the individual learner or institution). It is, however, fairly difficult to be much more precise than to say that the available evidence suggests that DIALANG is ‘reasonably’ useful, as far as those aspects of usefulness listed above are concerned. Perhaps the important thing to note at this point is the lack of evidence about truly serious problems or deficiencies with regard these aspects – with the few exceptions mentioned above.
Perhaps the overall usefulness of DIALANG can best be evaluated by reference, first, to the inevitable tension between practicality and the other aspects of test usefulness, and, secondly, to the impact that the system has had. First, DIALANG took a long time and considerable amount of money and work to develop. However, these obvious challenges to the practicality of the system did not prevent its creation, and, thus, we can say that in that sense the system passed its first, major practicality hurdle that could very well have prevented its design and completion (see Bachman and Palmer’s definition of practicality). Despite certain technical impracticalities and challenges, substantial numbers of learners, teachers and institutions and organizations use DIALANG. Interestingly, many institutions use the system for placement purposes even though it is by design not developed for that purpose and the practical problems in obtaining the results for whole groups of test takers are considerable. In spite of the clumsiness of DIALANG for such purposes, it nevertheless enjoys considerable popularity as a placement tool across Europe – because the overall usefulness based on the construct validity, authenticity etc of the system is apparently regarded by these users so high that it overcomes whatever impracticalities are involved in using it.

The second approach to estimating the overall usefulness of DIALANG is to consider the impact it has had. The survey by the present author of DIALANG users in Finland and Germany showed that a significant majority of the informants reacted positively to the system. The information they provided also shed light on the reasons for their positive views. Usually, the reasons given by them related to the stated aims of the system (e.g. helping them find out about strengths and weaknesses in their proficiency), which suggests that the impact of DIALANG at the level of individual learners is more or less what it was expected to be – at least for most of the learners surveyed.

Since such studies as the one carried out by the present author are not based on representative samples of users, they cannot inform us about the extent to which DIALANG is known and used in Europe and beyond, which is another potential indicator of a test’s impact. Fortunately, we have access to two important statistics: the number of visitors to the DIALANG website and the number of tests initiated in the system. These show that hundreds of thousands of learners have taken DIALANG tests or visited the website – and even more must have been informed about the system by their teachers and institutions. Such large-scale use indicates that considerable numbers of users find DIALANG useful enough for their
needs and purposes, both individual and institutional. It is difficult to say if the present number of people impacted by DIALANG matches the unspecified expectations of the European Commission and/or the DIALANG partnership. Certainly, the potential target group of DIALANG – European language learners above the age of 16 or thereabouts – is vastly more extensive than those who appear to have used DIALANG. Nevertheless, it appears that DIALANG has made a significant contribution to language education in Europe also at the practical level in tertiary and adult education, especially in certain countries such as the Netherlands, Germany, France, and Finland, but also elsewhere. Significant numbers of learners keep using DIALANG even after the official end of the Project in 2004, and importantly, the system keeps running with only the most minimal level of maintenance.

In the longer term, however, the most important type of impact of DIALANG is one that was not anticipated in the early phases of the Project – certainly, it was not listed among the deliverables of the Project in the applications for funding. This is the scientific impact of DIALANG on applied linguistics via language testing research and development. Before DIALANG, the diagnosis of second and foreign language proficiency was an almost totally neglected area, and it was not even fully realized how underdeveloped the field was compared to the diagnosis of first language development and problems, for example. After DIALANG, there are probably no serious language test developers and researchers who have not heard about DIALANG and who are not aware of its significance to the field, even if they personally may not work on diagnosing proficiency.

At the time of writing the present paper, the future of DIALANG as a freely available system for anybody who cares to use it is uncertain. Attempts to find a permanent home, maintainer and further developer have so far failed, in spite of interest shown, at times, by different organizations and commercial enterprises. The permanent closure of DIALANG is a real possibility – and it would be a great loss to many institutions and individuals who have found it useful – but its scientific impact alone has already earned it a place in the ‘hall of fame’ of language testing.
APPENDIX 1

Learners’ overall views on the best and worst aspects of DIALANG

The following is a description and analysis of the responses of the 557 learners using DIALANG who were surveyed by the author in 2004-06. The study used a fairly extensive questionnaire that elicited the users’ views on DIALANG in general and on all different types of feedback on offer in the system. The results of the study were made use of in the three empirically based articles comprising the present study (Huhta 2007a, 2007b, submitted), in the present paper, and in a conference presentation by the author (Huhta 2006) but no comprehensive report on the survey has been published. This Appendix contains an account of the users’ responses to the key open-ended questions in the questionnaire on the overall quality of DIALANG. The information presented in this Appendix provides background information that hopefully helps the reader to contextualise the empirical studies reported in Huhta 2007a and 2007b, in particular, in the overall survey of DIALANG users that has not been published to date, although it was referred to as Huhta (forthcoming) in the article on the VSPT. Appendix 1 and Appendix 3 also report on some findings of the survey study that are not covered in any of the empirical articles and these findings are thus intended to complement the evaluation of the usefulness of DIALANG carried out above, in the main part of this synthesis document. In other words, Appendices 1 and 3 serve to complement the three empirical articles and the present synthesis paper.

Description of the main results on the overall quality of DIALANG

The following is an account of the 557 users’ responses to six key questions concerning the quality of DIALANG as a whole and the quality of its feedback. These questions were all open-ended, and the users responses were coded into a number of categories (see the text below for details on the categories used). The analysis is thus a very straightforward content analysis that focuses on (1) the content of the responses, which then determined into which category /-ies the responses were coded, and (2) the frequency of responses in each category. The latter will be reported as percentages of the 557 respondents who mentioned something related to a particular category. In calculating the percentages, the users’ responses to any of
the three questions were combined, because a learner could say something positive about DIALANG in a question focusing on problems, or vice versa. What was of interested, was to find out what proportion of learners mentioned particular things, irrespective of the actual question in which the response was given. Some respondents obviously mentioned more than one thing at each question, in which cases the respondent’s replies were included in all the relevant response categories (e.g., a respondent could say that self-assessment and freedom of choosing the time of testing were good points about DIALANG, and his/her responses would be included in two categories, i.e., ‘self-assessment’ and ‘freedom / flexibility’).

The six key questions reported on here are:

(1) When you think about Dialang as a whole, what was best / most interesting / most useful about it?

(2) When you consider Dialang as a whole, what was worst / least useful about it, or something that still needs further development?

(3) Is there something that should be added to Dialang, or something that should be different? Other comments?

(4) Please describe why you think that those parts that you marked as the most useful in the above table are useful for you:

(5) Please describe why you think that those parts that you marked as the least useful in the above table are not useful for you:

(6) Have you suggestions on how we could improve Dialang feedback?

Questions 1-3 were included in all questionnaire versions. Also questions 4-6 were included in almost all versions except for a couple of groups of students who were presented with a shortened version of the questionnaire in the early stages of data collection. Questions 4 and 5 were preceded in the questionnaire by a table listing all types of DIALANG feedback; the respondents were asked to mark the three most useful (and the three least useful) types of feedback in the table (those rankings are not reported here, however, only the informants’ free responses to questions 4 and 5).
What was best about DIALANG?

Overall, the learners surveyed in the study had a very clear positive opinion of DIALANG. Almost all (95%) of those who replied to the questionnaire answered the question about the best features in DIALANG (question 1 above; i.e., 528 informants out of 557). Only 4 (0.7%) of those who replied to this question said there was nothing good about the system, which contrasts with the responses to question 2 about the negative aspects of DIALANG (see below for details).

The feature that was most often mentioned as the best thing in DIALANG was that the learners could find out their level of proficiency by using the system – about half of all respondents in the survey explicitly referred to the level in their responses. When we look at the feature that the learners mentioned first in their list of good things about DIALANG (or as the only feature, if they only mentioned one), this type of response stands out as the most frequent answer (23%). Further 26% of them replied by stating something quite similar, although using somewhat vaguer expression such as “You can find out how good you are” or “how good your skills are”, or simply “You can test your skills”. Thus, the chance of testing oneself, finding out about one’s proficiency, and in particular, about one’s level of proficiency were clearly the most important positive feature of DIALANG, in the opinion of the users surveyed here.

The second group of features that a substantive number of users singled out as being best in DIALANG consisted of two closely related sets of replies. The first was a more general statement about the opportunity to see where one had errors or to see what the correct answers were (26% of all users). The other was the chance to see the item level results immediately, i.e. the immediate item review function (23%). About half of the users thus appreciated the most detailed type of feedback on offer in the system, although it was seldom the first, or only, thing they mentioned.

The many-sidedness of the system (19%) and the whole, integrated concept behind the system (5%) were also mentioned quite frequently. The former refers to the variety of different languages, skills and tasks on offer to the user, whereas the latter seems to express the same idea but from a slightly different point of view.
The advice on how to improve your language learning was mentioned by almost every fifth respondent (18%). Also, quite a few stated in more general terms that the system helps them to improve or learn languages (12%) or that it enabled them to see their strengths and weaknesses so they would know what needs improving (11%). It is sometimes difficult to distinguish between these three response categories but the main point here is to show the main types of things that many of the users found good, useful and interesting in DIALANG rather than make any finer distinctions between obviously closely related types of answers.

Table 1. The best features in DIALANG and its feedback

<table>
<thead>
<tr>
<th>Features mentioned by users</th>
<th>Mentioned</th>
<th>Mentioned first / only</th>
</tr>
</thead>
<tbody>
<tr>
<td>You can find out your level</td>
<td>283</td>
<td>51% 23%</td>
</tr>
<tr>
<td>You can find out how good you are / how good your proficiency is / You can test your skills</td>
<td>145</td>
<td>26% 17%</td>
</tr>
<tr>
<td>You can see the correct answers / where you had errors</td>
<td>145</td>
<td>26% 1%</td>
</tr>
<tr>
<td>Immediate feedback on items</td>
<td>129</td>
<td>23% 5%</td>
</tr>
<tr>
<td>Many-sidedness of the system (languages, skills, tasks)</td>
<td>103</td>
<td>19% 11%</td>
</tr>
<tr>
<td>You get advice</td>
<td>102</td>
<td>18% 2%</td>
</tr>
<tr>
<td>Self-assessment / self-assessment feedback</td>
<td>94</td>
<td>17% 5%</td>
</tr>
<tr>
<td>You can learn / practice (when doing Dialang tests)</td>
<td>78</td>
<td>14% 3%</td>
</tr>
<tr>
<td>One or all of the Dialang tests or tasks themselves</td>
<td>72</td>
<td>13% 6%</td>
</tr>
<tr>
<td>That you get feedback (in general)</td>
<td>71</td>
<td>13% 4%</td>
</tr>
<tr>
<td>Helps to learn / improve language (in general)</td>
<td>66</td>
<td>12% 0%</td>
</tr>
<tr>
<td>You can see your strengths and weaknesses / what to improve</td>
<td>59</td>
<td>11% 1%</td>
</tr>
<tr>
<td>Clarity of instructions or information</td>
<td>38</td>
<td>7% 2%</td>
</tr>
<tr>
<td>Ease of use</td>
<td>36</td>
<td>7% 3%</td>
</tr>
<tr>
<td>Neutral / objective / realistic / real / correct / reliable information or result</td>
<td>35</td>
<td>6% 1%</td>
</tr>
<tr>
<td>It motivates / is encouraging / makes you believe in yourself</td>
<td>32</td>
<td>6% 0%</td>
</tr>
<tr>
<td>The whole system / whole concept / new idea / comprehensive (integrated) approach</td>
<td>27</td>
<td>5% 2%</td>
</tr>
<tr>
<td>Concrete feedback / information</td>
<td>20</td>
<td>4% 0%</td>
</tr>
<tr>
<td>International / common standard or yardstick for comparison</td>
<td>14</td>
<td>3% 1%</td>
</tr>
<tr>
<td>Quick way to test / quick results</td>
<td>11</td>
<td>2% 0%</td>
</tr>
<tr>
<td>You can monitor your progress with Dialang</td>
<td>11</td>
<td>2% 0%</td>
</tr>
<tr>
<td>Can be done any time / anywhere / by anybody / it's freely available</td>
<td>10</td>
<td>2% 0%</td>
</tr>
<tr>
<td>The fact that it is computerised</td>
<td>8</td>
<td>1% 0%</td>
</tr>
<tr>
<td>No time limit</td>
<td>7</td>
<td>1% 1%</td>
</tr>
<tr>
<td>Free of charge</td>
<td>6</td>
<td>1% 0%</td>
</tr>
<tr>
<td>Individual / personal feedback</td>
<td>5</td>
<td>1% 0%</td>
</tr>
<tr>
<td>Can be done independently / without a teacher</td>
<td>4</td>
<td>1% 0%</td>
</tr>
<tr>
<td>Other reason</td>
<td>119</td>
<td>23% 5%</td>
</tr>
</tbody>
</table>
Interestingly, 14% of all respondents replied that they appreciated being able to actually learn or practice a language when taking and studying the DIALANG test tasks. Many test takers did not thus appear to see DIALANG as (only) a language test but also as a context and opportunity to learn.

The DIALANG tests and test tasks themselves were also specifically mentioned as being good parts of the system (13%). Of the tests, listening and the Vocabulary Size Placement test were singled out as particularly good and interesting. When test tasks were referred to, adjectives such as interesting, realistic and varied were typically used.

Clarity and ease of use were both mentioned by 7% of respondents, which is somewhat reassuring, as much attention was paid on these qualities of the system in the design stages.

**What was bad in DIALANG? What could be improved?**

Next, the learners’ responses to two open-ended questions about the problems and areas for improvement in DIALANG are analysed. The two questions are treated together because they are related: there is some overlap in the wording of the questions, and many learners who reported some feature in the system to be problematic also identified that feature as something that needed improving. It should be noted, however, that quite a few learners replied only to one of these two questions, and some suggestions for improvement did not seem to mean that the learner regarded the particular point as a problem. For example, suggestions to add new languages to the system seem to imply a ‘problem’ in the system only indirectly.

It can be noted that whereas almost all respondents (94%) had something positive to say about DIALANG when directly asked, considerably fewer of them replied when they were asked to say what was bad or problematic in the system (65%) or what could be added or changed in the system (42%). The learners’ positive attitude to DIALANG was also evident in the fact that 17 (or 3%) stated there was nothing bad about the system when they were asked about its problems, and 26 (almost 5%) replied nothing needed adding to or changing
in the system, and further 2% said they did not know what to reply to the question about how
to improve the system. (For the sake of comparison, only 4 learners had replied to the
question about DIALANG’s good sides by quoting some problem with it.). Furthermore, the
list of different advantages mentioned by the respondents was far longer than the
corresponding list of problems.

The above is in line with the results of the analysis of the respondents’ Likert-scale
responses: overall, the learners surveyed here regarded DIALANG in positive way. Because
the sample is not statistically representative, this result, although encouraging, cannot be
confidently generalised to all users of the system in the two countries surveyed. However,
for the main aim of this study was to gain a more detailed understanding of the ‘strengths
and weaknesses’ of DIALANG and the learners’ reactions to the feedback that the system
provides, the present data inform us about the most salient benefits and problems of the
system, in the learners’ opinion.

Two things stand out in the learners’ responses on problems and areas for improvement in
DIALANG: content of the test or system, and the Vocabulary Size Placement Test. The
VSPT was clearly viewed as problematic whereas test content was an area where learners did
identify some problems but also suggested quite a few improvements.

**Test content**

Test content is a rather wide category so a more detailed account of what is included in it is
in order. Of the 152 respondents (27% of all) who mentioned content problems or suggested
improvements to content, a quarter (36) wanted to have more tests tasks available in the
system. Many of these reported that when retaking particular tests, they were given exactly
the same test tasks as before. This observation is quite accurate and reflects the state of the
DIALANG system: only one suite of tests was developed for each language in the project
that created the system. Thus, for example, the reading test of English in DIALANG has
three versions: easy, intermediate, and difficult. There is only one easy, one intermediate and
one difficult test versions, and furthermore, there is extensive overlap in terms of items
between the easy and intermediate versions on the one hand, and between the intermediate
and difficult versions, on the other hand. This obviously severely limits the usefulness of
DIALANG – in its present form, at least – for repeated testing, which some learners might want to do in order to monitor their progress.

Table 2. Worst aspects of DIALANG and suggestions for its improvement

<table>
<thead>
<tr>
<th>Features mentioned by users</th>
<th>Worst in DIALANG</th>
<th>Suggestions for improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mentioned</td>
<td>Mentioned first / only</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>%</td>
</tr>
<tr>
<td>Test content</td>
<td>85</td>
<td>15</td>
</tr>
<tr>
<td>Vocabulary Size Placement Test</td>
<td>81</td>
<td>15</td>
</tr>
<tr>
<td>Technical problems</td>
<td>63</td>
<td>11</td>
</tr>
<tr>
<td>Some part / aspect of feedback</td>
<td>54</td>
<td>10</td>
</tr>
<tr>
<td>Length (of tests, feedback or instructions)</td>
<td>32</td>
<td>6</td>
</tr>
<tr>
<td>Interface or layout</td>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>Functioning of the system</td>
<td>23</td>
<td>4</td>
</tr>
<tr>
<td>Self-assessment</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>Instructions</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>Answers not accepted (scoring)</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>Navigation</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>Other feature</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>85</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>88</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>88</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>33</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Percentages in Table 2 have calculated from all the 557 respondents.

Table 3. Worst aspects of DIALANG content and suggestions for its improvement

<table>
<thead>
<tr>
<th>What is wrong with content / How it could be improved</th>
<th>n</th>
<th>% of all 557 respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test content in total</td>
<td>141</td>
<td>25</td>
</tr>
<tr>
<td>More tasks needed (same tasks repeated in the tests)</td>
<td>36</td>
<td>7</td>
</tr>
<tr>
<td>Test of writing</td>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>Difficulty level not right</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>Add new languages</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>No speaking / add speaking</td>
<td>9</td>
<td>2</td>
</tr>
</tbody>
</table>

The second most frequent complaint or suggestion for change concerned the test of writing (19 or 3%), which was criticised, not surprisingly, for not letting the user to actually write very much. This is also a design feature of the system, and the Project was very much aware of the controversial nature of an attempt to test writing skills with tasks that can
automatically be scored by a system that does not contain any built-in artificial intelligence that could deal with longer responses. However, since DIALANG was designed to be a diagnostic rather than, e.g., a certification test, the use of indirect writing tasks that can be scored in a rather mechanical way is defensible, at least until we know more about diagnosing foreign/second language writing (Alderson, 2005; Pelgrum, 2001). If the writing tasks can be designed to focus on particular aspects of writing, it may indeed be possible to use such, even indirect tasks for diagnostic testing. It is also possible to design automatically scorable, indirect writing tasks that come quite close to simulating real-life writing tasks. An example of such a task would be a letter writing tasks where the test taker has to fill in one or more gaps in an otherwise complete letter that he/she on the screen. The gaps could target specific points in the letter that are essential in terms of, say, sociolinguistic appropriateness, such as the salutation and ending phrase. Or the gaps could target certain vocabulary and grammar items. In any case, the consequence of the decision to use indirect writing tasks in DIALANG resulted in many writing tasks closely resembling regular grammar, vocabulary or reading items.

There are in fact two somewhat different conclusions that one can draw from the above finding. On the one hand, it was somewhat surprising to find out that only 3% of the informants in the study singled writing out as an aspect of DIALANG in need of improving. It may be that an average test taker is more or less happy with the way in which DIALANG tests writing, perhaps because the item types they encounter are the familiar ones – multiple choice and gap-filling – and they may not easily see beyond that. It should be mentioned here that the justification for using indirect writing tasks in DIALANG is not available to the users on the DIALANG website or in the system help files. Thus, their apparent lack of criticism of the way writing is tested in the system is very unlikely to be based on an awareness and acceptance of the argument that such test methods may be defensible for diagnostic purposes. The second conclusion from the finding is that an explanation of the reasons why writing is tested the way it is might in fact be needed. At least some test takers were unhappy with the fact that they needed to write next to nothing in order to pass the writing tests. Clearly, these learners were judging DIALANG against the regular writing tasks that they had encountered in their lives before and failed to see why DIALANG was using such a weird looking approach to testing writing.
Some learners also felt the **difficulty of the tasks was not right for them**. Usually, those who mentioned this complained that the tests, or part of the tests, were too easy for them. These are interesting comments because an attempt is made in DIALANG to provide the test taker with the level of test that suits his/her skills best (either the easy, intermediate, or difficult test version). As explained in e.g. (Alderson & Huhta 2005) DIALANG uses a Vocabulary Size Placement Test and self-assessment as a way to guide the user to the most appropriate level of test. Clearly, this does not succeed in all cases.

There are at least two major issues when it comes to the difficulty of tests in DIALANG that the present author and the Project in general are aware of. The first has to do with the availability of items for all CEFR levels tested, and the second concerns the accuracy of the placement procedures used in DIALANG. Let us first deal with the issue of the difficulty of items – and, hence, the tests – available in DIALANG. Predicting the difficulty of comprehension items can be quite difficult; indeed Alderson (Alderson 2005) found out the item writers of English in DIALANG were often wrong in predicting the level of the items, although there was a positive overall correlation between the item writers’ estimate of the difficulty of the items and the empirical difficulty based on piloting the items with learners. Further analyses of the pilot tests and the results of the standard setting exercises within the Project both indicated quite clearly that, in general, the current DIALANG tests have very few C1 or C2 level items. Thus, even the most difficult versions of the three levels of tests available in DIALANG contain few items that really challenge the language abilities of advanced learners whose proficiency is at C1 or C2. Consequently, the tests may feel somewhat easy for such advanced learners, and may explain some of the comments on the general easiness of the tests.

It is however likely that the negative comments on too easy tests are based on the failure of the placement procedure to assign the learner to the optimal test version. The most likely source of problem in this respect is the Vocabulary Size Placement Test (VSPT). The issue is thoroughly examined in one of the articles (Huhta 2007b) that comprise the present thesis, and it is also briefly summarised elsewhere in this document / article. Suffice it to say that there is indeed enough reason to believe that the placement procedure does not work for learners using a particular test taking strategy for the VSPT, which results in them receiving way too low scores, and thus, them being given too easy a test.
In the suggestions for improvement, in particular, two further types of comments on test content are worth mentioning. Some learners wished there were more languages in addition to the 14 currently in the system (e.g. Russian, Latin, Chinese, Japanese, Arabic, and sign language). There is one obvious gap in DIALANG, namely the lack of speaking tests, and this was commented on by some learners, although you might perhaps have expected that more than 9 of the 557 respondents paid attention to this gap. It is possible that since the information about what is available in DIALANG and what is not is easily accessible to the users on the website and in the test selection menu that they face in the early stages of taking DIALANG tests, they do not perceive this as a particular problem – at least it is not a (unpleasant) surprise to them. As far as these test takers are concerned, many of them were informed of the lack of speaking tests in the pre-test information session given by the current author or their teacher.

**Vocabulary Size Placement Test**

The Vocabulary Size Placement Test (VSPT) topped the list of problems together with test content. Both were mentioned by 15% of all respondents as one of the least good aspects of DIALANG. Full 13% mentioned the VSPT as the first, or the only, problem in their list of issues with the system, which was in fact higher than the corresponding figure for test content (see Table 2).

The reasons for why so many users were critical of the VSPT are explained elsewhere in this article and in a separate article (Huhta, 2007b).

**Technical problems**

After the content of the tests and the VSPT the next most problematic thing about DIALANG were different kinds of technical problems: A total of 11% (63) of the informants singled them out, and they were also quite common as the only or the first problem mentioned (11%).

The technical problems mentioned included such absolutely crucial issues as failures to install or start the programme when the users first attempted to do that at home or that the system crashed or hung in the middle of a test session with the loss of all information accumulated by that point. Other, less serious, but still annoying problems included a failure
of part of the item content to display on the user’s screen, which made answering to the particular item more or less guesswork. Also, problems with audio recordings were quite common (audibility in general and volume level), although some of these are in fact due to problems with the user’s computer settings rather than with the DIALANG software. In any case, they may indicate issues in the way audio recordings are presented in an assessment system intended for a lay user of computer technology.

As with any other Internet-dependent computer programme, DIALANG has had problems with the reliability and speed of service across the Internet, as indicated by several users’ comments. Sometimes a network failure somewhere between the user’s computer and the DIALANG server(s) cut a test session short. Or, perhaps more commonly, users in certain areas or local networks experienced slowness in the delivery of test items and other content. Obviously, the listening tests with their big audio files are the first to suffer in such circumstances.

None of the technical problems listed by the informants of the present study are news to those who worked in the development of the system for years, myself included. In fact, the majority of the messages to the DIALANG project members from users of DIALANG across the world during the operational period of the system in 2001 – 2004 (and beyond that, in fact) concerned various technical questions or issues. For example, the problem of certain items not displaying properly on the computer screen in certain Windows XP systems was well known to the developing team but turned out to be unsolvable (at the time, at least) because of incompatibility between Java, the programming language of DIALANG, and Windows XP. Years of experience in advising DIALANG users suggests that in the real world, technical problems are probably more common and more serious than the results of the current study indicate. The informants in this study had obviously managed to use DIALANG either in their educational institution or at home, so they had already overcome the first major issue – installation of the system. Also, the bandwidth of most educational institutions from which the informants came from was sufficiently good, which probably prevented some of the problems due to slow connections from emerging.
Problems with feedback

DIALANG takes considerable pride in being the first large-scale language assessment system that provides such a wealth of different types of feedback and information to its users. With this background in mind, it is particularly interesting to examine the users’ reactions to the feedback and if they found any problems or deficiencies in it. In the following, the users comments about the following parts of feedback are discussed:

- test result (CEFR level + description)
- review of individual test items either immediately or after the test
- self-assessment and test result feedback and comparison
- information about self-assessment (reasons why SA and test may not match)
- extended descriptions of the CEFR levels for reading, listening, and writing
- advice on how to improve reading, listening, and writing

Feedback from the VSPT is not covered here as it is already discussed elsewhere and because in the users’ often integrated in their comments the VSPT test and the feedback they got from it.

Ten percent (or 54) of the informants mentioned some problem with feedback, and although only 5% offered suggestions about its improvement, quite a few of the responses about problems with the feedback also contained advice to the developers of the system on how it might be improved.

Clearly the most common complaint, and conversely, the most obvious direction for improvement was the generality of feedback: Much of it was criticised to be too general and self-evident. This concerned, in particular, the test result (just a level; the user already knew his/her level), the SA and test comparison (too vague and general to be of use) and advice (too self-evident). Thus, the most common wish was to get more detailed and more extensive feedback than what is on offer in the present version of DIALANG (about 9% of all respondents wanted this). They also wanted to get more personal feedback (3%). More specific wishes included explanations for the different wrong answers (presumably in multiple choice items), not just the information about the correct answer. In line with the
demand for more detailed feedback, several users (2%) wished to get more detailed advice on how to improve their proficiency.

The review of individual test items and the user’s answers to the items divided the informants’ views to some extent. Depending on their preference, those who preferred to see the result of their answer right after replying to an item (immediate item review) considered the possibility of reviewing the items after the test was over unnecessary and not useful. Those who preferred the post-test item review, for their part, considered the immediate item review too intrusive and distracting.

To conclude, the intention of the designers of DIALANG was to offer a wide range of different kinds of feedback. Clearly, that has been a major reason for the generally very positive reception of DIALANG among the participants of this study, and more generally. However, as the users comments reported here indicate, there is still quite some way to go. Although there are aspects of feedback that are somewhat personal and detailed, that is clearly a major area in which work is needed in the design of assessment systems such as DIALANG.

**Table 4.** Learners’ views on how DIALANG feedback could be improved

<table>
<thead>
<tr>
<th>How feedback could be improved</th>
<th>n</th>
<th>% of all 557 respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>More detailed / more extensive feedback</td>
<td>50</td>
<td>9</td>
</tr>
<tr>
<td>More personal feedback</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>Reasons for wrong answers</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>Better / different functionality of feedback</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>Better / more detailed advice</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Access to feedback afterwards / after all tests</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Easier access to advice</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Shorter feedback</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Need teacher to interpret feedback</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

**Length**

DIALANG is a fairly extensive system: Taking language more than one language test in it and studying all the feedback it offers takes a considerable amount of time. Thus, it was not surprising to see a fair number of critical comments on the length of the tests, feedback, and, perhaps more surprisingly, instructions in DIALANG (6% of the informants). The issue of
length is interesting in that we just saw above how quite a few users wanted to get more feedback. Length, thus, appears to divide opinions to some extend. Probably depending on your reasons for taking a test such as DIALANG and your previous experience with taking tests, especially on the Internet may affect your perception of DIALANG being either suitable or too long in terms of time it requires from the user.

DIALANG is a fairly extensive system for a good reason. First, it aims at being a reliable and valid measure of the language skills of reading, writing, listening, structures, and vocabulary included in it. This alone requires a certain minimum number of items to be used in each test. Secondly, it aims at being a diagnostic language tests. In order to diagnose, it is necessary that the test attempts to test different aspects or parts of the skill area in question (cf. (J. C. Alderson, 2005), and in order to get even a reasonably accurate picture of such sub-skills or sub-areas requires that each of them is tapped by enough items. Thus, truly diagnostic tests are bound to be long and extensive, and probably much more so than the current version of DIALANG.

There is one use of DIALANG, where its length may indeed be a handicap: placement. The use of DIALANG for placement purposes in educational institutions is probably the most common way of using it, at least when the number of test takers is considered. The fact however, that it is so widely used for this purpose may indicate that institutions value an instrument that is available and dependable, although it may be longer than they ideally would like to have.

Other issues and problems

The interface of DIALANG and the layout of the screens and items was something that caused a negative reaction in some users (5%). Also, the way the system functioned in general was an issue with some users; since it consists of several parts and a typical test taking session has several parts in it, it was a challenge to the designers to make it clear to the users what the main steps were and how to proceed from one step to the next. As is the case with most computer programmes, a user needs a certain amount of practice with DIALANG to navigate confidently in the system.
Self-assessment was also something that was not to everybody’s taste, as about 4% of the respondents regarded it as one of the less good aspects of DIALANG. Some of them apparently did not like to do self-assessment at all, others found the particular approach used in DIALANG not useful for their purposes. Still others found something wrong in the instructions available in the system (3%) or in the scoring of open-ended questions. The latter is an issue that is very hard to tackle in a foolproof way in any automated scoring system not using an intelligent scoring / rating algorithm. In DIALANG, all acceptable answers to a short-answer item have to ‘hard-wired’ into the system, i.e., all acceptable answer options (including misspelled but acceptable variants of the answers) have to be manually fed into the scoring part of the system. It is almost an impossible task to arrive at a 100% comprehensive scoring key for reading and listening items, in particular, even if based on extensive piloting, as was the case with some of the languages in DIALANG. Thus, inevitably, some users will find, if they use one of the item review functions offered by the system, that a (possibly) good answer to a particular item is not accepted by the system.

43 (8% of all respondents) stated that feedback in DIALANG was good enough as it was, and further 11 (2%) said they did not know how it might be improved.
APPENDIX 2

Own contribution in the two joint articles

The two joint articles included in my thesis (Huhta & Figueras 2004 and Alderson & Huhta 2005) both cover the development of DIALANG; the latter also reports some results of the piloting of the DIALANG tests and self-assessment instruments. Before giving an account of my contribution to the two articles it is thus important to specify my role and contribution in the DIALANG project that developed the assessment system.

Clearly, a project such as DIALANG is joint enterprise by dozens of participants, and even the number of those who made significant contributions to the system over several years must be 10 – 12, including the three persons whose names appear as the authors of the two above mentioned articles. All major components of the DIALANG system were developed over a considerable period of time and they were modified and revised sometimes several times (e.g., test items, wording of feedback) before incorporation into the operational system. Typically, one or two persons were responsible for designing the draft version of a component (a test item, text for feedback or instruction), which was then reviewed and revised by several other project members both before and after empirical trialling (see Alderson 2005 as the most detailed description available of the development of, e.g., the tests, feedback and self-assessment components).

Although all work on DIALANG was more or less collaborative and no individual could make independent decisions on anything, it is however possible to describe in broad terms the areas in which an individual made the most significant contributions. In some cases it is also possible to say who was responsible for producing the first version of particular texts or other components of the system.

The following table lists the positions that I held in the project and describes my tasks and responsibilities in each role:
Below, I will elaborate on the aspects of DIALANG that I contributed most; I focus on those aspects that are covered in the articles included in my thesis.

In terms of amount of work, my biggest contribution to DIALANG was in the quality control of the test items between their initial drafting and piloting; this took place in stages, depending on the language, between 2000 and 2004. This included the coordination of the review and revision of the draft items (together with the Dutch colleague José Noijons from CITO) in 14 languages by 14 teams of reviewers, to ensure that as many items as possible would ‘survive’ the piloting and data analyses. I was the main author of extensive item review guidelines that aimed at ensuring adequate and comparable review of items across the languages. The fact that the results of the piloting reported in Alderson & Huhta (2005) and Alderson (2005) were quite good suggests that the preceding work on the items by the reviewers and their coordinators had been successful. Thus, work on the test items...
contributed to the validity of the language tests in DIALANG, which is a prerequisite for the validity and potential usefulness of much of the feedback offered by the system.

Another important area that I worked on in DIALANG is the feedback. During the first phase of the project (1996-1999), I participated in a work group that had been created to design the self-assessment and feedback components of the system – hence its name ‘Work Group on Self-assessment and Feedback’. As a member of the group, I obviously contributed to all work on these two areas but it is probably fair to say that the points where I made the clearest individual contributions by drafting the first version of the texts or by making initiatives were the following:

(a) Creation of ‘Information about self-assessment’. The idea for this part of feedback came from me, and the first version was also drafted by me.

(b) Drafting of the ‘Extended descriptions of levels’ and ‘Advice on how to improve’ for writing. The drafting of these components of feedback was divided between work group members, and I had the task of producing the first version for the skill of writing.

(c) Contribution to the overall design of DIALANG feedback. This includes at least the addition of ‘Information about self-assessment’ to feedback and the way the post-test item review / feedback works. It should be noted that ‘Information about self-assessment’ was not planned to be the only such general package of information in DIALANG. Similar information about e.g. learning strategies was also planned to be included in the system but due to mainly practical reasons this plan was never completed.

In addition to my involvement in the design of feedback for DIALANG, this part of the system has continued to interest me, and much of my research on DIALANG after the launch of the beta version of the system in 2001 has been on feedback that it gives. The purpose of this research has been two-fold: a scientific interest to understand how the feedback works and a more practical need to develop the system in the future.

The third strand of my work on DIALANG concerned the coordination of the translations of the feedback texts and the instructions and help files used by the computer system underlying the testing system. The main concern there was to ensure that the translations were as
equivalent between the languages as possible, which can be seen as a way to make the system as comparable as possible between the different languages. To achieve this, I co-authored with Steve Fligelstone (Lancaster University) translation guidelines for the 13 teams that translated the texts from the English original. The work also included systematic cross-language checks between translations in the languages that I had at least some knowledge of.

The fourth type of contribution to the system that I made concerned the planning of the overall system as a member of the coordinating centre (1996-1999) and then as a member of the team of area / sub-project coordinators (test development and translations being my responsibility) (2000-2004). During the first phase of the project, I drafted the first version of the overall DIALANG Assessment Framework (a document that specified which areas of content in the interim version of the Common European Framework were to be used as the basis for DIALANG) and the first version of the test specifications for grammar.

**Own contribution in the writing of the two joint articles**

In the article with Charles Alderson, my contribution was smaller than my co-author’s. My contribution was fairly considerable in the first sections of the article (introduction, description of the system, test development) and in the conclusions; in fact, we mostly wrote these sections jointly, and it is impossible to disentangle the contributions of the two. We often used pieces of text produced by us for other purposes as a starting point for drafting sections of text. My co-author was just completing a book on DIALANG at the time of writing the article, so the empirical results of the piloting reported are a selection of those he included in his book, and he was mainly responsible for that section. As I described above, my contribution to the development of the system, its tests and the gathering of the data reported in this article, was, however, considerable.

The particular (joint) writing process makes it difficult to distinguish the contributions of the two authors in the above article, expect for the reporting of the empirical results, which was clearly my colleague’s responsibility. In the article Huhta & Figueras (2004), on the contrary, it is somewhat easier to draw the line between the two writers because we adopted a slightly different approach to writing the article. The two of us first drafted specific sections of the article, which we then joined together and only then started editing the whole draft. On the whole, it appears that I contributed more to this article than my co-author, at least in terms of
length of text. Most of the text from page 66 (Using the CEF) to page 74 (up to Lessons learnt and the way forward) were drafted by me, except for the section on the European Language Portfolio on page 70. Obviously, in this article, too, there are many ideas and pieces of text that one writer contributed to the text initially drafted by the other.
APPENDIX 3 – Part 1

Concept map of the advantages of DIALANG self-assessment based on learners’ sorting (the non-modified output from the Permap multidimensional scaling programme)
APPENDIX 3 – Part 2

Concept map of the advantages of DIALANG self-assessment based on learners’ sorting (the ‘opened’ output from the Permap multidimensional scaling programme where the clusters have been opened to allow the identification of the advantages of self-assessment in DIALANG). For the meaning of the numbers in the map below, see the Appendix in Huhta (submitted).
APPENDIX 4

Summary of the content of the six articles that form the thesis

All the articles deal with DIALANG, which is a computerised assessment system that provides diagnostic feedback on the strengths and weaknesses in the learner’s language proficiency in 14 different languages. As a whole, the articles provide validity arguments and evidence for the system, and thus, they can be thought of as a set of linked validity studies.

Articles 1 and 2 give an account of the rationale, background and development of DIALANG, in particular of how the Common European Framework of Reference was used as the main basis for the system. They describe what steps were taken in the design phase of the system to ensure as high quality as possible; some language testers (e.g. Weir 1993, 2005). Weir (1993) refers to this stage as a priori or theory-based validation. Also article 5 concerns the theoretical underpinnings of the system. Diagnostic testing, which is what DIALANG attempts to do, is a very messy concept, as discovered by Spolsky (1992) and Alderson (2005). Thus, in article 5 the author attempts to analyse and clarify the meaning of ‘diagnosis’ in educational measurement and to try to relate it to other, very similar types of testing. The article also reviews some literature on diagnostic assessment.

The other articles (including parts of article 2) report on different types of empirical research into DIALANG. Article 2 reports on the analyses of the piloting of the English tests of the system, and thus provides evidence of the reliability and validity of the language tests which for the core of the system and on whose quality the validity of much of the feedback depends. Articles 3 and 4 focus on users’ reactions to DIALANG and the feedback it gives. Article 4 gives a brief overall account of reactions to all parts of the quite wide variety of feedback on offer in the system and then focuses on the users’ views on self-assessment feedback. Article 3 deals with the most controversial subsystem of DIALANG, namely the Vocabulary Size Placement Test and analyses how the users’ comments shed light on certain problems in the system and, thus, on its validity. The last paper, article 6, analyses conceptions of self-assessment in more detail. By using methodology borrowed from other fields of science, the author analyses the benefits and problems associated with self-assessment and self-assessment feedback in DIALANG.


This article gives an overview of DIALANG and, in particular, how it makes use of the Common European Framework of Reference (CEFR) in its design. Self-assessment and the different parts of feedback are reviewed to see how they draw on the CEFR. A lot of the article is devoted to a description of the underpinnings of DIALANG with reference to pedagogy, life-long learning and autonomous, self-directed learning.

The article gives a more detailed overview of the parts of the DIALANG system and of their development: item writing, piloting, standard setting. The analysis of pilot data and the IRT-based calibration of the test items (and self-assessment statements) are also described. Overall quantitative results of the piloting and calibration are given. Some empirical results of the standard setting procedures, such as inter-judge agreement and relationship between judgements and empirical item difficulty. In addition, the correlations between different skill tests and between the tests and self-assessments are reported.


The article focuses on qualitative evidence on the validity, and possible directions for improvement, of the Vocabulary Size Placement Test (VSPT) used in DIALANG to direct the user to the most appropriate level of test. The evidence consists of users’ comments in questionnaires and interviews on the VSPT. The VSPT is an example of an unusual test format: in it the user has to decide if the words presented to them are real or not. The study reports on the purpose and development of the VSPT and then presents the users’ reactions to the VSPT results and to the test format itself. The users’ reactions to this part of DIALANG were more extreme – both negative and positive – than to other aspects of the system. The main finding was that the scoring algorithm underlying the VSPT is not optimal for certain learners and results in an underestimation of their proficiency. Some quantitative analyses give support to this conclusion.


The article reports on a study into language learners’ reactions to feedback on self-assessment of language proficiency, which they get from the computerized DIALANG assessment system. Feedback includes (1) a comparison of self-assessments and test results and (2) information about possible reasons for a mismatch between the two. In the study, 557 learners replied to questions about, e.g., the interest and usefulness of the feedback. The article reports on how users’ reactions to the new self-assessment feedback compared with other types of feedback, including traditional test feedback, i.e. the result and feedback on individual test items. Also reported are findings on how users’ background (e.g. age, gender, proficiency) were related to their evaluation of the interest and usefulness of feedback. The article concludes with a brief account of further research into the justifications the learners gave to their views on assessment feedback, which makes use of particular concept-mapping methodology.


There is considerable controversy in the literature as to what counts as diagnostic assessment. This article attempts to clarify the meanings of diagnosis and, in particular, its relationship to
formative assessment and other learning-oriented types of assessment. Work on both language testing and more general educational measurement is covered. The article also considers the position of DIALANG in relation to other diagnostic assessment and to other closely related types / purposes of assessment. In addition to treatment of concept and definition of diagnostic and formative assessment, the article reviews research on several topics related to both types of assessment: effects of assessment on learning, ability to diagnose performance, nature and practices of formative assessment, diagnosis of comprehension skills, and the content and construct of assessment.

(6) Huhta, Ari (submitted; now being revised)
Concept-mapping – an approach to understanding self-assessment

The study aims at understanding the concept of self-assessment by analysing language learners’ and language teachers’ views of the pros and cons of self-assessment as it is applied in DIALANG. The data for the study come from open-ended responses in the questionnaire study referred to in article 3 above; the questionnaire let the respondents to tell about the reasons why they thought self-assessment and self-assessment feedback of DIALANG were either useful or useless for them. A similar set of pros and cons of DIALANG self-assessment was gathered from language teachers using DIALANG in their teaching. The present study focuses on the advantages of self-assessment, which were analysed with the help of concept mapping procedures used in business, health care and general education (Trochim 1989). Sixteen teachers and language majors independently sorted the pros into groups. Their groupings were then analysed with the help of multidimensional scaling to find out the conceptual groups that emerge from their combined groupings. The outcome of the study is a concept map for the advantages of self-assessment in DIALANG that illustrate the dimensions of the pros and also how different dimensions relate to each other in terms of closeness.
References


YHTEENVETO

Väitöskirjassa esitetään yhteenvento ja analyysi DIALANG-järjestelmän laatua koskevista tutkimustuloksista ja muista siihen liittyvistä todisteista. DIALANG on 14-kielinen, tietokoneella internetin kautta suoritettava kielitaidon arviointijärjestelmä, joka antaa kielenoppijolle palautetta heidän kielitaitonsa tasosta ja sen vahvuudesta ja heikkouksista. Lisäksi järjestelmässä on mahdollista harjoitella kielitaidon itsearviointia ja saada palautetta itsearvioidun ja testissä osoitettun kielitaidon vastaavuudesta. DIALANG antaa käyttäjilleen myös vinkkejä siitä, miten kielitaitoa voi kehittää edelleen.

DIALANGilla on useita toisiinsa liittyviä tarkoituksia. Se antaa yksittäiselle kielenoppijalle tiedot hänen kielitaitonsa tasosta ja sen vahvuksista ja heikkouksista. Se kehittää oppijan tietoisuutta kielitaidosta, arvioinnista ja kielenoppimisesta; lisäksi järjestelmä neuvoo miten kielitaitoaa voi kehittää edelleen. DIALANG auttaa myös kielenopettajaa, joka haluaa saada tietoa oppilaittensa kielitaidosta sekä oppilaitostaa, joka tarvitsee välinettä jonka avulla sijoittaa oppilaita eritasoisille kielikurssseille. Viimeksi mainittu ei ollut DIALANGin alkuperäisiä tavoitteita, mutta testin käyttö taso- ja sijoittamistestinä on ilmeisesti nykyisin sen yleisin käyttötapa.


Ongelmallisin osa DIALANGia näyttäisi olevan sanastoava mittaava tasotesti, joka suoritetaan testin aluksi ja jonka perusteella oppija saa suoritettavakseen joko perus-, keski- tai ylimmän
tason testiversion siinä taidossa (esim. tekstin ymmärtämisessä), jonka hän on valinnut suoritettavaksi. Tasotestin tuloksen laskeva algoritmi näyttää rankaisevan vastausten arvaamisesta liian paljon, mikä johtaa joissakin tapauksissa liian matalaan tasotestin tulokseen ja sitä kautta liian helpon testiversion antamiseen oppijalle. Internetin kautta suoritettavalla tietokoneohjelmalla DIALANG kärkii välillä erilaisista teknisistä ongelmista, eivätkä kaikki halukkaat oppijat ja oppilaitokset välittämättä päätse käyttämään sitä tai he kärsivät ajoittain erilaisista teknisistä ongelmista.


Pitemmällä tähtäimellä DIALANGin suurin hyöty on todennäköisesti sen vaikutus kielitaidon ja kielen oppimisen diagnosointia koskevaan tutkimukseen soveltavassa kielitieteessä. DIALANGin ansiosta soveltavat kielitieteellisistä heräsivät huomaamaan, kuinka huonosti toisen ja vieraan kielen oppimisen ongelma ja niiden diagnosointia tunnetaan, varsinkin verrattuna äidinkielien oppimisen ja sen ongelmien diagnosointiin, joka on varsin pitkälle kehitetty (Alderson 2005). Havainto onkin jo johtanut lupaaaviin, uudenlaisiin avauksiin toisen / vieraan kielen oppimisen ja sen ongelmien tutkimuksessa ja tulevaisuudessa voimme odottaa entistä tarkempaa tietoa kielenoppimisen diagnosoinnista ja parempia diagnostisia testejä.