Collaboration in Language Testing and Assessment
To the memory of Dr. Felianka Kaftandjieva
# Table of Contents

Foreword  
*John H.A.L. de Jong*  
9

Chapter 1. Issues and Challenges in Combining SLA and Language Testing Research  
*Riikka Alanen, Ari Huhta, Scott Jarvis, Maisa Martin and Mirja Tarnanen*  
15

Chapter 2. Replicating Results from a CEFR Test Comparison Project Across Continents  
*Jamie Dunlea and Neus Figueras*  
31

Chapter 3. Adaptations of CEFR Descriptors to Local Contexts  
*Karin Vogt*  
47

Chapter 4. Co-operation in Europe: ‘Local’ Practices and ‘Global’ Requirements  
*Carole Sedgwick*  
63

Chapter 5. Working with the CEFR: The Bilkent Linking Project  
*Carole Thomas*  
79

Chapter 6. A Collaborative Approach to National Test Development  
*Gudrun Erickson and Lisbeth Åberg-Bengtsson*  
93

Chapter 7. Comparing the Hungarian School-Leaving Examination with International Language Examinations  
*Györgyi Együd, Zoltán Kiszely and Gábor Szabó*  
109

Chapter 8. Local Institution, Global Examination: Working Together for a ‘Co-certification’  
*David Newbold*  
127

Chapter 9. Matura’s Rocky Road to Success: Coping with Test Validity Issues  
*Slobodanka Dimova*  
143

Chapter 10. Synergies and Tensions in Computerised Language Testing  
*Anders Johannessen Fikke and Hildegunn Lahlum Helness*  
159
Chapter 11. Translation Assessment Methodology and the Prospects of European Collaboration

June Eyckmans, Winibert Segers and Philippe Anckaert

Chapter 12. Preliminary Collaborative Steps in Establishing CEFR Sign Language Levels

Laura Sadlier†, Beppie van den Bogaerde and Joni Oyserman

Chapter 13. Collaboration in Understanding Results – Self-assessment of EFL Writing

Anne Dragemark Oscarson

Chapter 14. Effects of CLIL Teaching on Language Outcomes

Lisbeth M. Brevik and Eli Moe
Foreword

John H.A.L. de Jong
Pearson / VU University Amsterdam

The selection of articles in this book, dedicated to the memory of Felianka Kaftandjieva, has been gathered by Dina Tsagari and Ildikó Csépes around the theme of collaboration. Felly has collaborated with many of the authors in this book and has most certainly influenced them in their attention to detail, their respect for measurement rigour and their enjoyment in research discoveries. She taught us to savour the reward of finding meaningful results achieved by painstakingly searching for truth in data. The editors and the authors of this book by collaborating on its publication show their wish to honour her memory by bringing the lessons she taught us into practice.

The articles were originally presented at conferences of the European Association for Language Testing and Assessment (EALTA) from 2008 to 2010 in Athens (Greece), Turku (Finland) and The Hague (The Netherlands). The collection illustrates the international and inclusive character of EALTA, involving researchers in large scale testing as well as in classroom assessment.

As scientific disciplines develop they tend to diversify into a growing number of branches that specialise in ever smaller areas in order to study these areas in more detail. General linguistics as it developed around the beginning of last century subdivided into branches like psycholinguistics, sociolinguistics and applied linguistics around the middle of last century. The latter then again split up into the studies such as child language, second language acquisition and language testing. Language testing into listening, speaking, reading and writing. Studying at increasing levels of granularity is a necessity brought about by the expanding human knowledge, but occasionally we have to zoom out again to understand how our fragmentary understanding relates to the real phenomenological world. This reverse direction has also occurred within language testing where we have seen how initial discrete point testing was replaced by communicative four skills testing, which in its turn is currently being exchanged for integrated skills testing to get at the language how it is used in real life. Specialists in listening comprehension therefore find themselves having to collaborate with specialists in the testing of speaking, reading experts with writing experts. Similarly second language acquisition scholars seek collaboration with language testers.

With the ease of communication created by electronics and the relatively decreasing cost of international travel we also see increasing collaboration between scholars from research centres worldwide.

1 John.DeJong@Pearson.com
Indeed the first chapter in this volume by Riikka Alanen, Ari Huhta, Scott Jarvis, Maisa Martin and Mirja Tarnanen exemplifies both these forms of collaboration. Four authors from the University of Jyväskylä, Finland collaborated with one researcher from Ohio University in the USA. They point at a lack of co-operation between language learning and language assessment research and contend that both fields come together in defining the construct of L2 proficiency. To illustrate their point they report on the CEFLING project, a study on the relationship between linguistic features of writing performance and functionally defined language proficiency levels. The functional description is taken from the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). Their study aims to detect which linguistic developmental phenomena go in parallel with the set of levels of increasing proficiency defined in the CEF. Their study shows how our knowledge and understanding can be advanced by combining the depth and the attention to detail of SLA with the rigour and the larger numbers typically used in language testing.

In the next chapter Jamie Dunlea, based in Japan, and Neus Figueras, born and raised in Spain collaborate on finding out whether the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) applies as well outside of its original European context as it does inside. For their study the authors have chosen to evaluate the relationship between the EIKEN tests developed in Japan and the CEFR. The EIKEN tests are a seven-level set of tests made and administered by the Society for Testing English Proficiency (STEP), a non-profit foundation established in Japan in 1963. In order to properly introduce their study they provide a thorough discussion of standard setting methodologies. Preliminary studies based on research in Japan had established a hypothesized relation between particular scores on the EIKEN tests and the CEFR. The present study used judgements from teachers based in Catalonia and found slightly higher cut-offs than the original cut-offs found in Japan, but within their predefined interval of acceptable difference. The authors conclude therefore that the difference is small and that the interpretation of the CEFR across the two cultures is closer than some critics of the usage of the CEFR outside the European context would have expected.

In the third chapter Karin Vogt starts out by discussing the quality of the set of performance and ability descriptors that came with the publication of the CEFR. She points at several inconsistencies and inadequacies and proposes a method for extending and improving the original set of descriptors to apply to the vocational sector. Therewith Karin introduces yet another form of collaboration, that is, collaboration between professional language testing and language testing in the professions. The chapter reports on developing new descriptors in three specific professional contexts and presents a thorough study using both qualitative and quantitative methods to validate the newly developed
descriptors. The chapter is an example of how the CEFR can be applied as intended: a framework of reference allowing development beyond the illustrative set of descriptors provided in the 2001 publication.

Carole Sedgwick reports in Chapter 4 how she lost her initial enthusiasm for the ideal of the Bologna Declaration (1999) which is to enhance transnational mobility of European citizens by creating comparability of degrees in higher education. Based on her research she fears the richness of local diversity might be threatened by the centralizing tendencies in the Bologna Declaration. She based her study on thesis writing as it is practised and valued in two European countries: Hungary and Italy and found that local academic traditions tend to be undervalued in an approach that attempts to homogenize the approach to thesis writing. Carole’s study exemplifies the European dilemma between unity and diversity. Wishing to recognize the equality of languages, the European Union tries to withstand the natural development of English becoming a lingua franca. On the other hand, the Union’s ideal of human mobility is assumed to require a homogenization of tertiary degrees across the European Union, but, as Carole points out, in thesis writing this leads to standardization on an Anglophone model.

In Chapter 5, Carole Thomas presents an example of post-hoc linking of an existing exam to the CEFR. The object of study is the Certificate of Proficiency in English (COPE) developed and administered by Bilkent University in Turkey. Engaging in this activity revealed initial uncertainty among teachers about the levels and required gaining greater knowledge and understanding of the CEFR. An important side-effect was that by involving teachers in the process of standardization they were trained on the CEFR. This collaboration has resulted in a growing number of teachers with experience and understanding of the CEFR.

A similar process of collaboration in the Swedish context is described by Gudrun Erickson and Lisbeth Åberg-Bengtsson in Chapter 6. In Sweden there is a long tradition of involving teachers in standardized assessment. In fact the teachers are entirely responsible for grading their students when exiting secondary education, but a national standardized examination is provided to assist teachers in this task. This examination is developed in close collaboration between teachers and experts. Students too are involved in piloting the exam tasks and in providing feedback on their experience with them.

Györgyi Együd, Zoltán Kiszely and Gábor Szabó in Chapter 7 present a critical discussion of the national school-leaving exam in Hungary. They compare this exam to two other exams accredited by the Hungarian state: the exam from the European Consortium for the Certificate of Attainment in Modern Languages (ECL) and Cambridge First Certificate in English (FCE). The results from administering the three tests to pilot groups suggest that the reading tasks on all three exams are quite similar in difficulty but that the
writing tasks on the school-leaving exam are significantly easier. These results are remarquable as the Hungarian Accreditation Board for Foreign Language Examinations (HABFLE) sets quite strict requirements on the alignment of international exams to the CEFR. Apparently the national school-leaving exam is not held to meet these requirements.

In Chapter 8, David Newbold describes how the University of Venice has managed to find a way to satisfy both the need for external recognition for English language exam results and relevance of the examination with respect to a local curriculum. As a result of the Bologna Declaration (1999) Italian Universities experienced the need to revise their curricula and for foreign languages they choose to model these on the CEFR. The revision of the curriculum for English required reviewing the examinations. They choose one of the internationally operating language test providers (Trinity) and collaborated with them to produce a localized version that is recognized both by the university and by the external test provider. Winner in this effort of co-certification is obviously the test taker, who taking a local exam, obtains an internationally recognized certificate at the same time.

Slobodanka Dimova in Chapter 9 presents an interesting record of the introduction of a new examination system in the Republic of Macedonia. As part of secondary education reforms the national “Bureau for Development of Education” decided to introduce a Matura exam to replace the traditional school-based examination. Although positive washback on instruction and enhanced credibility of the secondary school diploma were expressed as main goals, the introduction of the Matura also had a mixed set of further objectives including the control of teaching curricula based on educational standards. Objections against the Matura came from the general public and from students fearing a reduced probability of passing the exam. Dimova suggests that a major cause of the upheaval about the new exam was that authorities failed to involve stakeholders in the design and implementation of the innovation. From her study based on surveys conducted with students and teachers it would appear that these stakeholders felt there was a discrepancy between the curriculum as taught and the exam. She concludes that collaboration with stakeholders in the design phase could have been beneficial for a better understanding: the developers knowing more about the expectations of the students and the students realizing the intentions of the developers.

Chapter 10 provides a report by Anders Johannessen Fikke and Hildegunn Lahlum Helness from the University of Bergen, Norway on three parties involved in moving a computer-based test to a new platform. The authors discuss sources of tensions in this collaborative effort involving the contracting authority, the test developers and the software engineers. The chapter shows how ultimately success can be achieved by surmounting technological hurdles, but most of all by a cooperative attitude of parties involved.
June Eyckmans, Winibert Segers and Philippe Anckaert in Chapter 11 advocate the collaboration between translation trainers and test developers because they see the need to further professionalize translation assessment practices. With the advent of the communicative approach to language learning, teaching and assessment, translation went out of the window with grammar-translation method. Thirty years later we see a renewed interest in translation because of the needs of our international society. The authors describe the current situation in translation assessment and point at the shortcomings from a measurement perspective of the Translation Services-service requirements issued by the European Committee for Standardization. It is therefore that they call for standardization in measurement principles and methodology. In order to achieve this they suggest developing a Common Framework of Reference for Translation Competence and the adaptation of the EALTA Guidelines for Good Practice to also cover Translation testing.

Yet another new field is introduced in Chapter 12 by Laura Sadlier and Beppie van den Bogaerde who report on Irish and Dutch scholars in sign language (SL) collaborating on investigating the necessary adaptation of the CEFR to make it appropriate for learning, teaching and assessment of SL’s. Clearly no adaptation is required in the context of the written skills, but for the spoken skills the use of the CEFR would require replacing the auditory-verbal concepts by visual-gestural ones. The authors describe current developments in creating assessment instruments for testing SL skills and present examples from tests developed in Ireland and in the Netherlands.

Chapter 13, authored by Anne Dragemark Oscarson, revisits collaboration between teachers and learners. The author reports on a study conducted in Sweden where students were required to self-assess their writing skills based on writing exercises and self-assessment instruments. Teachers and students appreciated the self-assessment, but correlations between teacher grades and learners self-assessment were moderate for overall writing (explaining one-third or less of the variance) and low for specific writing tasks. Nevertheless the self-assessment exercise is seen as positive as learners acquire greater insight in the learning task. In focus interviews students did express their appreciation and mentioned they became more aware of the learning process and its requirements. In addition, the students indicated that the self-assessment equipped them with life-long learning skills.

In the last chapter, Chapter 14, Lisbeth M. Brevik and Eli Moe from Norway describe how researchers benefitted from their collaboration with teachers to investigate the effect of Content and Language Integrated Learning (CLIL). The researchers allowed teachers and students to choose whether they wished to participate in a CLIL or in a control group. Parallel tests of listening and reading were developed and administered one at the beginning of the experiment and the second at the end of the same school year. Both groups
showed substantial gain in ability as measured by the tests, but the CLIL showed slightly more increase. The researchers grapple with the usual problem of lack of representativeness of small samples in classroom-based research on teaching methods and found that the CLIL group started the experiment at a level where the control group ended. A promising finding, however, was that the group showing the largest gain in ability during the experiment was the weakest performing CLIL group, which may suggest that CLIL is most beneficial for weaker students. The authors admit however that the n-size (35) for such analyses on subgroups was rather small and suggest that this inherent disadvantage of the small n-size in classroom-based research may be mitigated by setting up many small studies in parallel, thus calling for more collaboration.

Common themes across a number of papers in this volume are the usage of the CEFR and the positive influence of the rigorous data-driven approach from language testing on other branches of applied linguistics. Gone are the days of the single-case examples in describing linguistic phenomena. The concepts of validity, reliability and rater agreement seem firmly established in linguistic research outside of language testing and principles of good practice such as the EALTA Guidelines for Good Practice in Language Testing and Assessment (EALTA, 2006) are frequently mentioned as example to be adopted from the language testing field. It is encouraging to learn from this volume that collaboration across disciplines, across borders and among stakeholders does make a difference and leads to improving the quality of research.

**References**


Recent reviews of research on language learning and language assessment reveal a lack of co-operation between the two fields. Combining their strengths, however, could help to address problems which are difficult to tackle from one point of view only. Yet how does such cooperation between language testers and SLA researchers take place in practice? What are the benefits for both parties and what are theoretical and methodological issues that need to be considered? In this chapter, we discuss the methodological and theoretical issues involved in doing and combining SLA and language testing research. For some concrete examples, we will draw on the CEFLING project, a study on the relationship between linguistic features of writing performance and functionally defined language proficiency levels. Finally, we will attempt to show how serious attention to good language testing practice improves the quality of the study and the validity of the findings by ensuring that the elicitation and rating of learners’ performances are dependable.

Key words: second language learning, language testing, task-based assessment

1. Introduction
In recent years, there has been a renewed interest in combining two rather different fields of research of second language acquisition (SLA) and language testing. At the outset, these two fields do not seem to be very far from each other; after all, they both share an interest in second or foreign language
proficiency (L2 proficiency). While SLA research strives for a description of SLA acquisition by collecting and analyzing L2 data from learners, language testing mostly aims for a fair assessment of L2 users’ communicative ability. Language testing research investigates different aspects of language testing practices from construct building and validity to rater reliability. Although from the SLA research perspective, language testing is firmly occupied with the here-and-now of L2 use, this is of course not the case. In order to be able to generalize from L2 users’ performance at the testing situation, language testing research rests on the bedrock of communicative language ability: only by positing the existence of such a construct is it possible to make the assumption that test takers’ use of L2 in the testing situation transfers to other situations.

In fact, the construct of L2 proficiency emerges as a powerful link between these two fields of research. It is no wonder that researchers and scholars in both fields of research (Bachman & Cohen, 1998; Byrnes, 1987, Alderson, 2005; Hulstijn, 2007; Norris & Ortega, 2003, 2009; Sigott, 2000) have identified the research area surrounding the construct of L2 proficiency – how it is conceptualized and operationalized, among other things – as the one that could be a source of fruitful co-operation between the two fields. In this chapter, we will raise some of these issues and challenges as they emerged during a research project combining both SLA and language testing research. The aim of the project CEFLING (Alanen, Huhta & Tarmanen, 2010; Martin, Mustonen, Reiman & Seilonen, 2010) is the description of the linguistic features of written Finnish L2 and English L2 in Finland placed at various levels of proficiency according to the Common European Framework of Reference (CEFR) scale for writing. More specifically, we will discuss issues related to the research design in which the same L2 tasks were used for both L2 data elicitation and L2 proficiency assessment purposes at the same time. We will also propose that the benefits that this type of research co-operation could bring may depend on the aims and practices of the particular field of study: some types of SLA research may benefit from language testing research more than other types, and vice versa.

2. SLA research meets language testing

2.1 Measurement of L2 proficiency

In SLA studies targeting the effect of outside factors such as L2 teaching methods or quality of input and interaction, the participants’ level of L2 development is usually determined in some way. Quite often researchers use measures such as length of exposure to L2, or scores of well-known proficiency tests (e.g. TOEFL) (see e.g. a review of L2 proficiency measures used in SLA studies by Tremblay & Garrison, 2010). In these studies, SLA research quite often uses an entirely different set of tasks for data elicitation, with the proficiency measure serving only as a background or control variable since the
focus most of the time is on the detailed description of what the participants were able to do in L2.

Language testing is concerned about the reliability and validity of assessment: whatever tests, tasks or items are used, they should measure L2 proficiency as validly and reliably as possible. From this perspective, it should come as no surprise that language testers may find the way SLA research sometimes characterizes L2 learners as ‘beginners’, ‘intermediate’ or ‘advanced’ rather vague or even haphazard. This criticism, though sometimes warranted, is also too harsh, mainly because it overlooks the rather different goals that SLA research has. For SLA research is usually interested in the development of particular features in the learners’ L2: complexity, accuracy or fluency, crosslinguistic influence, or how learners use articles, case endings, prepositions, discourse markers, tense and aspect, vocabulary items, how they apologize etc. While SLA research may use L2 proficiency as background or control variable, these measures are rarely enough for research purposes since they yield little data about the development of specific linguistic features. Perhaps because of this, researchers often use data elicitation methods specifically designed for their particular needs, which may, however, have issues with reliability and validity, at least in terms of the standards applied in most serious language tests.

The level of detail in the analysis is consequently somewhat different in these two fields of research. In SLA research, the performances of individual learners are typically studied in detail while in language testing, this varies. After all, testing is mostly concerned with the measurement of constructs, while research is above all interested in constructs themselves (see e.g. Fulcher & Davidson, 2007; Sigott, 2000). Prototypically, proficiency testing (e.g. international language examinations), many placement and achievement tests, and most tests used in national and international studies about language learning have a very broad view of proficiency (Alderson & Huhta, 2010). However, there are forms and purposes of testing and assessment that aim at getting a fairly detailed picture of a learner’s proficiency. This is the case in much of formative assessment and in diagnostic assessment (see e.g. Alderson, 2005; Huhta, 2008). In this respect, formative or diagnostic assessment can be rather similar to SLA research when it comes to analyzing learners’ linguistic performance.

2.2 Using CEFR levels in determining learners’ language proficiency
There are few studies so far that have attempted to use the CEFR scales to determine L2 users’ proficiency. Many language tests have been aligned with the Framework levels post-hoc and there are even some tests whose design was based on it, such as DIALANG (Alderson, 2005; Alderson & Huhta, 2005) but the direct use of CEFR scales to determine, for example, learners’ speaking or
writing level has been surprisingly uncommon in research so far. The framework has attracted a fair amount of criticism, mainly because it has been sometimes misused as a model for communicative language ability (Fulcher, 2004; Fulcher & Davidson, 2007), or its proficiency scales have been uncritically and without the due process of validation or calibration used for purposes that they were not meant for (Weir, 2004). However, the CEFR scales are widely used in Europe as a framework for describing and comparing language examinations, curricula, courses and textbooks. In Finland, the CEFR scales form the basis for the 10-point Finnish National Core Curriculum scales for second and foreign languages. The National Certificate language examination system in Finland uses scales that have been linked with the CEFR scales as well. When a decision was made in 2006 to set up a research project to investigate the linguistic features of L2 Finnish and L2 English learners at various levels of L2 proficiency, these were among some of the practical reasons when we decided to use the CEFR scales to operationalize the levels of L2 proficiency.

In the CEFLING project, L2 Finnish and L2 English data were collected from young and adult L2 learners by using a set of communicative L2 writing tasks: Task 1 was an email message to a friend, Task 2 was an email message to a teacher, Task 3 was a complaint to an internet store, Task 4 was an opinion piece, and Task 5 was a story. For logistical reasons, Tasks 1 and 2 were alternates. The young learners’ performances were rated by using two scales, the CEFR and the Finnish National Core Curriculum for Basic Education (2004) scales. The adult learners’ performances were rated with the National Certificate examination rating scales which are, as mentioned above, very similar to the 6-point CEFR scale (in fact, we used the operational ratings awarded to the learners when they had taken the NC examination). The data collected in the project were used to build an L2 Finnish and L2 English learner corpus for the analysis of linguistic features (within the limits presented by the data set).

There were other reasons for the adoption of the CEFR levels: there were renewed calls across Europe for a systematic study of L2 development that would attempt to take into account the issues and concerns of L2 testing (see e.g. Hulstijn, Alderson & Schoonen, 2010; see also the SLATE network at www.slate.eu.org), mainly brought on by the large-scale adoption of the CEFR by a great number of European countries. However, the decision to use the CEFR scale in a study combining both SLA and language testing perspectives brought along with it another, potentially problematic issue, that of holistic vs. analytic approach in L2 research. The CEFR scale is by nature holistic. Can it or any other scales similar to CEFR be considered appropriate for describing L2 users’ proficiency? After all, the level descriptions are so general that each level may include a great variety of performances.
In language testing, the choice between holistic and analytic rating is a very old problem (Hamp-Lyons 1991; Huot 1997; Kroll 1998). Similarly, in SLA research there are branches that prefer holistic approaches to the various issues in SLA rather than analytic studies. However, the kind of SLA research that aims at discovering and describing critical features of L2 systems can be characterized as fundamentally analytic by nature, if by analytic we mean its tendency to break down the whole of L2 performance into parts, i.e., linguistic features it purports to describe. This does not mean that SLA research is not concerned with the systemic relationship of parts to the whole; it means, though, that the way it goes about describing this relationship includes a detailed examination of linguistic evidence. In this respect, as alluded to above, diagnostic or formative approaches to language assessment share the same concern for a detailed analysis of learner performance.

Another potential problem that the use of the CEFR scale might give rise to concerns the linearity of L2 development. This is, however, a red herring: neither the CEFR scales – nor any other scale – take it for granted that L2 learning is linear. This is a misconception brought on by the scale itself even though it is just one way of describing a particular point in time during the long process of the development of L2 proficiency. It was not designed for nor can it easily describe development over time since by nature it is a rather general way of describing proficiency. However, the way it is used in CEFLING presents a well-founded attempt to systematically categorize L2 users’ performances according to their ability to use language for communicative purposes in L2 tasks designed to be as authentic as possible.

It is important to be clear about what the CEFR scales are actually used to measure. So far, we have skirted the problem by talking about how the CEFR levels can be used to place learners on various levels of L2 proficiency. In CEFLING, the CEFR scales were in actual practice used to rate learner performances on communicative L2 writing tasks: in other words, the scale was also used as a measure of communicative adequacy of specific task performances. As Pallotti (2009) suggests, communicative adequacy should be considered both as a separate performance dimension and as a way of interpreting linguistic measures such as complexity, accuracy and fluency. As Pallotti (2009) notes, there are few studies as of yet that have applied such holistic scales to the description of communicative adequacy (De Jong, Steinel, Florijn, Schoonen & Hulstijn, 2007; Gilabert, Kuiken & Vedder, 2010).

In the future, the data collected in CEFLING will hopefully give researchers a chance to look at not only the linguistic features of the CEFR levels but also shed light on the linguistic basis of the ratings. It is also important to remember that the CEFR levels were used for dual purpose in CEFLING: on the one hand, the ratings were used to categorize L2 performances into various levels of proficiency; on the other hand, they were used to capture one aspect of
L2 performance, that of communicative adequacy. L2 proficiency is of course a much broader construct than communicative adequacy. However, we hope that a detailed analysis of L2 performances collected in the project may help us to elucidate what the differences and similarities are. Some caution must be exercised here. In language testing, it is a common practice to extend the rating based on test or task performances to the individual L2 user, thus it is common to speak of A2 learners or intermediate learners. Similarly, in SLA research it is common to use L2 proficiency as a background or control variable and talk about ‘beginning learners’ etc. However, to be precise, the proficiency ratings in CEFLING only apply to L2 writing performances. Although it is tempting to talk about level A2 writers etc., it would be more accurate to talk about level A2 performances.

2.3 Variability of performances across tasks and time
Designing and selecting tasks that are relevant for both SLA and language testing research is particularly challenging. On the one hand, tasks should successfully elicit specific kinds of L2 data; on the other hand, tasks should also be communicative so that the communicative adequacy of task performances can be assessed. It was assumed from the beginning that learner performances across L2 tasks would vary along these both dimensions. There is a slight difference in the way SLA research and language testing view variability, however. In SLA research, variability of learner language is a well-known phenomenon, which researchers make every effort to understand and systematize. For example, learners’ linguistic performance varies depending on the task or other languages they know, and as every L2 teacher knows, a learner’s linguistic performance may vary even during the same task. Variability is a sign of L2 development: no variability in production means little observable change (for better or for worse).

For much of language testing, the variability of learners’ performance in different contexts and at different points in time is a problem to be tackled in some way. The variability of an individual learner’s performance across tasks is a major concern to language test designers as their aim is often to get a generalizable picture of test takers’ abilities. It is a well-established fact in language testing that the testing method (e.g. a multiple choice test or an oral interview) has a ‘method effect’ on the test results because some learners may be more familiar with or better-suited to doing some tasks than others (e.g. Bachman & Palmer, 1982). Therefore, reliance on only one testing method or task type is not recommended, but rather, tests use as many different task types as is practically possible (see e.g. Sigott, 2000). Unfortunately, tests can rarely include as many different task types as their designers would wish. It is particularly difficult to employ several different speaking and writing tasks because these tests tend to take a lot of time so two to three productive tasks
per skill are the most that typical language tests can include. In CEFLING, for example, the participants typically carried out all four tasks in class, usually on two different occasions. The performances were then rated by four raters after an intensive training period (for details, see Alanen, Huhta & Tarnanen, 2010).

The assessment of communicative adequacy imposes conditions of its own on the type of tasks: from the outset, learners’ level of L2 proficiency may either help or hinder them in their attempts to carry out L2 tasks successfully. In language testing, it is therefore common to use different tests and tasks for beginning, intermediate and advanced learners. From the SLA perspective, on the other hand, the use of different tasks for different learners can be viewed as a potentially confounding variable. In research combining both SLA and language testing perspectives, there are a number of solutions to this problem: for example, one can simply ask all learners, regardless of their age or proficiency level, to do all types of task, or, one can try to match tasks with the test taker’s ability. In CEFLING, an attempt was made to combine both approaches: all learners were asked to do a set of four different tasks (with one of the tasks having an alternate version); at the same time, the type of tasks that the participants were most likely to have encountered and the level of their L2 proficiency were carefully estimated in advance so as to make the tasks as suitable for them as possible (see Alanen, Huhta & Tarnanen, 2010).

Figure 1. The median ratings of proficiency levels of L2 English learner performances in Task 3 and Task 5 (0=A0, 1=A1, 2=A2, 3= B1, 4= B2).

On the other hand, the test must be able to differentiate among learners at different proficiency levels; in other words, the variability between learners is not only expected but desirable. As Figure 1 shows, roughly one third of the L2 English learners’ performances on Task 3 (‘Write an email to an internet store complaining about a defective computer game’) and Task 5 (‘Tell a story about
the funniest, scariest or most exciting experience of your life’) in CEFLING were rated at the proficiency level A2. Thus, it seems that the tasks were able to differentiate between learners; yet, at the same time, they were not too difficult or easy but targeted a range of proficiency levels appropriate for these particular L2 learners. The arithmetic average of the median scores from four raters for Task 3 was 2.04 (sd=0.823, N=212), while for Task 5 it was 2.01 (sd=0.842, N=210).

From the SLA perspective, it was expected that there would be some variation in the language used by individual learners, but would there be differences in the linguistic features produced by L2 English learners at different proficiency levels? A quantitative analysis of the number of main verbs in past tense used by learners in both tasks reveals differences between proficiency levels both between and within tasks (Figure 2).

Figure 2. Relative frequency of past tense forms to all simple verbs in each task at proficiency levels A1-B2 in Task 3 and Task 5.

<table>
<thead>
<tr>
<th></th>
<th>Task 3</th>
<th>Task 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>12.9 %</td>
<td>32.0 %</td>
</tr>
<tr>
<td>A2</td>
<td>22.0 %</td>
<td>51.1 %</td>
</tr>
<tr>
<td>B1</td>
<td>23.8 %</td>
<td>63.2 %</td>
</tr>
<tr>
<td>B2</td>
<td>27.4 %</td>
<td>70.3 %</td>
</tr>
</tbody>
</table>

As one might expect from the prompt, learners produced more past tense forms in Task 5 than Task 3; after all, they were supposed to tell about an event in the past in Task 5. On the other hand, in both tasks the learners at more advanced levels of proficiency used more past tense than present tense forms.

Consequently, it seems possible to bring out proficiency-based differences in L2 performance by using this kind of research design, although the analysis conducted here was a fairly simple and straightforward frequency analysis of a linguistic form. It is far more difficult, however, to track the development of individual learners, since most of them performed only four tasks, and not all of their performances were included in the first linguistic analyses: at the initial
stage of research (where these results come from), only those performances were included in the linguistic analyses that received a fairly unanimous judgment from the raters (three quarters of the L2 English raters had to agree on their assessment for the performance to be included in the data set as a sample for a particular proficiency level).

In addition, despite the fact that learners carried out as many as four tasks, the research design was still cross-sectional, using one-shot testing. As far as the time dimension is concerned, typical language tests are one-shot measures of performance that tell about learners at the point when they were tested. Studies using one-shot testing are susceptible to outside interference; learners may, for example, behave atypically on the day of the test and fall below their normal level of performance due to e.g. illness. The language portfolio (e.g. European Language Portfolio, Kohonen, 2000; Little, 2009) is a rare example of a type of assessment that can capture learners’ development over time, at least to an extent; in fact, portfolio assessment suffers less from the time limitation alluded to above than one-shot tests. A longitudinal research design may also be used. In fact, a new project funded by the Academy of Finland (2010-2013) called TOPLING – Paths in SLA (www.jyu.fi/topling) was set up, with Maisa Martin as the principal investigator, in order for us to be able to follow the same learners for a longer period of time. In this project, written L2 data will be collected from L2 Finnish, L2 English and L2 Swedish learners over a period of three years.

3. Inclusion vs. exclusion of data from learners’ performances
As was mentioned above, some performances were not included in the data set for linguistic analyses simply because the raters did not agree to a sufficient degree as to what their level of performance was. After all, a key reason for making use of language testing expertise in this particular SLA research study was to ensure that the learners’ level of proficiency could be established reliably. Hence, the reliability of the ratings, and thus of the data, is an important consideration, as it cannot be assumed that trained raters can automatically do a flawless job. The question then arises of how best to aggregate the raters’ (partially) different ratings to CEFR levels for the learners.

Several different methods can be applied and at least three of them have been experimented with in the CEFLING project. For in SLA studies that rely on only one source of information about the participants’ language proficiency, such as one test result, one rater or other single piece of information, there is probably no other option than to include all participants that fit the criteria in the study.

In studies such as CEFLING that employ multiple ratings of all performances it is possible to study the dependability of the ratings and, thus, the quality of the data, and, for example, leave out cases that cannot be placed