Designing and assessing L2 writing tasks across CEFR proficiency levels

Riikka Alanen, Ari Huhta and Mirja Tarnanen
University of Jyväskylä

With the advent of the Common European Framework of Reference (CEFR) for the learning, teaching and assessment of modern languages, there have been renewed calls for the integration of the research perspectives of language testing and second language acquisition across Europe. The project Cefling was set up in 2006 with this purpose in mind. In the project our aim is to describe the features of language that L2 learners use at various levels of language proficiency defined by the CEFR scales. For this purpose, L2 Finnish and L2 English data were collected from young and adult L2 learners by using a set of communicative L2 writing tasks. In the course of the project, the different understandings of what the purpose of an L2 writing task is needed to be reconciled not only in the minds of researchers but also in research design. In what follows, we will discuss the issues involved in designing and assessing L2 tasks for SLA and language testing purposes by using the design and assessment procedures in the project as a case in point. We will also present some of our findings to illustrate how statistical procedures such as multifaceted Rasch analysis can be used to examine task difficulty.

1. Introduction

Until quite recently, there have been relatively few empirical studies combining the research perspectives of language testing and second language acquisition. Beginning in the 1990s (see e.g. Bachman & Cohen, 1998), the number of such studies has steadily grown although it has remained fairly small, in particular in task-based research. With the advent of Common European Framework of Reference, CEFR, (Council of Europe, 2001) for learning, teaching and assessment of modern languages, there has been an increasing interest in setting up studies combining both research perspectives across Europe. Integrating the two research perspectives is not without difficulty, however; almost inevitably, compromises must be made. In this chapter, we will discuss a number of issues relevant to task design and assessment in research approaches attempting to combine the goals and practices of SLA research in task-based research, on one hand, and language testing, on the other hand, by using the theoretical and

EUROSLA MONOGRAPHS SERIES 1
Communicative proficiency and linguistic development, 21-56

methodological decisions taken in one such research project as an illustration. The project in question is called *Cefling – The linguistic basis of the Common European Framework levels: Combining second language acquisition and language testing research*; it is a project set up to study the linguistic features of the proficiency levels described by the CEFR scales (see Martin, Mustonen, Reiman, & Seilonen, this volume). The chapter ends with an illustrative analysis of task variability learner performance, and a discussion of how quantitative and qualitative analysis of task performance can help researchers to evaluate task design.

SLA research is, of course, primarily interested in the development of L2 proficiency, complexity, accuracy and fluency, in particular. Language testing, on the other hand, is occupied with the development of reliable and valid measures for assessing communicative language ability or language proficiency. It is primarily interested in how successful the items used in language testing are, and, depending on the type and goals of the language test, also in the communicative adequacy of tasks (see Pallotti, 2009).

Task emerges as a key notion linking both SLA research and language testing practice. It is a key unit both in L2 data elicitation and measurement. In SLA and language teaching, task is regarded as a programmatic or even curricular unit, a type of meaning-based activity L2 teaching should be focused on or organized around (see e.g. Bygate, Skehan, & Swain, 2001; Ellis, 2003; Samuda & Bygate, 2008; Van den Branden, 2006; Van den Branden, Bygate, & Norris, 2009). It is also a key unit in performance based assessment of L2 proficiency. As Brindley (1994/2009) defines it, task-based language assessment (or task-centered as it was called then) is

> the process of evaluating, in relation to a set of explicitly stated criteria, the quality of the communicative performances elicited from learners as part of goal-directed, meaning-focused language use requiring the integration of skills and knowledge. (p. 437)

In task-based assessment, task performance can be assessed according to its communicative adequacy, i.e., on how well the learner is able to use language to accomplish task requirements. Communicative adequacy is commonly evaluated by rating scales; yet, as Pallotti (2009) notes, there are surprisingly few studies attempting to look at complexity, accuracy and fluency in terms of such scales in task-based SLA research. De Jong, Steinel, Florijn, Schoonen, & Hulstijn (2007) and a number of other studies in this volume (e.g. Gilabert, Kuiken, & Vedder; Martin et al., both this volume) are among the first to approach the issue from this particular perspective.

The notions that are of particular importance for this chapter are *task*, and how *L2 performance* on a particular task are perceived and operationalized in

SLA research and language testing. In a number of ways, some of the issues fundamental to task-based language assessment are central to this chapter, as well. To slightly modify the list originally devised by Norris (2002, p. 337), in this chapter we will discuss the following issues: 1) Why are participants asked to perform communicative L2 writing tasks in the first place? 2) What exactly do researchers want to know on the basis of their task performances? 3) How can tasks be selected or designed, and performances judged, so that researchers can know these things? and 4) What are going to be done with judgments of participants' task performances, once they have been elicited?

In what follows, we will first briefly describe the characteristics of language testing practice and language testing research relevant for SLA research. It is useful to be aware of what sensible language testing *practice* entails because that is crucial for the quality of whatever data elicitation and collection instruments an SLA researcher is using. Awareness of what goes on in language testing *research*, for its part, is also useful for ensuring the quality of the instruments but here the contribution of testing relates more to the conceptual level of the entire measurement process (cf. Norris & Ortega, 2003, p. 720). In the latter part of the chapter, we will discuss issues relevant to task design and assessment, and finally, show, by using the judgment data from the Cefling project as an illustration, what can be done to analyze and evaluate the participants' task performance.

2. SLA research and language testing: goals, purposes and practices

L2 development is a complex, dynamic process; however, it is mostly investigated through L2 products, slices of L2 performances elicited at certain points of time under a set of specific circumstances. There are a number of aspects in L2 development that have been the focus of study over the years, including particular linguistic structures (e.g. negation, question formation, tense and aspect). Increasingly, since the 1990s, the notions of complexity, accuracy and fluency have come to be used to define and describe L2 performance and L2 proficiency, or CAF for short, from the SLA perspective (see the special issue of *Applied Linguistics* on CAF in SLA research edited by Housen and Kuiken, 2009).

Compared with SLA research, language testing has a large number of very different purposes, and a number of different decisions can be made on the basis of the results of assessments. It is beyond the scope of this article to give a detailed account of language testing but it is useful to be aware of some of its main purposes and how they might relate to SLA.

The key question that determines the quality and trustworthiness of language assessment instruments is the degree to which testing (or assessment more generally) adheres to professional guidelines on good practice that have been

developed both for measurement in general – for example *The standards for educational and psychological testing* (American Educational Research Association, 1999) – or language assessment – e.g., *EALTA Guidelines for good practice in language testing and assessment* (European Association for Language Testing and Assessment, 2006) or *ILTA Guidelines for practice* (International Language Testing Association, 2007). Such guidelines strive to ensure that assessment is carried out reliably, validly and fairly, without which test results would be meaningless. In the context of SLA research, poorly designed data collection instruments would cause findings to lack interpretability and generalizability (Norris & Ortega, 2003, p. 717). Thus, following the principles of test design and validation developed in such fields as language testing, educational measurement and psychological testing will help SLA researchers make sure that they can depend on the data that they gather with their instruments.

There is no clear-cut way of categorizing purposes and types of decisions made on the basis of assessment but one simple division can be made between assessment related to specific language courses or curricula and proficiency testing that is detached from any teaching (see e.g. Bloom, Hastings, & Madaus, 1971; Huhta, 2008; Millman & Greene, 1993; Weigle, 2002). The latter is often carried out by examination organizations that certify learners' level of proficiency (e.g., Educational Testing Service, Cambridge ESOL, and the Goethe Institut are examples of such organizations). One specific type of proficiency testing relates to research in applied linguistics. Here, the aim of assessment is to enable applied linguists to gather information about learners' language skills as reliably and validly as possible (Huhta & Takala, 1999). As Douglas (1998) notes, when a language test elicits linguistic features of performance, it functions as an "SLA elicitation device" (p. 141). In an early discussion of the issues connecting language testing and SLA research, Byrnes (1987) points out how "data from proficiency testing should be able to provide information about the interrelationship between posited developmental stages and variational features, particularly in instructed SLA" (p. 48).

There are two main strands of research combining the SLA and language testing perspective (see also Hulstijn, Schoonen, & Alderson, this volume). On the one hand, researchers can collect L2 performance data from existing language tests and examination systems and analyze them for a number of linguistic features (e.g. Banerjee, Franceschina, & Smith, 2004; Norris, 1996, as cited in Norris & Ortega, 2009; Salamoura & Saville, this volume). On the other hand, they may choose a task-based approach and design a set of communicative tasks and rate them for communicative adequacy and at the same time analyze the linguistic features of learner performance. In what follows, we will discuss task-based approaches to SLA research and language testing by taking the solutions developed in the Cefling project as a case in point.

3. Tasks in SLA research and language testing

One of the key choices that needs to be made in any research attempting to combine both SLA and language assessment perspectives concerns the elicitation and measurement of L2 performance. In SLA research, performance data have always been collected in naturalistic conditions, that is, in the context of real language use. Alongside this strand of research, there is also a strong tradition of research relying on analytic tests and discrete point data collection instruments such as structured exercises and completion tasks (see e.g. Hulstijn, 1997). Beginning in the mid-1990s, a new, task-based research tradition investigating the multiplicity of cognitive and interactive factors on task performance (see e.g. Bygate, Skehan, & Swain, 2001; Robinson, 2001; Skehan, 1998; Skehan & Foster, 1997) began to emerge. This line of research has produced a number of studies and hypotheses about the influence of features such as task complexity or task difficulty on L2 development.

Task has been defined in a number of ways. In this chapter, it is defined as "an activity which requires learners to use language, with emphasis on meaning, to attain an objective" (Bygate, Skehan, & Swain, 2001, p. 11). A task typically involves holistic language use: "through engaging with the task, learners are led to work with and integrate the different aspects of language for a larger purpose" (Samuda & Bygate, 2008, p. 8). The learner's performance on the task can be assessed by focusing on specific features (such as grammatical accuracy or lexical complexity, or fluency or any number of features specified in the ALTE evaluation grids, for example). However, an essential feature of task is that it has a goal and an outcome: there is an objective that learners have to complete, and to do that, they have to use language (cf. Brindley, 1994/2009). How successful and efficient learners are in achieving the task's goal is called communicative adequacy: Pallotti (2009) notes that communicative adequacy is a dimension that most task-based studies in SLA have rarely looked at. For a learner to achieve the purpose of the task, i.e., to be communicatively adequate, it is not necessary for him or her to use correct, target-like language (Skehan, 2001, p. 167).

As Pallotti (2009, p. 597) goes on to point out, in open tasks, adequacy can be assessed by using qualitative ratings such as the CEFR scales. This is an approach adopted by the Cefling project (see also Gilabert et al., this volume). In the Cefling project, a key role of language testers was to ensure that the tasks and language proficiency ratings needed in the project were designed according to good language testing practice and that the quality of the ratings and data collection instruments was empirically ascertained.

There are, of course, all kinds of language tests, ranging from discrete point multiple choice tests to tests imitating real life communicative situations, and

their emphasis is not always only on meaning or even carrying out a particular communicative task. However, one of the central aims for language testing practice and research has been the development of tests and test items that measure language proficiency or communicative language ability as reliably and validly as possible (see e.g. Bachman, 1990; Bachman & Palmer, 1996).

The nature of language abilities has received considerable attention over the decades and is one of the main contributions of language testing research to applied linguistics (see e.g. Bachman & Cohen, 1998). Whether such abilities-oriented approaches to L2 proficiency can be used successfully to predict real-life task performances is another matter: scholars like Skehan (2001), for example, remain rather skeptical of whether the "codifying nature of the underlying competence-oriented models" (p. 167) can be used to make such predictions and have preferred to create alternative models of test performance (see also McNamara, 1996). On the other hand, language testers are well aware of the effect that different contextual factors, test taking strategies, test-taking processes and characteristics of the tasks and measurement instruments in general have on test performance, and studies focusing on such features are considered important in both language testing and SLA research. Yet, one might argue that one of the key purposes of most language testing research – when it relates to large-scale, high-stakes tests in particular – is designing tests and tasks that are as impervious as possible to such contextual factors; after all, the aim of such tests is to be able to generalize the test performance to other contexts.

Such motives may also reflect on the way L2 proficiency and L2 performance are conceptualized in language testing: preferably, both should be as stable as possible because in that way their measurement is easier, more reliable, generalizable and valid. Yet, from an SLA perspective, tasks need to be such that they elicit L2 performance that is variable enough. In usage-based approaches to SLA, in particular, L2 proficiency and L2 performance are considered inherently dynamic (e.g. de Bot, Lowie & Verspoor, 2005; Larsen-Freeman, 2002, 2009). The degree to which L2 performance is regarded as relatively stable and/or more or less systematically influenced by task structure or cognitive features varies from approach to approach. In the case of the Cefling project, what both SLA and language testing researchers share is a common understanding of L2 proficiency as something based on (or even having its origins in) communicative L2 use. Evaluating learner performance on communicative L2 tasks emerges as the common factor that both SLA research and language testing share an interest in.

Different types of tasks elicit different L2 performance. From the SLA perspective, to rely on just one kind of task in L2 data elicitation may lead into serious error, or at least yield only incomplete findings on the nature of L2 development, whether it is CAF, DEMfad (Martin et al., this volume), or a particu-

lar linguistic structure that is the focus of research. Similarly, from a language testing perspective, a range of different tasks should be used if the researcher wants to obtain a generalizable picture of learners' (writing or other) skills: there should be sufficient and valid evidence about learners' proficiency unless one is interested only in learners' ability to do one or two particular tasks (see e.g. Norris, Brown, Hudson, & Bonk, 2002). Finally, after L2 performances have been collected, the tasks should also be scrutinized to check whether they were functioning the way they were supposed to – in terms of their difficulty or complexity, for example – or whether they elicited the type of language that was expected.

In what follows, we will highlight some of the issues to be considered before and after L2 data elicitation when designing tasks for both SLA and language testing purposes by using data from Cefling as a demonstration. We will pay particular attention to the issues related to task design and assessment.

4. L2 writing proficiency from the SLA and language testing perspectives

Research on the development of L2 writing proficiency – L2 writing is used here as a cover term for both second and foreign language writing – has a fairly long history (see e.g. Grabe, 2001; Matsuda, 2005). A number of measures have been developed to capture various aspects of L2 development in writing, including complexity, accuracy and fluency. Wolfe-Quintero, Inagaki, & Kim (1998) reviewed a number of constructs and measures used to operationalize them. Recently, Norris and Ortega (2009) closely analyzed the notion of (linguistic) complexity and its measurement, taking a critical view of some of the suggestions made by Wolfe-Quintero et al. (see also Ortega, 2003; Pallotti, 2009). In sum, it appears that researchers are gradually beginning to take into account the multidimensional and multicomponential nature of L2 proficiency and development both in designing research as well as interpreting the findings: the constructs and measures which seem particularly adapted for capturing the growth of proficiency across, for example, the beginning stages of L2 writing proficiency, may not be as suitable for the more advanced levels of writing, or for L2 speaking, for that matter, or vice versa. Language testing has, of course, for a long time looked at L2 proficiency as multicomponential, although this conceptualization has been used more for improving test construction than for understanding L2 development.

There are a number of studies that have looked at L2 writing proficiency by examining data from the already existing language tests such as IELTS or Cambridge ESOL examinations (see e.g. Banerjee et al., 2004; Salamoura & Saville, this volume). However, not much attention has been paid to the effect

of task type, task complexity or task structure on learner performance in studies of task-based L2 writing from the SLA perspective, in contrast to L2 speaking. In one of the first studies of this kind, Kuiken and Vedder (2008) examined the effect of task complexity on syntactic complexity, lexical variation, and accuracy in the written productions of low-proficiency and high-proficiency L2 Italian and L2 French learners. Students' L2 proficiency was assessed by using a separate cloze-test; as a task, students had to write a letter to a friend helping them choose a holiday destination. No indication of significant interaction between task complexity and lexical variation and syntactic complexity was found in the study; however, an increase in task complexity led learners to produce a text which was more accurate. As Kuiken and Vedder (2008) conclude, this finding can be interpreted as a function of an increased control of the L2 system that a more complex task may require from learners rather than support for any of the existing models of task performance.

In fact, as Kuiken and Vedder (2008) note, the relationship between task type or task complexity and L2 writing performance is not at all clear: as an example, they point to a study by Hamp-Lyons and Mathias (1994) showing that, contrary to expectations, students' L2 performance was lower on personal and expository writing tasks, which were judged as easier by experts, and better on argumentative and public tasks, which were rated more difficult.

Various types of L2 writing tasks have been used in research as data collection instruments. However, it appears that at least in the U.S., there is a difference between the types of writing tasks most commonly used in foreign language and ESL writing classes. In her review, Reichelt (1999) notes that in the former, the focus is typically on creative and expressive, non-academic writing while the latter has tended to involve essays or compositions or other argumentative or descriptive texts commonly used in academic contexts. That L2 writing tasks of personal or expressive nature – essays on hobbies, family life, friends, holidays, and personal letters – are preferred in foreign language classrooms is probably true for other countries as well.

Some of the studies on L2 writing development include studies which have an explicit link to language testing (whether any of these studies can be regarded as task-based in the sense that today's research uses the term is unclear). Valdés, Haro, & Echevarriarza (1992) analyzed the ACTFL scale for writing in detail and then examined short essays written by novice, intermediate and advanced L2 Spanish learners attending a college-level language program. Their findings suggested that the students' L2 proficiency interacted with their L1 writing skills so that more proficient L2 learners were able to write more competent and coherent essays. Henry (1996) investigated the early L2 Russian writing development of novice and intermediate L2 learners by analyzing their essays for fluency, syntactic fluency and accuracy and by contrasting them to the

ACTFL writing proficiency descriptors. For various reasons, neither study used the ACTFL proficiency scales to rate the students' texts; instead, either years of study (Valdés et al., 1992) or a type of global assessment (Henry, 1996) was used to determine the writers' proficiency level. A separate attempt was made by Henry (1996), however, to evaluate learner performances as to their communicative adequacy by asking the judges to decide whether the essays would be understandable (p. 315).

As Skehan (2001) notes, in task-based SLA research using rating scale measures is not typical; rather, researchers have tended to use various operationalizations of constructs such as CAF. Similarly, Pallotti (2009) notes that there are few studies using qualitative ratings to evaluate the communicative adequacy of tasks. Studies reported by Wigglesworth (1997, 2001) are an early exception: Wigglesworth studied variability in L2 speaking performances across five different tasks. Tasks targeted at different proficiency levels were rated with a rating scale used to assess the L2 English proficiency of adult immigrants to Australia. Wigglesworth examined task variability by looking at learners' performance on tasks and their evaluations of task difficulty, and conducted multifaceted Rasch analyses on the data by using the statistical modeling program FACETS (e.g., Linacre, 2010) (see also McNamara, 1996). Her findings reveal a complex interaction of a number of factors such as task structure and task conditions (interlocutors' actions and familiarity with the topic).

In sum, there are few studies so far attempting to use specific proficiency scales to assess the level of task-based L2 writing performances and then utilizing those judgments as independent variables in order to determine particular and general linguistic features typical for those levels. In their review of prototype task-based performance tests called ALP (see Norris, Brown, Hudson, & Yoshioka, 1998), Norris et al. (2002) found, among other things, that careful simulations of L2 communication tasks could effectively elicit a wide range of L2 performances, and that average performance patterns based on the rating scales reflected expected differences among the ability levels of participants. Findings such as these support the idea that ratings of learner performances on a set of communicative tasks can be used as an indication of learners' ability to accomplish such tasks, and that task-independent ratings can be used as an indication of learners' general abilities in performing the range of test tasks (Norris et al., 2002, p. 415).

In the Cefling project (see Martin et al., this volume), L2 Finnish and L2 English data were collected from young and adult L2 learners by using a set of communicative L2 writing tasks. Learner performances were rated by using two scales, the CEFR (young learners) and the National Certificates (adult learners) examination scales for writing, and the Finnish *National Core Curriculum for Basic Education* (2004) scales (young learners). 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). In the future – work on corpora is still in progress — the data collected in the project will 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.[1]

5. Designing and selecting communicative L2 writing tasks

Designing and selecting tasks that are relevant for both SLA and language testing research is particularly challenging. Tasks should elicit L2 data that can be analyzed for differences in linguistic features, while it should also be possible to rate task performances based on the communicative adequacy of those performances. The latter dimension imposes conditions of its own on the type of tasks that can be used in data elicitation: from the outset, learners' level of L2 proficiency either constrains or supports their ability to carry out L2 tasks successfully. In language testing, this has been taken into account by using different tests and tasks for e.g. beginning, intermediate and advanced learners.

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 was carefully estimated in advance so as to make the tasks as suitable for them as possible.

Since our intention in the project was to collect data for research purposes from learners from all proficiency levels, from both adult and young learners, our task was challenging indeed. However, what both helped and constrained us in the design and selection of L2 tasks was the existence of a data set collected from adult L2 learners available from the National Certificate (NC) language examination system. The NC is based at the Centre for Applied Language

1 It is important to note that initially, only those learner performances were included in the corpora that received CEFR ratings that were sufficiently reliable (this was done because the ratings were used as an independent category variable in the study). Tracking the performances of individual language learners across different tasks, no matter how they were rated, will be possible, however, at later stages of research.

Studies at Jyväskylä; it has stored L2 data in digital format from test takers in several languages in its data base since 2004. The data base is available for researchers through a web portal at the Finnish Social Science Data Archive (http://www.fsd.uta.fi/english/index.html). The tasks used to collect these data served as a starting point – whatever data were going to be collected from teenaged language learners had to match the existing data base as to the type and nature of the task.

A considerable amount of time and effort went into designing or selecting tasks that would reflect a variety of features relevant for the development of communicative L2 writing. It was also necessary to take into account the scales that were going to be used for L2 writing assessment: the CEFR and the Finnish National Core Curriculum scales. The NC tasks were designed for adults, and since cognitive, interactive and learner factors are influenced by age and experience gained through activity in a wide variety of social contexts (see Vähäpassi, 1982; Weigle, 2002), they were not necessarily suitable for younger L2 learners writing in school context.

The issues that needed to be taken into account included the proficiency level the task was aimed at, the topic and domain of the tasks, as well as the genre and functions of the language we expected the tasks to generate. In many ways, the decisions made in the project concerning the nature and type of tasks reflect a striving for communicative authenticity and adequacy of tasks; yet, an attempt was also made to make sure that tasks would elicit particular linguistic structures (e.g. locative expressions, verb forms, relative clauses, questions, negation).

It was also felt that the tasks should be communicative and that they should have some measure of authenticity in terms of text types and processes needed in completing the tasks. In Finland, it seems that users mostly engage in communicative L2 writing outside the classroom (e.g. on the Internet) (Luukka et al., 2008).

Based on such considerations, a set of tasks was designed and extensively piloted by administering tasks to 7th graders in a number of schools. Piloting the tasks was an essential part of the process and yielded much valuable information. The final set of tasks consisted of five communicative tasks representing a variety of text types, functions and register; most tasks belonged to the personal domain. 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 (see Table 1 for the features of the tasks and Appendix 1 for the tasks themselves). For logistical reasons, Tasks 1 and 2 were alternates: it was felt that both teachers and students were easily able to fit four tasks in their lessons during one term but no more than that. In the end, the participants did either Tasks 1, 3-5 or Tasks 2-5.

One last issue we want to raise has to do with the nature of prompts used in data elicitation. Quite often, prompts for L2 writing tasks of this type include target language data: the instructions may be in English, or the task includes newspaper articles or letters to the editor or other such items in the target language. However, from an SLA perspective, in order to obtain a reliable description of what the linguistic repertoire of learners from each proficiency level is, it may be better to exclude such prompts (see e.g. Grant & Ginther, 2000). On the one hand, it is difficult to say what effect, if any, such recycled fragments of L2 might have in the statistical analysis of data; in the worst case, it could potentially seriously distort the analysis of such dimensions of L2 performance as CAF. Be that as it may, at least for the L2 English tasks, a decision was made to ensure that the task prompts contained as little L2 input as possible.

Table 1. The domain and register and the functions and linguistic structures that the tasks in Cefling were expected to elicit.

Tasks	Domain, register and functions
Task 1 Email message to a friend	• *Domain and register*: personal life, informal • *Functions*: apologizing, argumentation or expressing obligation/necessity (answering to *why* question, negation), requesting, giving information • *Linguistic structures*: questions, negation, tense, locative expressions
Task 2 Email message to a teacher	• *Domain and register*: personal life, school, informal or formal • *Functions*: argumentation or expressing obligation/necessity, asking for information • *Linguistic structures*: questions, negation, tense, locative expressions
Task 3 Email message to an internet store	• *Domain and register*: personal, public, formal • *Functions*: introducing oneself, complaining, requesting correction, suggesting solution • *Linguistic structures*: questions, negation, aspect
Task 4 Opinion	• *Domain and register*: personal, everyday life, school, informal or formal • *Functions*: expressing an opinion, arguing for or against • *Linguistic structures*: tense, aspect, locative expressions
Task 5 Story	• *Domain and register*: personal life, informal • *Functions:* describing and narrating, argumentation or expressing an opinion, liking or dislike • *Linguistic structures:* tense, agreement

6. Designing the assessment procedure

In much of language assessment, the focus is very much on finding out what L2 learners can and cannot do communicatively, i.e., their ability to do things with the language. The main interest in many types of language assessment is to place learners at certain levels of L2 proficiency. In the kind of tasks used in the project, this always involves human raters since they are the ones who decide – based on a set of descriptors – how well L2 learners succeed in completing the tasks (see Weigle, 2002). Such a ratings process is always subjective in the sense that it depends on how raters understand, interpret, and put into practice the scales and their descriptors. Usually, but not always, raters are also given a set of benchmarks, a set of performances that serves as prototypical examples of specific performance levels.

The selection of scales and training raters in how to use them is crucial for a research project using the proficiency levels as independent variables, in particular. Two issues about rating scales should be mentioned at this point. In our view, none of the CEFR scales is a proper rating scale comparable to most scales specifically designed for rating purposes (see Alderson, 1991). Using CEFR scales for rating is therefore a challenging exercise and it is in principle uncertain to what extent particular CEFR scales actually enable reliable rating to take place even though the careful design of these scales gives cause for some optimism (North, 1996/2000). In comparison, the Finnish curriculum scale for writing appears more 'rater friendly', with references to linguistic features and to deficiencies in learners' performance. However, no published research appears to exist yet on how well the CEFR writing scales or the Finnish curriculum scales actually work for rating purposes. One of the aims of the Cefling project was to investigate how the scales function as an evaluation tool of learner performances.

The second issue is the effect on the ratings of the linguistic features of the rated performances. On what features do raters base their ratings? Paying too much attention to linguistic features could introduce circularity in the reasoning: proficiency levels are determined on the basis of linguistic features, and these features are, in their turn, used in defining the levels. The choice of use-oriented CEFR scales was intended to minimize this danger: they focus on communication with practically no references to specific linguistic features. The use of several raters instead of just one may also address this problem, as it is likely to reduce the effect on the ratings of very linguistically-oriented raters. While it makes sense to try to minimize linguistically-influenced rating of performances when the aim is to place learners on proficiency levels on the basis of their ability to use language, we believe it is ultimately impossible to totally remove the effect of linguistic features from any rating of language performances.

Two proficiency scales – the CEFR scale and the Finnish National Core Curriculum scale – were selected for placing learners' performances in the Cefling study. The CEFR scale used in the Cefling project is a combination of six CEFR scales for writing (see Appendices 1 and 2). The Finnish National Core Curriculum (NCC) scale is an adaptation of the CEFR scale for the purpose of defining targets for learning, teaching and assessment in the foreign language curricula for primary and secondary education in Finland (see Appendix 3). It is the official reference scale that teachers in the Finnish schools should apply when assessing their pupils' foreign and second language proficiency. There are no language-specific versions of the NCC scale but the same scale is used for target-setting and assessment in all foreign and second languages covered by the national curriculum. As regards content, the NCC differs from the CEFR scales in that it has no genre-specific level descriptors for different text types. Importantly, the level descriptors of the NCC scale make explicit references to general linguistic characteristics such as accuracy and complexity, vocabulary and structures, while the CEFR scales do not (see Hildén & Takala, 2007).

Good rating scales are necessary but not sufficient for ensuring reliable and valid rating of learner performances. Design of the entire rating process, training of the raters and selection of benchmark examples are also important. There are a number of recommendations in language testing literature as to how such a process should unfold (e.g. Alderson, Clapham, & Wall, 1995). Of course, it also helps to have experience in organizing large-scale rater training, and to have raters who have previous rating experience in using similar scales. For example, the training of raters in Cefling was a multi-stage process that consisted of one full session, a brief update meeting and self-study. The training process was also used to create the final, official benchmarks that were used in the final phase of task assessment.

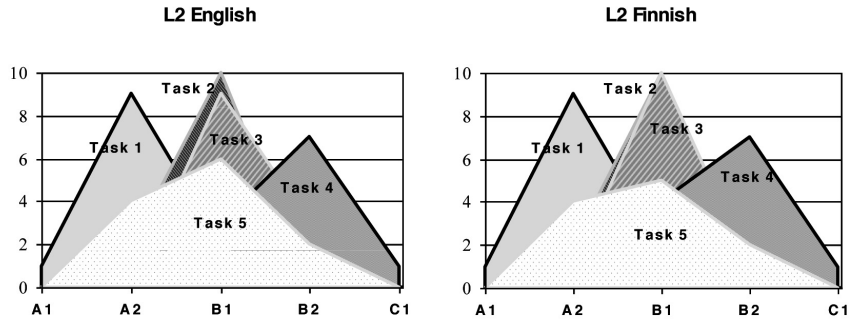
7. Evaluating task difficulty based on rating data

In this section, we will examine the tasks in more detail, focusing on the 7th-9th graders' actual performances on L2 English tasks in terms of the ratings they received on the tasks. The final data collection began in the autumn term of 2007 and lasted well into the spring term of 2008. A total number of 3427 L2 Finnish and L2 English performances were collected from 7th to 9th graders from schools. A number of scripts (N = 1789) were selected for assessment by a team of trained raters (N= 9 for English, N = 11 for Finnish). Because of the large number of texts, rating was divided among the team members so that each script was rated by four (L2 English) or three (L2 Finnish) raters. Incomplete

ratings were not considered a major problem since multifaceted Rasch analyses (see below) are capable of handling incomplete rating designs as long as there are enough links between raters and tasks in the design. The rating of the performances was based on the learners' original handwritten scripts, which were photocopied and delivered to the raters together with the instructions, scales, and the benchmarks.

To begin with, the raters' perception of tasks was placed under scrutiny. Before the actual rating of learner performances, the suitability of the tasks for L2 learners at various CEFR and NCC levels was assessed by 12 of the raters. For each L2 Finnish and L2 English task, the raters were asked to indicate the level they thought the task would be most and second-most suitable for, as well as the lowest and highest proficiency level that the task could be used for. Figure 1 shows the level the tasks were best suited for in both L2 English and Finnish according to the raters. The frequency distributions are very similar for both languages. The raters (N=12) rated Task 4 as most suitable for B2, while Task 1 was considered as the easiest for both languages – most raters considered it as the most suitable for A2. Tasks 2, 3 and 5 were considered best suited for B1, with Task 5 more skewed towards A2.

Figure 1. Perceived task difficulty by raters (n=12) for L2 English and Finnish across all tasks: the level the task is most suited for.



In evaluating the interaction of tasks and judgements of learners' proficiency level, various statistical analyses were used. Figure 2 shows how the L2 English and L2 Finnish task performances were rated. The box plots show the arithmetic means of ratings of L2 English and L2 Finnish performances and their variance across the five tasks.

Figure 2. The ratings given for L2 English (on the left) and L2 Finnish (on the right) performances across the five tasks (T1-T5); dots represent outliers in the data set.

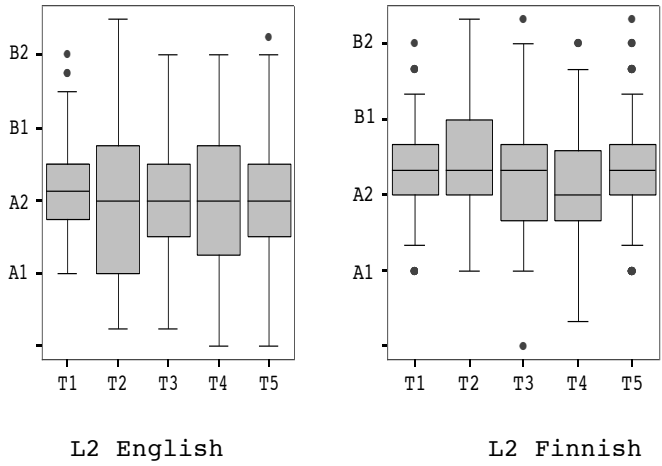
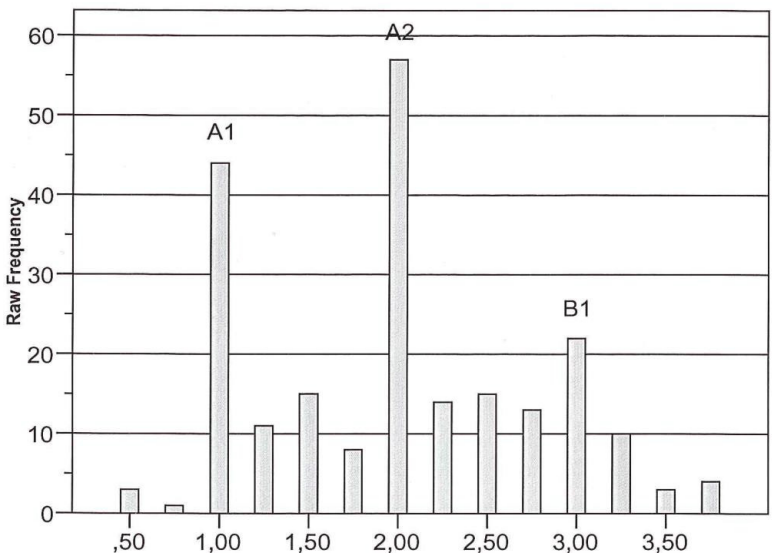


Figure 3. The median of median ratings on Tasks 1 – 5 in L2 English.



Looking at these results alone, it appears that the level of task performances in both L2 English and L2 Finnish was most often A2, with L2 Finnish learners receiving slightly higher ratings apart from Task 4. The distribution of the medi-

an ratings for performances in L2 English support this (see Figure 3). A2 is the most frequent median rating, while A1 appears to form a second peak in the distribution.

Figure 2 also reveals that there was greater variance among L2 English learner performances in terms of the ratings they received than among L2 Finnish learners. To gain a more in-depth understanding of task variability, multifaceted Rasch analyses were also calculated for the L2 performances by using FACETS (Linacre, 2010). A Rasch analysis takes into account a number of facets, or elements, in test performance, to check how the test taker's ability – in this case L2 learners' proficiency – interacts with other factors such the raters' relative severity or leniency, which makes it possible to analyze task difficulty from multiple perspectives. In language testing research, this method has been used, for example, to study different types of performance tasks (McNamara, 1996), the rating of L2 oral discussion tasks (Bonk & Ockey, 2003), or the interaction of teacher, peer- and self-assessment on the rating of EFL writing tasks (Matsuno, 2009). Wigglesworth (2001) used it to analyze task variability in oral L2 performance data.

The facets included in the analysis were the ratings for L2 learners, the raters, and the tasks. Figures 4 and 5 show the order of task difficulty in L2 English and L2 Finnish, from the easiest down. Zero stands for the centre of the range of item (task) difficulty (default origin). The higher negative value indicates a higher difficulty level for the task.

Table 2. The results of the Rasch analyses (in logit points) run on Tasks 1-5 in L2 English and Finnish.

Logit points	L2 English	Logit points	L2 Finnish
.38	Task 3	.29	Task 2
.00	Task 5	.25	Task 1
.02	Task 1	.20	Task 3
-.10	Task 2	-.29	Task 5
-.27	Task 4	-.46	Task 4

Table 2 shows the order of task difficulty in L2 English and L2 Finnish, from the easiest down. In a multifaceted Rasch analysis, a common measurement scale, called a logit scale, is created that allows the facets of interest, such as learners' proficiency, task difficulty, and raters' severity, to be placed on the same scale and, thus, directly compared. The logit scale is an interval scale, i.e., the distance between the scale points is exactly the same across the scale, unlike the CEFR and NC rating scales, which are ordinal scales in which the distance

between, say, A1 and A2 is probably not the same as the distance between B1 and B2 or C1 and C2. Measuring language proficiency, task difficulty, and later severity on an interval, logit scale thus provides us with more precise information of these phenomena.

Figure 4. A multifaceted Rasch analysis of L2 English performances across Tasks 1-5.

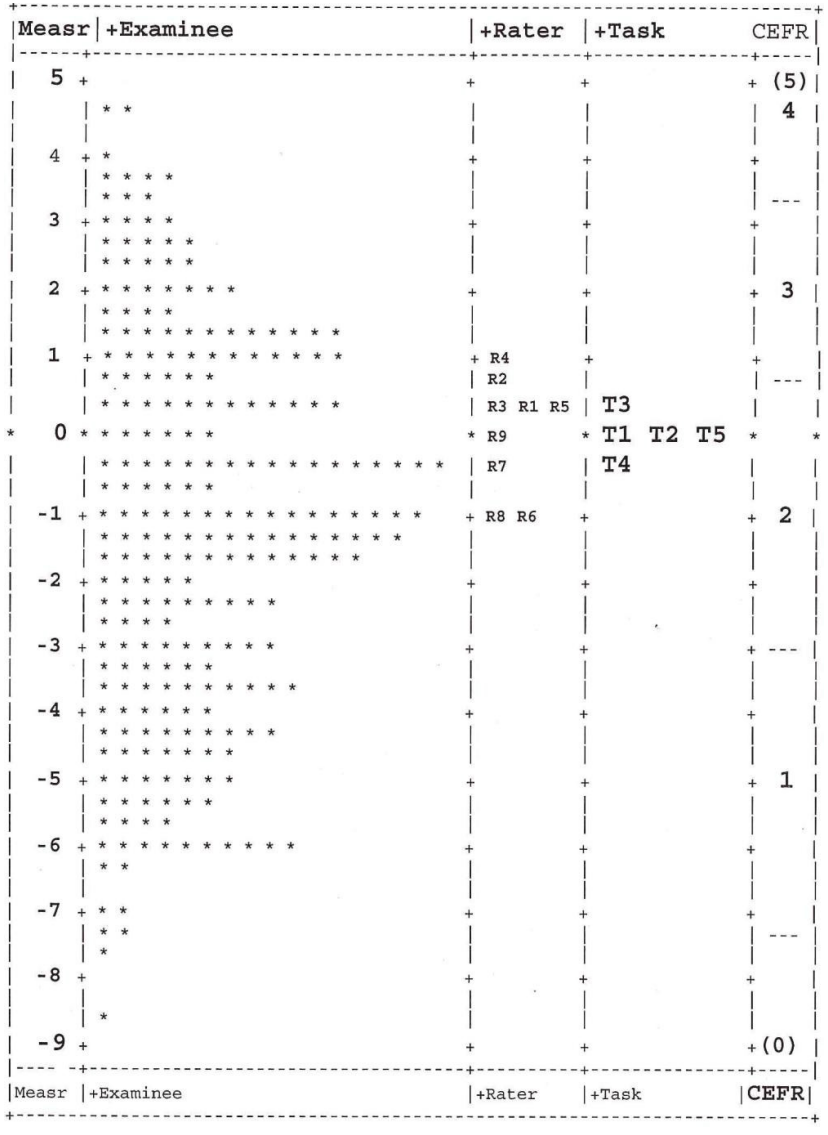
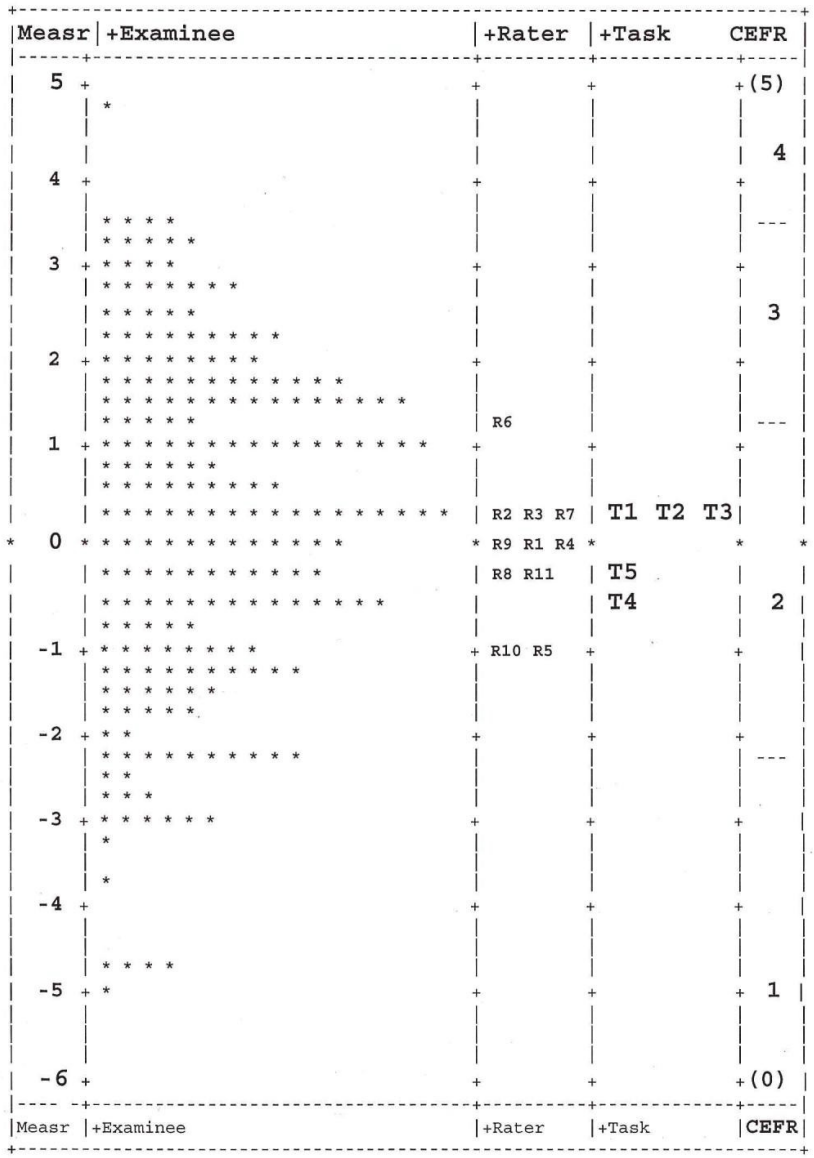


Figure 5. A multifaceted Rasch analysis of L2 Finnish performances across Tasks 1-5.



As Figures 4 and 5 also show, the tasks do not differ very much in terms of difficulty because the minimum and maximum logit values only differ by .65 points for English and .75 for Finnish. The analysis also reveals that there is both within group and between group variation across task performances. Most

notably, Task 2 seemed to be more difficult for L2 English learners than it was for L2 Finnish learners while Task 4 seemed to be difficult for both groups of L2 learners. However, the differences between tasks are rather small.

In sum, the mean L2 ratings and the results of multifaceted Rasch analyses both support the conclusion that the tasks designed for Cefling were quite successful in targeting the expected proficiency levels of the young language learners participating in the project. The statistical analysis of L2 learners' performances reveals that both L2 Finnish and L2 English learners received an average rating of approximately A2 in the 7th - 9th grades (aged 12-16) when the data were collected. Both in L2 English and L2 Finnish, the level of tasks corresponds to the borderline between A2 and B1. The relatively low variation in learner performances and task difficulty reflects the fairly narrow range of proficiency levels elicited by tasks. There could have been a greater number of higher level (B1 and above) performances in the data. The low number of such performances was likely due to the relatively young age of the participants and their limited experience as language learners. Adult performances included in the NC data base will likely yield more varied data.

8. Conclusion

In this chapter, we have been concerned with issues relevant for designing and assessing L2 tasks for SLA research purposes. What should be the most appropriate tasks when collecting evidence of the development of L2 writing from particular proficiency levels?

Task-based SLA research has traditionally been concerned with the effect of task features (structure, complexity, planning time etc.) on L2 performance and development. To gain reliable and valid results, task-based SLA research aims at controlling a number of such features and studying their effect on L2 performance (Norris & Ortega, 2009). An overall generalizable assessment of learners' level of L2 proficiency in terms of communicative success or adequacy has usually not been a major concern; sometimes, various a priori measures or categories have been used to estimate participants' level of language proficiency (course level, years of study), sometimes, it has been assessed by using L2 data elicited during the task performance. Language testing – performance-based testing in particular – has slightly different concerns: first, that the tasks should be designed with a particular proficiency level in mind; second, that the linguistic performance should be grounded within a particular L2 construct that can be assessed by using a reliable and valid rating scale; and third, more than one task and one rater should be used to elicit data. One of the advantages of projects like Cefling is that they have a rating system based on learner performances across several tasks.

A research design that uses several tasks and several raters allows researchers to be more certain about learners' level of proficiency. Rating learners' performances across a number of tasks with reference to e.g. the CEFR proficiency scale allows SLA researchers to define learners' proficiency with more precision and a firmer basis than by relying on the number of years studied or courses taken, for instance. Using more than one task and one rater follows sound measurement principles: several tasks cover learners' proficiency better than one task and this is likely to result in a more generalizable picture of the learner's proficiency – unless one is interested in performance on certain specific tasks only. Measurement instruments always introduce some method effect – or to put it differently, a learner's performance is always a combination of his or her skills and the effects of the task (and a host of other factors). Applying several tasks reduces the effect that any one task format has on the learner's performance. The use of several raters functions basically in the same way although this is usually discussed in terms of reliability: the effect of any one, possibly 'unusual', rater on the outcome is reduced when several are used. Furthermore, if there are enough raters, a rater who is too idiosyncratic can be removed from further data analysis by using the multifaceted Rasch analysis program FACETS.

The use of several tasks and raters also gives SLA researchers more options in how they define the data from which they draw conclusions about language learning. It is possible to include in the analyses only those learners or task performances whose rating fulfils specific quality criteria and leave out from the analyses those learners whose rating is considered too unreliable or otherwise problematic. That is, if one wants to describe the development of linguistic features across different proficiency levels, such as the CEFR levels, it is possible to use only those learners who have been successfully (reliably, validly) placed on specific levels. As a consequence, the validity of the findings about language development improves.

There are different ways to decide which methods represent particular proficiency levels reliably. The one that is used in the first Cefling analyses of linguistic features and their relationship to different proficiency (CEFR) levels is based on direct observation of rater agreement. For the linguistic analyses conducted in Cefling, only those samples of writing were chosen on which the majority of the raters had reached sufficient agreement: three out of four raters in English and two out of three in Finnish rated the texts as belonging to the same proficiency level. An additional criterion was also used: the remaining rater should not deviate from the others by more than one CEFR level up or down. If these criteria were not fulfilled, the sample was not included in the L2 data set to be used for linguistic analyses. At the moment, about 63% of the rated performances in English and 92 % of the rated performances in Finnish are included in the data set for linguistic analyses.

As long as the data set is large enough, it is possible to study the effect of applying other criteria for keeping or rejecting L2 writing samples on SLA data analysis. As is probably the case in most other comparable studies that use more than one rater to assign proficiency level to learners and their performances, it is possible to assign a proficiency level to all samples by using various statistical measures – such as the mean or median rating – and thereby include them in the analyses. More sophisticated analyses such as FACETS can also be used to create a value for each learner that is similar to the mean, for example, but one that also takes into account the difficulty of the tasks he or she has taken and the severity or leniency of the raters who happened to rate them. These 'ability values' could then be related to proficiency scales (such as the CEFR scale) by studying what the analyses tell us about the relationship between the analyzed learners, tasks, raters and the scale(s) used in the rating. Possibly some standard setting procedures would also be needed to confirm the translation of such an ability value scale into the scale in question (see e.g. Kaftandjieva, 2004).

Very little is known about the effect of applying different criteria for the inclusion or exclusion of data on the results of linguistic analyses. The extensive amount of L2 data collected in research projects such as Cefling or others with similar design enables researchers to check whether the linguistic analyses will remain the same when all of the cases – instead of only some 60-70% of the samples – are included and their CEFR level is determined in one of the possible ways described above.

Finally, the analysis of linguistic features – CAF or various linguistic structures – is needed for researchers to be able to tell whether tasks were successful in eliciting the kind of data they were expected to. The ultimate aim of SLA research is to shed light on the nature of L2 development and the dynamic processes that underlie L2 proficiency. In the case of Cefling, linguistic analyses are still in progress.

The linguistic analysis of the data is doubly important since – as we are acutely aware – the assessment of communicative L2 performances cannot be wholly separate from the linguistic features such as complexity, fluency, or an increasing accuracy of a given linguistic structure in the same performances. So how to live with the potential circularity built in our research design? This methodological conundrum has implications for our understanding of L2 development as well. The ultimate aim of the projects combining SLA and language testing perspectives may be to discover the linguistic features that characterize proficiency levels, regardless of how they are determined. Nonetheless, the question is whether the findings will be more like patterns and probabilities of occurrence rather than a list of features, yet to be discovered, that with a 100% certainty are always present at a particular level. And if such features are discov-

ered, what will they tell us about the nature of the assessment process and ratings? What do raters base their judgments on?

At the moment, such linguistic features remain to be fully discovered, but based on the findings presented in this volume it seems more likely that such overarching features, if present, will be rather general, on the order of nouns and verbs or words expressing (agentive or other) relations. These features will be present at level A1 with an increasing fluency and complexity of relations, and at higher levels of proficiency there will be an increasingly target-like use of L2 repertoire, to varying degrees. As Norris and Ortega (2009) point out, a finer understanding of the multidimensional and multicomponential nature of the development of L2 proficiency is not only desirable but also absolutely necessary on both theoretical and methodological levels. A conceptualization of L2 tasks as a unit of activity with multiple dimensions and components – such as task completion, linguistic accuracy, situational or discourse-pragmatic appropriateness – that learners can carry out with a varying degree of success from both communicative and linguistic perspectives is both a challenge and a necessity for future SLA and language testing research.

References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 71–86). London: Macmillan.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, American Psychological Association & National Council on Measurement in Education.

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

Bachman, L. F., & Cohen, A. D. (1998). *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Banerjee, J., Franceschina, F., & Smith, A. M. (2004). Documenting features of written language production typical at different IELTS band score levels. *IELTS Research Reports*, *7*, 249–309.

Bloom, B. S., Hastings, J. T., & Madaus. G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89–110.

Brindley, G. (1994/2009). Task-centred language assessment in language learning. The promise and the challenge. In J. Norris, M. Bygate, & K. Van den Branden (Eds.), *Task-based language teaching. A reader* (pp. 435–454). Amsterdam/Philadelphia: Benjamins.

Bygate, M., Skehan, P., & Swain, M. (2001). Introduction. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks. Second language learning, teaching and testing* (pp. 1–20). Harlow, England: Longman/Pearson Education.

Byrnes, H. (1987). Proficiency as a framework for research in second language acquisition. *The Modern Language Journal*, *71*, 44–49.

Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

De Bot, K., Lowie, W., & Verspoor, M. (2005). *Second language acquisition. An advanced resource book*. London: Routledge.

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. (2007). The effects of task complexity on fluency and functional adequacy of speaking performance. In S. Van Daele, A. Housen, M. Pierrard, F. Kuiken, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 53–63). Brussels: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten.

Douglas, D. (1998). Testing methods in context-based second language research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 141–155). Cambridge: Cambridge University Press.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

European Association for Language Testing and Assessment (2006). *EALTA Guidelines for good practice in language testing and assessment*. Retrieved from http://www.ealta.eu.org/guidelines.htm

Grabe, W. (2001). Notes toward a theory of second language writing. In T. Silva & P. K. Matsuda (Eds.), *On second language writing* (pp. 39–57). Mahwah, NJ: Lawrence Erlbaum Associates.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, *9*, 123–145.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, *3*, 49–68.

Henry, K. (1996). Early L2 writing development: A study of autobiographical essays by university-level students of Russian. *The Modern Language Journal*, *80*, 309–326.

Hildén, R., & Takala, S. (2007). Relating descriptors of the Finnish school scale to the CEF overall scales for communicative activities. In A. Koskensalo, J. Smeds, P. Kaikkonen, & V. Kohonen (Eds.), *Foreign languages and multicultural perspectives in the European context/Fremdsprachen und multikulturelle Perspektiven im europäischen Kontext* (pp. 291–300). DICHTUNG, WAHRHEIT UND SPRACHE. LIT-Verlag.

Housen, A., & Kuiken, F. (Eds.). (2009). Complexity, accuracy and fluency (CAF) in second language acquisition research [Special issue]. *Applied Linguistics, 30*(4).

Huhta, A. (2008). Diagnostic and formative assessment. In B. Spolsky & F. Hult (Eds.), *Handbook of educational linguistics* (pp. 469–482). Malden, MA: Blackwell.

Huhta, A., & Takala, S. (1999). Kielitaidon arviointi. In K. Sajavaara & A. Piirainen-Marsh (Eds.), *Kielenoppimisen kysymyksiä* (pp. 179–228). Jyväskylä: University of Jyväskylä, Centre for Applied Language Studies.

Hulstijn, J. H. (1997). Second language acquisition research in the laboratory: Possibilities and limitations. *Studies in Second Language Acquisition*, *19*, 131–143.

International Language Testing Association (2007). *ILTA Guidelines for practice*. Retrieved from http://www.iltaonline.com/index.php?option=com_content&view=article&id=122&Itemid=133

Kaftandjieva, F. (2004). *Standard setting. Section B of the Reference Supplement to the preliminary version of the Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment.* Strasbourg: Council of Europe.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, *17*, 48–60.

Larsen-Freeman, D. (2002). Language acquisition and language use from a chaos/complexity theory perspective. In C. Kramsch (Ed.), *Language acquisition and language socialization* (pp.33–46). London: Continuum.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*, 579–589.

Linacre, J. M. (2010). FACETS. Version 3.66.3. [Computer software]. Chicago: MESA Press.

Luukka, M.-R., Pöyhönen, S., Huhta, A., Taalas, P., Tarnanen, M., & Keränen, A. (2008). Maailma muuttuu – mitä tekee koulu? Jyväskylä: University of Jyväskylä, Centre for Applied Language Studies.

Matsuda, P. K. (2005). Historical inquiry in second language writing. In P. K. Matsuda & T. Silva (Eds.), *Second language writing research. Perspectives on the process of knowledge construction* (pp. 33–48). Mahwah, NJ: Lawrence Erlbaum Associates.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing*, *26*(1), 75–100.

McNamara, T. F. (1996). *Measuring second language performance*. London and New York: Longman.

Millman, J., & Greene, J. (1993). The specification and development of tests of achievement and ability. In R. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 335–366). Phoenix, AZ: Oryx Press.

National Core Curriculum for Basic Education (2004). Helsinki: Finnish National Board of Education.

Norris, J. M. (1996). *A validation study of the ACTFL Guidelines and the German speaking test* (Unpublished MA thesis). Honolulu: University of Hawai'i.

Norris, J. M. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing*, *19*, 337–346.

Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk, W. (2002). Examinee abilities and

task difficulty in task-based second language performance assessment. *Language Testing*, *19*, 395–418.

Norris, J. M., Brown, J. D., Hudson, T. D., & Yoshioka, J. (1998). *Designing second language performance assessments.* Honolulu: University of Hawaii Press.

Norris, J. M., & Ortega, L. (2003). Defining and measuring SLA. In C. Doughty & M.H. Long (Eds.), *Handbook of second language acquisition* (pp. 716–761). Malden, MA: Blackwell.

Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*, 555–578.

North, B. (1996/2000). *The development of a common framework scale of language proficiency* (Doctoral dissertation). London: Thames Valley University. (Reprinted 2000, New York: Peter Lang).

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*, 492–518.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics, 30*, 590–601.

Reichelt, M. (1999). Toward a more comprehensive view of L2 writing: Foreign language writing in the U.S. *Journal of Second Language Writing*, *8*, 181–204.

Robinson, P. (2001). Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics*, *22*, 27–57.

Samuda, V., & Bygate, M. (2008). *Tasks in second language learning*. Basingstoke: Palgrave.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks. Second language learning, teaching and testing* (pp. 167–185). Harlow, England: Longman/Pearson Education.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*, 185–211.

Vähäpassi, A. (1982). On the specification of the domain of school writing abilities. In A. C. Purves & S. Takala (Eds.), *An international perspective on the evaluation of written composition* (pp. 265–289). Oxford: Pergamon.

Valdés, G., Haro, P., & Echevarriarza, M. P. (1992). The development of writing abilities in a foreign language: Contributions toward a general theory of L2 writing. *The Modern Language Journal*, *76*, 333–352.

Van den Branden, K. (Ed.) (2006). *Task-based language education. From theory to practice*. Cambridge: Cambridge University Press.

Van den Branden, K., Bygate, M., & Norris, J. M. (Eds.). (2009). *Task-based language teaching. A reader*. Amsterdam/Philadelphia: Benjamins.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, *14*, 101–122.

Wigglesworth, G. (2001). Influences on performances in task-based oral assessments. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks. Second language learning, teaching and testing* (pp. 186–209). Harlow, England: Longman/Pearson Education.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy & complexity*. Technical Report #17. University of Hawai'i: Second Language Teaching & Curriculum Center.

APPENDIX 1
Task prompts in CEFLING (translated into English from the Finnish originals)

Task 1. Note to a friend	You've set up a meeting with your English-speaking friend at a café. However, something has come up and you have other things to do. Send an email message to your friend. • Explain why you can't come. • Suggest a new time and place. Remember to **begin** and **end** the message appropriately. Write **in English** in clear characters in the space below.	You've agreed with your friend that you will meet after school at a café. However, you have other things to do. Send an email message to your friend. • Tell why you can't come. • Suggest a new time and place. Write **in Finnish** in clear characters in the space below. Remember to **begin** and **end** the message appropriately.	more basic syntactic structure more frequent vocabulary more basic morphology
Task 2. Message to your teache	You've been away from school for a week. Soon you'll have an English exam. Your teacher, Mary Brown, speaks only English. Send an email message to the teacher. • Tell her why you've been away. • Ask two things about the exam. • Ask two things about the English lessons that were held during the week. Remember to **begin** and **end** the message appropriately. Write **in English** in clear characters in the space below.	You've been away from school for a week. Soon you'll have a Finnish exam. Send an email message to the teacher. • Explain why you've been away. • Ask two things about the exam. • Ask two things about other events that occurred during the week. Remember to **begin** and **end** the message appropriately. Write **in Finnish** in clear characters in the space below.	more basic morphology

>>>

Task 3. Message to an internet store	Your parents have ordered a PC game for you from a British internet store. When you get the game you notice that it doesn't work properly. You get upset and decide to write an email message to the internet store. In the message, say • Who you are • What your parents ordered • Why you're unhappy (mention at least two defects/problems) • How you would like them to take care of the matter • Give your contact information Remember to **begin** and **end** the message appropriately. Write **in English** in clear characters in the space below.	Your big brother has ordered a PC game for you from an internet store. The game works badly. Write an email message to the internet store and say • Who you are • Why you're writing (mention two problems about the game) • How you would like them to take care of the matter • Give your contact information Write **in Finnish** in clear characters in the space below. Remember to **begin** and **end** the message appropriately.	more basic syntactic structure more frequent vocabulary more basic morphology less detailed contextualization
Task 4. Opinion	Choose one of the topics and write about what you think about the matter. Give reasons for your opinion. 1. Boys and girls should go to different classes at school. 2. No mobile phones at school! Write **in English** in clear characters in the space below (continues on the reverse side).	Choose topic 1 or 2 and write for the school paper about your opinion on the matter. Give reasons for your opinion. 1. No mobile phones at school! 2. Parents get to decide how children use the internet. Write **in Finnish** in clear characters in the space below. Write at least five sentences.	more basic syntactic structure more frequent vocabulary more basic morphology more detailed contextualization
Task 5. Narrative	Tell about the scariest / funniest / greatest experience in your life. Choose one. • Tell what happened (what, where, when, and so on). • Tell why the experience was scary / funny / great. Write **in English** in clear characters in the space below (continues on the reverse side).	Tell about one scary or funny thing that has happened to you. • What happened. • Why the experience was scary or funny. Write **in Finnish** in clear characters in the space below.	more basic syntactic structure more frequent vocabulary more basic morphology less detailed contextualization

APPENDIX 2
CEFLING rating scales (based on the CEFR levels)

	OVERALL WRITTEN PRODUCTION	WRITTEN INTERACTION	CORRESPONDENCE & NOTES, MESSAGES, FORMS	CREATIVE WRITING & THEMATIC DEVELOPMENT
A1	Can write simple isolated phrases and sentences.	Can ask for or pass on personal details in written form.	Can write a short simple postcard. Can write numbers and dates, own name, nationality, address, age, date of birth or arrival in the country, etc. such as on a hotel registration form.	Can write simple phrases and sentences about themselves and imaginary people, where they live and what they do.
A2	Can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but' and 'because'.	Can write short, simple formulaic notes relating to matters in areas of immediate need.	Can write very simple personal letters expressing thanks and apology. Can take a short, simple message provided he/she can ask for repetition and reformulation. Can write short, simple notes and messages relating to matters in areas of immediate need.	Can write about everyday aspects of his/her environment, e.g. people, places, a job or study experience in linked sentences. Can write very short, basic descriptions of events, past activities and personal experiences. Can write a series of simple phrases and sentences about their family, living conditions, educational background, present or most recent job. Can write short, simple imaginary biographies and simple poems about people. Can tell a story or describe something in a simple list of points.

>>>

B1	Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence.	Can convey information and ideas on abstract as well as concrete topics, check information and ask about or explain problems with reasonable precision. Can write personal letters and notes asking for or conveying simple information of immediate relevance, getting across the point he/she feels to be important.	Can write personal letters giving news and expressing thoughts about abstract or cultural topics such as music, films. Can write personal letters describing experiences, feelings and events in some detail. Can write notes conveying simple information of immediate relevance to friends, service people, teachers and others who feature in his/her everyday life, getting across comprehensibly the points he/she feels are important. Can take messages communicating enquiries, explaining problems.	Can write straightforward, detailed descriptions on a range of familiar subjects within his/her field of interest. Can write accounts of experiences, describing feelings and reactions in simple connected text. Can write a description of an event, a recent trip – real or imagined. Can narrate a story. Can reasonably fluently relate a straightforward narrative or description as a linear sequence of points.
B2	Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesising and evaluating information and arguments from a number of sources.	Can express news and views effectively in writing, and relate to those of others.	Can write letters conveying degrees of emotion and highlighting the personal significance of events and experiences and commenting on the correspondent's news and views.	Can write clear, detailed descriptions of real or imaginary events and experiences, marking the relationship between ideas in clear connected text, and following established conventions of the genre concerned. Can write clear, detailed descriptions on a variety of subjects related to his/her field of interest. Can write a review of a film, book or play. Can develop a clear description or narrative, expanding and supporting his/her main points with relevant supporting detail and examples.

>>>

APPENDIX 2 >>>

	OVERALL WRITTEN PRODUCTION	WRITTEN INTERACTION	CORRESPONDENCE & NOTES, MESSAGES, FORMS	CREATIVE WRITING & THEMATIC DEVELOPMENT
C1	Can write clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion.	Can express him/herself with clarity and precision, relating to the addressee flexibly and effectively.	Can express him/herself with clarity and precision in personal correspondence, using language flexibly and effectively, including emotional, allusive and joking usage.	Can write clear, detailed, well-structured and developed descriptions and imaginative texts in an assured, personal, natural style appropriate to the reader in mind. Can give elaborate descriptions and narratives, integrating sub-themes, developing particular points and rounding off with an appropriate conclusion.
C2	Can write clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader to find significant points.	As C1	As C1	Can write clear, smoothly flowing, and fully engrossing stories and descriptions of experience in a style appropriate to the genre adopted.

APPENDIX 3
The Finnish National Core Curriculum scale for writing

A1.1	Can communicate immediate needs using very brief expressions. Can write the language's alphabets and numbers in letters, write down his/her personal details and write some familiar words and phrases. Can use a number of isolated words and phrases. Cannot express him/herself freely, but can write a few words and expressions accurately.
A1.2	Can communicate immediate needs in brief sentences. Can write a few phrases and sentences about him/herself and his/her immediate circle (such as answers to questions or notes). Can use some basic words and phrases and write very simple main clauses. Memorized phrases may be written accurately, but prone to a very wide variety of errors even in the most elementary free writing.
A1.3	Can manage to write in the most familiar, easily predictable situations related to everyday needs and experiences. Can write simple messages (simple postcards, personal details, simple dictation). Can use the most common words and expressions related to personal life or concrete needs. Can write a few sentences consisting of single clauses. Prone to a variety of errors even in elementary free writing.
A2.1	Can manage in the most routine everyday situations in writing. Can write brief, simple messages (personal letters, notes), which are related to everyday needs, and simple, enumerated descriptions of very familiar topics (real or imaginary people, events, personal or family plans). Can use concrete vocabulary related to basic needs, basic tenses and co-ordinate sentences joined by simple connectors (and, but). Can write the most simple words and structures with reasonable accuracy, but makes frequent basic errors (tenses, inflection) and uses many awkward expressions in free writing.

>>>

APPENDIX 3 >>>

A2.2	Can manage in routine everyday situations in writing. Can write a very short, simple description of events, past actions and personal experiences or everyday things in his/her living environment (brief letters, notes, applications, telephone messages). Commands basic everyday vocabulary, structures and the most common cohesive devices. Can write simple words and structures accurately, but makes mistakes in less common structures and forms and uses awkward expressions.
B1.1	Can write an intelligible text about familiar, factual or imaginary topics of personal interest, also conveying some detailed everyday information. Can write a clearly formulated cohesive text by connecting isolated phrases to create longer sequences (letters, descriptions, stories, telephone messages). Can effectively communicate familiar information in the most common forms of written communication. Has sufficient command of vocabulary and structures to formulate most texts used in familiar situations, even if interference and evident circumlocutions occur. Routine language material and basic structures are by now relatively accurate, but some more demanding structures and phrases still cause problems.
B1.2	Can write personal and even more public messages, describing news and expressing his/her thoughts about familiar abstract and cultural topics, such as music or films. Can write a few paragraphs of structured text (lecture notes, brief summaries and accounts based on a clear discussion or presentation). Can provide some supporting detail to the main ideas and keep the reader in mind. Commands vocabulary and structures required for a relatively wide range of writing. Can express coordination and subordination. Can write intelligible and relatively accurate language, even if errors occur in demanding situations, text organisation and style and even if the influence of the mother tongue or another language is noticeable.

>>>

B2.1	Can write clear and detailed texts about a variety of areas of personal interest and about familiar abstract topics, and routine factual messages and more formal social messages (reviews, business letters, instructions, applications, summaries). Can express information and views effectively in writing and comment on those of others. Can combine or summarise information from different sources in his/her own texts. Can use broad vocabulary and demanding sentence structures together with linguistic means to produce a clear, cohesive text. Flexibility of nuance and style is limited and there may be some jumps from one idea to another in a long contribution. Has a fairly good command of orthography, grammar and punctuation and errors do not lead to misunderstandings. Contributions may reveal mother tongue influence. Demanding structures and flexibility of expression and style cause problems.
B2.2	Can write clear, detailed, formal and informal texts about complex real or imaginary events and experiences, mostly for familiar and sometimes unfamiliar readers. Can write an essay, a formal or informal report, take notes for future references and produce summaries. Can write a clear and well-structured text, express his/her point of view, develop arguments systematically, analyse, reflect on and summarize information and thoughts. The linguistic range of expression does not noticeably restrict writing. Has a good command of grammar, vocabulary and text organisation. May make mistakes in low-frequency structures and idiomatic expressions and style.
C1.1	Can write clear, well-structured texts about complex subjects and express him/herself precisely, taking the recipient into account. Can write about factual and fictional subjects in an assured, personal style, using language flexibly and diversely. Can write clear and extensive reports even on demanding topics. Shows command of a wide range of organisational means and cohesive devices. Has a very wide linguistic range. Has a good command of idiomatic expressions and common colloquialisms. Has an extremely good command of grammar, vocabulary and text organisation. May make occasional mistakes in idiomatic expressions and stylistic aspects.